2024, 10 (1)

ARGUMENTA

The Journal of the Italian Society for Analytic Philosophy

First published 2024 by the University of Sassari

Argumenta is the official journal of the Italian Society for Analytic Philosophy (SIFA). It was founded in 2014 in response to a common demand for the creation of an Italian journal explicitly devoted to the publication of high quality research in analytic philosophy. From the beginning *Argumenta* was conceived as an international journal, and has benefitted from the cooperation of some of the most distinguished Italian and non-Italian scholars in all areas of analytic philosophy.

Contents

Editorial

Welcome to the *Metaphysical Issue*!

This is how I like to consider it, given that its three major parts are each devoted to as many important topics within current metaphysical research.

The first part consists of the Special Issue on *Epistemology of Metaphysics*, edited by Lorenzo Azzano, Massimiliano Carrara, and Vittorio Morato. Starting from the conviction that, without a properly developed epistemology, the prospects for a fully mature analytic metaphysics would not be complete, the editors clarify in their "Introduction" that epistemology may be the best way to prevent metaphysics—notoriously a highly abstract reflection—from becoming too distant from both scientific and everyday practice. Indeed, questions such as "How are we supposed to know whether metaphysical statements are true?", "Are they similar to mathematical or physical truths?", "Are they known a posteriori or a priori?", and the like, can pave useful ways to keep metaphysics from straying too far.

The second part of the present issue is a discussion of Jessica Wilson's *Metaphysical Emergence*, a book that is noteworthy in more than one respect. The book clarifies once and for all the correct uses that the term "emergence" should be put to and, as one of the discussants (Karen Bennett) emphasises, one of its many virtues is "its engagement with, and reliance upon, classic older work in the metaphysics of mind [such as that] by people like Terence Horgan, Jaegwon Kim, Andrew Melnyk, Sydney Shoemaker, and Stephen Yablo". And Wilson's book makes a significant addition to the classics.

Argumenta 10, 1 (2024): 3-4
ISSN 2465-2334

In the *Précis* provided by Wilson, the reader will find one of the most crystal-clear expositions of a central metaphysical question and, both in the eight discussion papers and in Wilson's replies, a lively example of a master philosophical discussion.

The third part of this issue is a discussion between Harold Noonan and Alfonso Muñoz-Corcuera of an article by Eric Olson that was published in this journal in 2019 devoted to the consequences for personal identity that Derek Parfit drew from the possibility of fission. Olson had criticized these consequences, stressing that they follow only if we make a specific assumption: Noonan here disagrees with this conclusion, and Muñoz-Corcuera disagrees in turn with Noonan, thus giving rise to a lively exchange at one remove from the original debate.

The present number also includes five articles that have already appeared in 'early view' (by Sofia Bonicalzi and Mario De Caro, Daniele Cassaghi, Giulia Lasagni, Karsten R. Stueber, and Steven Umbrello and Maurizio Balistreri). They have already made and will continue to make significant contributions to discussion in their respective fields.

The number is then rounded off by the section of Book Reviews. We are proud to offer readers three thoughtful reviews of as many interesting new books.

Finally, the editors of the Special Issue on the *Epistemology of Metaphysics* would like to thank Tobia Fogarin for his help in formatting the papers. For my part, I would like to thank all the colleagues who have acted as external referees, the members of the Editorial Board, the editors of the Book Reviews, the assistant editors, and the team of librarians from the University of Sassari. All of them have been very generous with their work, advice, and suggestions.

As usual, the articles appearing in *Argumenta* are freely accessible and freely downloadable, therefore it only remains to wish you:

Buona lettura!

Massimo Dell'Utri
Editor-in-Chief

Argumenta 10, 1 (2024)
Special Issue

Epistemology of Metaphysics

Edited by

Lorenzo Azzano, Massimiliano Carrara,
Vittorio Morato

The Journal of the Italian Society for Analytic Philosophy

Contents

Epistemology of Metaphysics: An Introduction

Lorenzo Azzano, Massimiliano Carrara,** Vittorio Morato***

* *University of Milan*

** *University of Padua*

1. Some Words on the Epistemology of Metaphysics

The widespread development of metaphysical debates in the last decades of analytic philosophy has been accompanied by deep-seated doubts about the very viability and ambitions of metaphysics. For instance, Hirsch's quantifier variance has brought back into the spotlight the Carnap/Quine dichotomy on the status of ontology and metaphysics—indeed many share the suspicion that many issues of metaphysics do not have the theoretical significance they are thought to have, and rather display the superficiality and arbitrariness of questions like "does a fist come into being when I close my fingers?". Therefore, "easy ontologies" now abound (Thomasson 2015).

Alternatively, Ladyman *et al.* (2007) have famously launched an assault against "scholastic metaphysics", too detached from any actual scientific research to be relevant, which is to be substituted by a "naturalized metaphysics"—although Paul (2012) has argued that metaphysics is not so distinct in methodology from science.

Metaphysics is seen as contiguous with science, either because it shares some methods and tools with science or because it aims to unify the sciences, as some proponents of naturalized metaphysics argue. This proximity might imply that metaphysical theories inherit some epistemic status from scientific theories. Since science undeniably provides knowledge, it's plausible to argue that metaphysics does too. Moreover, if the justification for scientific knowledge is empirical, then metaphysical knowledge could be empirically justified as well. However, it's not entirely clear whether the justification of scientific knowledge is purely empirical; nor is it clear whether using scientific methods in metaphysics necessarily means its knowledge is empirically justified. For example, mathematics is considered a part of science, but its justification is arguably not empirical. This raises the question: does metaphysics provide truths in the way physics does, or as mathematics does? Are metaphysical truths known a posteriori (based on experience and ob-

Argumenta 10, 1 (2024): 9—12
ISSN 2465-2334

DOI 10.14275/2465-2334/202419.int
First published: dd November 2024

servation) or a priori (based on reasoning without recourse to experience and observation)? Merely recognizing a sort of proximity of metaphysics to science does not seem sufficient to resolve this issue.

Friends of metaphysics may find some of these doubts perplexing. At the end of the day, some claims of "traditional" metaphysics appear, on the face of it, as perfectly clear statements concerning an outside reality and its many features (e.g. modal realism, mereological nihilism). Methodological doubts, at this stage, may take the form of an epistemological question: how exactly, are we supposed to know whether these statements are true?

More generally, how is the investigation of this outside reality meant to proceed? Ultimately, discussing the epistemology of metaphysics is an effort in understanding the very nature of metaphysics as a discipline, its subject matter and the resources required to investigate it.

This is an old question, of course; built as a sort of generalization from Benacerraf's Dilemma in the philosophy of mathematics, Peacocke (1999: 1) proposes the so-called "Integration Challenge" viz. "the task of reconciling a plausible account of what is involved in the truth of statements of a given kind with a credible account of how we can know those statements". In many cases, this amounts to a reconciliatory challenge between a certain metaphysics, perhaps accompanied by its own ontology, and the correspondent epistemology. On the one hand, the Integration Challenge appears more pertinent to the more "robust" conceptions of metaphysics; that said, it is far from obvious that "easy" conceptions of metaphysics do not have their own epistemological and methodological hurdles to solve; usually, "easy" conceptions of ontology and/or metaphysics require the existence of certain privileged epistemological paths as opposed to others, which may be source of further discussions.

Yet, despite Peacocke's insistence, the level of sophistication and development in the epistemology of metaphysics is not even remotely comparable to those of metaphysics *tout court*, nor to those of the epistemological debates in general. Only very rarely the epistemology of metaphysics has been recognized and pursued as a field of inquiry in and of itself: the recent surge of interest in the epistemology of modality is an exception to the rule; this is no surprise, given the special status that metaphysical necessities have in metaphysics. However, given the far-reaching and multi-faceted nature of contemporary analytic metaphysics, we expect the epistemology of metaphysics proper to vastly outstrip modal epistemology.

In this special issue we want to *bring the epistemology of metaphysics to the forefront*. The objective of developing the epistemology of metaphysics is of paramount importance: for without a properly developed epistemology, one might think that the prospects for a fully mature analytic metaphysics would not be complete.

2. The Papers

In this special issue we have collected ten papers that, from different angles, all are engaged with the different aspects, challenges and features of the epistemology of metaphysics.

These ten papers could be organised in two groups. Five of them (Bryant, Snellman, Strollo, Takho, Wirling) tackle general epistemological/methodological questions on the status of metaphysical inquiry. The other five (Cortesi,

Dohrn, Lee, Schoonen, Sgaravatti) all are all engaged with epistemological questions related to specific metaphysical debates, in particular modal metaphysics and grounding.

In "Naturalized Metaphysics without Scientific Realism" Amanda Bryant aims to show that the project of naturalizing metaphysics does not require realist assumptions and that the project of naturalizing metaphysics can come apart from the assumption of realism; in particular she explores how the naturalist program can cohere with even a strong form of scientific antirealism.

In "Between Science and Logic: Securing the legitimacy of Analytic Metaphysics", Andrea Strollo defends the view that analytic metaphysics (or at least a significant portion of it) has the same kind of legitimacy that naturalized metaphysics has. The legitimacy of analytic metaphysics is secure by its methodological and thematic continuity with logic. A nice effect of this view, according to Strollo, is that the rivalry between naturalized metaphysics *vs* analytic metaphysics should be reconceived as a distinction between two different disciplines: philosophy of science and philosophy of logic.

In "Metaphysics as a Science: A Sketch of an Overview", Lauri Snellman sketches a pragmatist methodology for metaphysics. In his view, metaphysical inquiry should be usefully conceived as the result of the interaction of a bottom-up methodology, whose main aim is the description of language-games of some metaphysical relevant words ("there is", "all", "none") with a top-down methodology whose main aim is that of developing conceptual schemes for use as starting-points for scientific research.

In "Laws of Metaphysics for Essentialists", Tuomas Takho first argues in favour of the view that metaphysical inquiry plays a genuine explanatory role by means of laws of metaphysics. Such laws should be understood, for Takho, as counterfactual-supporting general principles that are responsible for the explanatory force of non-causal, metaphysical explanations. Second, he argues for a unification of metaphysical and scientific explanation by means of the notion of general essence.

In "Understanding with Epistemic Possibilities: The Epistemic Aim and Value of Metaphysics", Ylwa Wirling proposes a radical reconceptualization of the epistemic aims of metaphysics. According to Wirling, we should conceive metaphysical inquiry in a way that makes compatible the claims that at least some instances of metaphysical inquiry are assessed positively and that metaphysical inquiry is intrinsically plagued by systematic and persistent disagreement between researchers. The solution she proposes is based on the specification of a non-factive notion of understanding, placing the value of metaphysical inquiry mainly in its epistemic role.

In "The Thesis of Experiential Revelation in The Philosophy of Mind: A Guide for The Perplexed", Fabio Cortesi defends the view that awareness of our own phenomenal mental states constitutes a peculiar kind of knowledge and that we have good reason to think that this knowledge be essence-revealing. Cortesi then evaluates the consequences of this view for a materialist framework about phenomenal consciousness and about reality in general.

In "The Feasibility Approach to Imagination as a Guide to Metaphysical Modality", Daniel Dohrn presents a novel approach to modal imagination as a means of knowing metaphysical possibilities. The starting point is the "natural

inclination" to use imagination in simulating solutions to everyday feasibility issues. According to Dohrn, there is a continuity between this natural use of imagination and the use of imagination in tackling philosophical possibility issues.

In "The Pragmatics of Metaphysical Explanation: An Epistemology of Grounding", James Lee aims to show that realist analytic metaphysicians, in particular those engaged in the grounding debate, need not fear epistemic explanations or explanatory practices in general. Lee's approach in developing his epistemology of metaphysical explanation is based on the use of so-called *contrast classes* in order to confer justification for beliefs about metaphysical relations such as grounding.

In "What Everett Couldn't Know", Tom Schoonen criticizes the epistemic side of so-called quantum modal realism (defended by Wilson 2020), according to which modal metaphysical space could be described in terms of the many-worlds interpretation of quantum mechanics. Schoonen's point is that, from an epistemic point of view, such a view is in a worse condition than Lewis's modal realism. While quantum modal realists have surely the advantage of being able to subsume the epistemology of modality under the general epistemology of science, they would not be able, according to Schoonen, of explaining the ordinary way in which modal knowledge is obtained, given that such ordinary modal knowledge cannot rely on the findings of experimental and theoretical physics.

In "Essence and Knowledge", Daniele Sgaravatti defends a hybrid modal-epistemic account of essence according to which an essence is a set of cognitively significant properties with a certain modal profile. Such an epistemic element in the notion of essence is what best explains the various epistemic roles such a notion is designed to play.

There are many paths that the epistemology of metaphysics might take. Some have already been partially explored, while many others still await adequate development. We hope to that this SI will contribute to the progress of some of them.

References

Ladyman, J., Ross, D., Spurrett, D., and Collier, J. 2007, *Every Thing Must Go: Metaphysics Naturalized*, Oxford: OUP.

Paul, L.A. 2012, "Metaphysics as Modeling: The Handmaiden's Tale", *Philosophical Studies*, 160, 1-29.

Peacocke, C. 1999, *Being Known*, Oxford: OUP.

Thomasson, A. 2014, *Ontology Made Easy*, Oxford: OUP.

Naturalized Metaphysics without Scientific Realism

Amanda Bryant

University of Calgary

Abstract

Abstract: It is often assumed that a commitment to scientific realism naturally, if not necessarily, accompanies a commitment to naturalizing metaphysics. If one denies that our scientific theories are approximately true, it would be unclear why one should index metaphysics to them. My aim is to show that the project of naturalizing metaphysics does not require realist assumptions. I will identify two success conditions for the project of disentangling naturalized metaphysics from realism: 1) the *narrow* success condition, which requires the antirealist to explain why naturalized metaphysics is preferable to non-naturalized metaphysics, and 2) the *broad* success condition, which requires the antirealist to explain why naturalized metaphysics is preferable to metaphysical quietism. I believe that the antirealist can meet these conditions. Although I will not defend any definitive way of meeting them, I will explore argumentative avenues open to the antirealist. In particular, I will consider some conceptions of naturalized metaphysics, discuss their antirealist-compatible expected payoffs, and consider whether those payoffs enable the antirealist to meet the success conditions of the project. I will find that the antirealist has several argumentative avenues open to them.

Keywords: Naturalized metaphysics; Scientific realism; Epistemology of metaphysics; Epistemic value; Facticity.

1. Introduction

It is common to think that a commitment to scientific realism at least goes naturally with, if not necessarily accompanies, the project of naturalizing metaphysics. It is *prima facie* puzzling why one who does not believe that our scientific theories are approximately true would, at the same time, insist that metaphysics should be indexed

Argumenta 10, 1 (2024): 13—33
ISSN 2465-2334

DOI 10.14275/2465-2334/20240.bry
First published: dd November 2024

to them.[1] If one thought, further, that our scientific theories were false or likely false, one might reasonably believe that it would spell doom for a metaphysics based on or derived from science. What would be the value of a metaphysics anchored to false science? My aim in this paper is to show that the project of naturalizing metaphysics can come apart from the assumption of realism—and to explore how the naturalist programme can cohere with even a strong form of scientific antirealism.

I am not the first to notice or question the assumption that scientific realism is a precondition for naturalized metaphysics. Guay and Pradeau note that "a majority of proponents of scientific metaphysics adopt scientific realism… [and many] of them even suggest that scientific realism is a *necessary* component of every project in metaphysics of science" (2020: 1852). While realism is, they say, "a perfectly legitimate… position" (2020: 1853), they suggest that the metaphysics of science "should perhaps not attach itself too rapidly" to it (2020: 1852). That is because, *inter alia*, realism is "demanding and difficult to demonstrate" (2020: 1854), and its truth or falsity is "not already settled" (2020: 1855). In their view, presupposing realism "leads to excluding without good reasons some possible avenues for metaphysics of science" (2020: 1854). In much the same spirit, I wish to indicate the presence of some antirealist-compatible avenues for naturalized metaphysics. I do so because I consider antirealism (like realism) rationally permissible and because I think, with Guay and Pradeau, that it would be unwise to needlessly foreclose available options. In addition to identifying such options, this paper will begin to explore them in greater detail. In particular, I will examine which adjustments to the naturalist's philosophical package are forced by the denial of realism.

One parameter that arguably needs adjusting is the doxastic attitude that the naturalist takes toward the theories of naturalized metaphysics. Belief in the truth of those theories is clearly not on the table for an antirealist naturalist. One well-explored alternative to belief is van Fraassen's (1980) notion of *acceptance*. To accept a theory is to believe that the theory is empirically adequate and to commit to using its language and explanatory resources in further research.[2] I flag the issue of doxastic attitudes as one that the antirealist naturalist needs to consider, but it will not be my focus here.

My focus will instead be on various conceptions of naturalized metaphysics, as well as its aims, prospects, and value. I will explore how those parameters can be adjusted to form a cohesive package with antirealism. I will identify two conditions for successfully disentangling naturalized metaphysics from the assumption of realism, which I will call the narrow and broad success conditions. The narrow condition requires the antirealist to explain why naturalized metaphysics is preferable to non-naturalized metaphysics; the broad condition requires the antirealist to explain why naturalized metaphysics is preferable to metaphysical quietism. I will not assume the

[1] One who thinks that naturalizing metaphysics is not about contact with scientific theories but rather scientific practices does not appear to face the same *prima facie* puzzle (Waters 2014, 2017, 2018, 2019). An approach that attends to the complexity and plurality of scientific practices might sit more obviously well with certain localized antirealisms (ex. Ereshefsky 1998, 2018). This is an interesting and fruitful avenue of inquiry but not one that I will explore further here.

[2] For relevant applications of this notion, see Elgin 2017, Beebee 2018, and Rosen 2020.

burden of definitively meeting these conditions on behalf of the antirealist, but I will highlight a number of argumentative routes they might take.

Section 2 defines scientific realism and antirealism. Section 3 gives a general definition of naturalized metaphysics and discusses why it is often assumed to go hand-in-hand with realism. Section 4 outlines the narrow and broad success conditions for the project of disentangling naturalized metaphysics from the assumption of realism. Section 5 outlines the sorts of philosophical packages that are open to the antirealist. It considers specific conceptions of naturalized metaphysics, their envisaged payoffs, the compatibility of those payoffs with antirealism, and finally, whether the payoffs would enable the antirealist to meet the narrow and broad success conditions. I will identify a number of combinations that could, with further argument, do the trick. Section 6 concludes.

2. Scientific Realism

There are many substantively distinct formulations of *scientific realism*. Some are axiological, in that they concern the aims of science (van Fraassen 1980), while others concern its actual accomplishments (Boyd 1983, Devitt 1997, Psillos 1999). Some are ontological, in that they concern the mind-independent existence of the unobservables posited by science; some are semantic, in that they concern the truth or successful reference of scientific theories; and some are epistemological, in that they concern knowledge or justified belief with regard to scientific theories (see Chakravartty 2007). The general spirit of the view is captured by its slogan formulation, which states that *our best current science is approximately true.* This slogan is loaded; each of its constitutive notions—'best', 'current', 'science', 'approximately true'—is vague and requires elucidation. Realists have devoted substantial effort to that task, with special attention to the meaning of 'best' (often cashed out in terms of maturity) and of 'approximately true'.[3]

The slogan formulation is a *wholesale* formulation (Magnus and Callender 2004) in that it generalizes about science on the whole. For my purposes here, it will be important to construe realism broadly, so that it includes both wholesale varieties and more *selective* varieties—that is, varieties that attach realist commitment to systematically identifiable parts of science. It is important to do so because it has already been established that the naturalist can do without wholesale realism. The arguable progenitors of recent interest in naturalized metaphysics, Ladyman and Ross, are not themselves wholesale realists but rather selective ones. Thus, I will define scientific realism (or just 'realism') in the following disjunctive way.

> *Scientific realism*: Either our best current science is approximately true or significant parts of it, which are identifiable in a non-*ad-hoc* way, are.

Scientific antirealism (or just 'antirealism') is likewise formulated in a variety of substantively distinct ways. I will define it as the negation of realism.

> *Scientific antirealism*: It is the case neither that our best current science is approximately true nor that significant parts of it, which are identifiable in a non-*ad-hoc* way, are.

[3] See for example Hunt 2011; Psillos 1999; Smith 1998; Weston 1987, 1992; and Worrall 1989.

On this characterization, the antirealist believes in the substantial falsity of our best current science. Some might find the strength of this formulation unpalatable. There are certainly humbler forms of antirealism. For instance, one might adopt the view that we cannot know or be justified in believing our best current science and that the best policy is to suspend judgment. In adopting such a view, the antirealist would play the role of the skeptic. By no means do I wish to assert that the strong form of antirealism is the most attractive or defensible one. I have defined antirealism in this strong way because doing so presents the greatest challenge to my present aims and so makes for a more significant outcome if I am successful. The challenge is to show how someone who believes that science is substantially false could at the same time believe that naturalizing metaphysics is desirable or even requisite, perhaps even for epistemic reasons.[4] I am optimistic that the challenge can be met, which is, I think, important and interesting. If I am right, then it should be comparatively easy to square more modest forms of antirealism with the naturalist programme in metaphysics.

3. Naturalized Metaphysics and the Assumption of Realism

Just as there is a heterogeneous family of scientific realisms and antirealisms, the view *that metaphysics ought to be naturalized* has been cashed out in a number of distinct ways. As a terminological note, I will reserve the term 'naturalist' for one who adopts that view, which is a local form of methodological naturalism, not to be confused with numerous other non-equivalent senses of the term.[5] Moreover, while others may wish to preserve distinctions among the following terms, I will consider 'naturalized metaphysics' (and, equivalently, 'naturalistic metaphysics') to be co-extensive with 'scientific metaphysics', 'metaphysics of science', 'science-guided metaphysics' and 'scientific ontology'. While these terms have been characterized in different ways, they typically mean something like the following:

> *Naturalized metaphysics:* Metaphysics that engages with science in some substantive way.

The naturalist's immediate challenge is to define 'metaphysics' and 'science' in a way that makes the view contentful (Chakravartty 2017, Williamson 2013). For my purposes here, I will take the respective academic institutions and their activities as rough proxies for what is intended by the terms. There are differing conceptions of the appropriate *modes* of engagement, higher and lower bars for what counts as an adequate

[4] Compare what McKenzie calls *the progress problem*: "the science upon which contemporary [science-guided metaphysics] relies is overwhelmingly likely to be false, meaning that a metaphysics based on it is likely to be false also. Given that—unlike in science itself—there is also no clear sense in which metaphysical claims can at least be said to be 'making progress,' the epistemic value of a present-day metaphysics that is based in current science becomes very difficult to discern" (2021: 436). See McKenzie 2020 for greater detail.

[5] The view that metaphysics should be naturalized is *local* in that it pertains to metaphysics only; it is *methodological* in that it is a methodological prescription; it is a form of *naturalism* in that it prescribes engagement with science. For more discussion of local and non-local methodological naturalisms and how they differ from other forms of naturalism, see Bryant 2020b. See also Papineau 2014.

degree of engagement, and different views about the precise *object* of engagement—that is, which sciences ought to be privileged and why. However, the common denominator is that metaphysics should not float entirely free of science; it should not be what I have called "free range" metaphysics (Bryant 2020a).

One natural rationale for thinking that metaphysics should not float entirely free of science is the belief that the domain of metaphysical fact does not float entirely free of the domain of scientific fact. On such a view, *contra* Kant, it is not that there are two discrete levels of reality, the empirical and the properly metaphysical, only the former of which is revealed by science. The domain of metaphysical states of affairs is not distinct from nature and in principle epistemically inaccessible via the methods of science. Rather, science and metaphysics have, at least to some extent, a shared target of inquiry. Since the domains of metaphysical and scientific interest overlap to some extent, science is to a proportional extent a source of evidence relevant to metaphysical matters.[6]

Moreover, the thought continues, science is a *good* source of evidence concerning such matters. This is where realism finds its natural entry point. In explaining what makes science an especially *good* source of pertinent evidence, it is tempting for the naturalist to appeal to realism directly or indirectly. She might invoke realism relatively directly by claiming that science gives us a *true* picture of reality, generates *knowledge* of it, or reveals *facts* about it—these factive notions being signals of realist commitment. Alternatively, she might invoke realism indirectly by gesturing toward properties of science that frequently motivate realism, such as its unparalleled success (e.g. Ladyman and Ross 2007: 7). In sum, since the naturalist motivates her project by appeal to the goodness of scientific evidence, and since realism offers a straightforward basis for considering scientific evidence good, it is natural to assume that realism accompanies the project of naturalization.[7]

Indeed, many philosophers draw the connection explicitly. For instance, according to Hawley, whether one is optimistic or pessimistic about the prospects for making metaphysical progress on the back of scientific progress is, in large part, "parasitic upon debates and decisions about scientific realism" (2006: 468). She explains: "it should come as no surprise that anyone who is sceptical about the ability of science to give us knowledge of quarks and quasars will be sceptical about whether science

[6] A reviewer worries that the just-so story I am telling on behalf of the naturalist fails if science and metaphysics operate at such different levels of description that they are incommensurable. We know they are not incommensurable, since science demonstrably speaks to metaphysics by informing and standing in evidential relations to it (existing naturalized metaphysics supplies the proof). If they are partially incommensurable, then one of the limits on naturalized metaphysics will be the limits of commensurability. It's up for debate where those limits fall, but this has the air of a feature rather than a bug.

[7] While I have suggested that realism naturally enters the scene in a justificatory capacity, for the role it plays in giving the naturalist reason to positively assess or privilege scientific evidence, others have imagined the relationship between realism and naturalism somewhat differently. For instance, Devitt does not see realism as playing a justificatory role with respect to naturalization but as an inevitable outgrowth of an antecedent commitment to naturalism. In his view, "when we approach our metaphysics empirically, Realism is irresistible" (1999: 96).

can give us knowledge of universals and possible worlds" (2006: 454). Conversely, optimism about science may translate into optimism about naturalized metaphysics:

> [D]ifferent naturalisers will take different approaches. But one attractive option is to see the naturalising metaphysician… as a kind of scientific realist, who uses inference to the best explanation to move from the empirical successes of a scientific theory to the accuracy of the metaphysical picture embedded in the theory. (Hawley 2018: 189)

Ladyman and Ross draw a connection between the truth-conduciveness of science and that of naturalized metaphysics: "[t]he naturalistic metaphysician… is optimistic about the possibility of bringing metaphysical hypotheses into closer conformity with objective reality to the extent that these hypotheses non-trivially unify bodies of *established scientific knowledge*" (my emphasis 2013: 109). The reference to scientific knowledge—and, indirectly, the suggestion that a metaphysics that engages with that knowledge has a better shot at 'conforming to objective reality'—indicates realist commitment. Schrenk also comments on the connection between naturalized metaphysics and realist commitment: "Philosophers who engage with the metaphysics of science tend to sympathize in one way or another with science itself… they see science… as the single most important, most reliable path to truth" (2017: 296). In that way, he says, "Scientific Realism is at least an ally to metaphysics of science" (2017: 298). The phrase is apt; scientific realism is often taken to be at least an ally to naturalized metaphysics if not a presupposition of it (as in Esfeld 2009).

4. Success Conditions for Naturalized Metaphysics Without Realism

We have seen that scientific realism figures into one obvious rationale for naturalized metaphysics, so it should come as no surprise that a naturalist who avows antirealism would need some alternative account of its rationale. If one thinks that our best current science is substantially false, then why bother with a metaphysics that is anchored to it? There are two preliminary explanatory challenges for the antirealist naturalist, which I will refer to as, respectively, *narrow* and *broad* success conditions. These will be success conditions for the project of disentangling naturalized metaphysics from the assumption of realism.

The first is well-encapsulated by a passage from Chakravartty. He remarks that the naturalist:

> …must assume that some parts of scientific theories are likely to be retained over time across theory change, and furthermore, that we are in a position to identify at least some of these parts. Without some such identification… the scientific ground of naturalized metaphysics would inevitably shift significantly in time… [and] one would have no good reason to suspect that metaphysics done in conjunction with it at any given time is preferable to metaphysics that is alien to it. (2013: 39)

This line of reasoning shows that realism plays an important explanatory role for the naturalist: it justifies her preference for her own metaphysical approach. The antirealist naturalist needs to explain the preferability of her approach in some other way. Here we have our first success condition.

> *Narrow success condition*: The antirealist naturalist must give some reason why naturalized metaphysics remains preferable to non-naturalized metaphysics notwithstanding the falsity of science.

Virtually all naturalists in my sense of the term are strongly committed to the superiority of naturalized metaphysics to non-naturalized metaphysics. To maintain that commitment, they will need to meet this narrow success condition.

Regarding the formulation of the condition, I acknowledge that without explicit precisification, 'preferable' isn't particularly contentful. The formulation immediately raises the question, 'preferable *how*?'. This indicates that to determine whether the condition is satisfied, we need a criterion of preferability. The same will be true of the second success condition. I have intentionally left this open because I wish to canvas, in an exploratory spirit, some of the many and varied reasons an antirealist might have for preferring naturalized metaphysics, as well as the sorts of epistemic and non-epistemic criteria of preferability they might apply. I leave it to the reader to judge which of these reasons and criteria are compelling—but the heterogeneous results are, I think, deeply interesting.

Still, one might worry that this open approach renders my ultimate conclusion—that there are plenty of argumentative avenues open to the antirealist naturalist—unsurprising. That there are plenty of avenues open to the antirealist naturalist is a consequence of the permissive way I define success.[8] One might hope to see, for instance, a specifically epistemic restriction on the kind of preferability that must be shown—that naturalized metaphysics has distinctively epistemic payoffs even when paired with the assumption of strong antirealism. I invite readers who share this concern to interpret the success conditions epistemically. The approach discussed in section 5.1 won't obviously meet the conditions so interpreted, but more promising options will be discussed in 5.2 and 5.3. Readers who are content to take a more exploratory approach can, with me, leave the success conditions open to a wider range of interpretations.

The second success condition emerges when one considers the anti-metaphysical spirit of many antirealisms, including van Fraassen's constructive empiricism.[9] Van Fraassen's constructive empiricism famously commits not to the truth of scientific theories but instead to their empirical adequacy. He comments that "the assertion of empirical adequacy is a great deal weaker than the assertion of truth, and the restraint to acceptance *delivers us from metaphysics*" (my emphasis 1980: 69). The constructive empiricist has, as part of their philosophical temperament, a generalized aversion to metaphysics. For many constructive empiricists, an anti-metaphysical temperament

[8] I thank a reviewer for bringing this criticism to my attention.

[9] In this paragraph, I use 'antirealism' in a broad sense, not in my narrow sense. I take van Fraassen's constructive empiricism to be one of the humbler and more skeptical kinds of antirealism discussed above.

is part of what disposes them to constructive empiricism in the first place. While there is nothing essentially anti-metaphysical about antirealism in principle, it does often have an anti-metaphysical spirit. One might wonder whether those antirealists who seek to avoid metaphysics are comparatively wise, given the epistemically risky nature of metaphysics. Perhaps the antirealist is better off not bothering with metaphysics, naturalized or not. This thought leads us to the second success condition of the project.

> *Broad success condition*: The antirealist naturalist must give some reason why pursuing naturalized metaphysics is preferable to avoiding metaphysics entirely notwithstanding the falsity of science.

They should, in other words, say something about the value of naturalized metaphysics that would make it worth the epistemic risk it entails.

One might be tempted to dispense immediately with the broad success condition by invoking the familiar claim that *metaphysics is unavoidable.* More than one philosopher has remarked upon the apparent indispensability of metaphysics to human thought (see Kant's *Prolegomena* 4:367 and Peirce *CP*, 1.129). They argue that metaphysics inevitably emerges in our thinking whether we like it or not, and without recognized standards or criticism it takes its own haphazard shape. One who is compelled by remarks such as these might reason that, since we are bound to do metaphysics implicitly anyway, we might as well come out in the open and do it *explicitly* and *in the best possible way*, which the naturalist believes is in concert with science.

Yet the skeptic might respond that, while we cannot entirely avoid metaphysical concepts and assumptions, that does not mean we should jump headfirst into the organized pursuit of metaphysics. While we cannot avoid metaphysical concepts and terms, for instance, we can, as a regulative ideal, do our best to minimize how much metaphysical theorizing we actively do. Thus, the skeptic may argue, the broad success condition should be understood as a demand to know why we should attempt naturalized metaphysics rather than declining to pursue organized metaphysics and instead adopting metaphysical quietism *so far as possible*. What, in other words, does the naturalist think makes organized metaphysics worth saving?

The realist naturalist should, of course, have an answer to that same question. However, their optimism about the capacity of science to generate knowledge, some of which concerns metaphysical matters, gives them a straightforward path to an answer: we should save organized metaphysics because we can pin our hopes for metaphysical knowledge on the back of our trust in scientific knowledge. The path for the antirealist is not so straightforward. If one has reason to reject our best science, then one has reason to reject the naturalized metaphysics drawn from it.[10] But if naturalized metaphysics is not plausibly approximately true, then why bother with it at all?

[10] I do not mean to suggest that the truth of antirealism would logically entail the falsity of naturalized metaphysics. It would not. It is just that the naturalist thinks that the relevant epistemic relations and properties track: if science is justified, then naturalized metaphysics is also justified (albeit to a lesser degree); conversely, if science is rationally disbelieved, then so too is naturalized metaphysics. Of course naturalized metaphysics could turn out to be true notwithstanding the falsity of the science it was based on, but this would be wildly coincidental, and nobody would be justified in expecting it.

Some suggest that a satisfying answer is not forthcoming. For instance, Ladyman and Ross approvingly cite van Fraassen's belief that, in their words, "a metaphysics that is not at least broadly true… is worthless… [and] should be abandoned" (2013: 109).[11] One of my aims here is to challenge that sort of view. In the remainder of the paper, I will explore some candidate values in virtue of which naturalized metaphysics would be worth pursuing even if it were false. I will identify a number of payoffs relative to which an antirealist could meet the narrow and broad success conditions I have outlined. My aim is not to defend or privilege any particular one but rather to explore options. I will conclude that there are plenty of ways in which antirealism can sit comfortably with the naturalist programme.

I do not believe the antirealist naturalist who explores these sorts of alternate values should be viewed as cornered and desperate—or as pursuing what should be regarded as "a last resort", as McKenzie puts it (2020: 24). There is rich and fertile philosophical terrain to be trodden here concerning non-factive epistemic aims and values, how they relate to and trade off against one another, what happens when we privilege some over others, and their potential to shift ongoing philosophical dialectics. So I take the antirealist naturalist to be at an exciting juncture in theoretical space.

5. Naturalized Metaphysics: Conceptions and Expected Payoffs

In what follows, I will consider whether naturalized metaphysics can have antirealist-compatible 'payoffs' (that is, whether it can promote truth-independent values) that give the antirealist naturalist the resources to meet the narrow and broad success conditions. Since there are many specific conceptions of naturalized metaphysics, I will proceed by surveying a small sample of them, examining the explicit rationale and expected payoffs of each, and considering whether each assumes realism. Where realism is assumed, I will consider whether that same metaphysical project could be pursued for other reasons, with antirealist-compatible payoffs in view. In each case, I will consider whether the antirealist-compatible payoffs could enable the antirealist naturalist to meet the narrow and broad success conditions I have outlined.

5.1 Quinean Ontology

I begin with Quine, whose conception of naturalized metaphysics is familiar. On Quine's conception, naturalized metaphysics involves deductively deriving ontological commitments from regimented science. Regimenting a scientific theory involves clarifying and simplifying it by translating it into logical notation.

Quine is a realist, and he takes naturalistic philosophy to begin with realist assumptions. Quine views the scientific method as the "way to truth" and the "last arbiter of truth" (1960/2013: 21). Moreover, according to Quine, "[t]he naturalistic philosopher begins his reasoning within the inherited [i.e. scientific] world theory as a going concern. He tentatively believes all of it, but believes also that some unidentified

[11] Note that this is a departure from their claim, to be discussed below, that naturalized metaphysics is *probably false* yet still desirable insofar as it is the "best metaphysics we can have at [time] t" (2007: 2).

portions are wrong" (1981: 72). This is consilient with the realist's commitment to approximate truth.

What does Quine take to be the expected payoff of his form of ontology, and does it hinge on the assumption of realism? The ontologist's regimentation of science is rationally considered truth-conducive, according to Quine. He remarks that "simplicity, in a theory that squares with observation sentences so far as its contacts with them go, is the best evidence of truth we can ask; no better can be claimed for the doctrines of molecules and electrons" (1960/2013: 230). Further, he claims that "[t]he quest of a simplest, clearest overall pattern of canonical notation is not to be distinguished from a quest of ultimate categories, a limning of the most general traits of reality" (1960/2013: 147). In other words, the Quinean ontologist aims to derive a true ontology. The expected payoff here is a true metaphysics—or, at least, one whose truth is best evidenced.

While realist commitment factually underlies Quine's naturalistic approach to metaphysics, we can ask whether realism is essential to it. If one did not expect science to supply us with a true ontology, why else might one pursue Quinean ontology? I suggest that one place to look for an answer is applied ontology. Applied ontology attempts to discern the ontological commitments of concrete domains in science, industry, and government and how they can be systemized into classification systems that enable consistent representation of information.[12] Quine's conceptions of ontology and ontological commitment are foundational to applied ontology (Smith 2003, 2014). Indeed, in many of its applications, applied ontology does something generally resembling the Quinean project: it uses the tools of logic to limn the ontological commitments of the sciences. Granted, such projects go well beyond Quine's vision of ontology in terms of the rich suite of sophisticated methods they implement (Arp et al. 2015). Nevertheless, they can be understood as an extension of the Quinean tradition.

Applied ontology also responds to some pressing practical needs, including, among others, those generated by big data. Scientists are to an increasing extent dealing with vast quantities of rich data. However, their datasets are often "characterful" in the sense that they have diverse contents and structures and are full of gaps, inconsistencies and uncertainties (Cooper and Green 2016). This makes the data immensely difficult to work with and draw conclusions from. In such contexts, applied ontology serves the purpose of cleaning up problematic datasets, by making the data more consistent, complete, and ultimately more useful. This is a pragmatic payoff. Moreover, it is antirealist-compatible, because it does not require the truth of science.

The final question to consider in relation to Quinean approaches is whether the suggested truth-independent payoff would give the antirealist naturalist the resources to meet the narrow and broad success conditions. That is, would the pragmatic benefit of making large scientific datasets more useful make the sort of Quinean applied ontology that I have described: 1) preferable to non-naturalized metaphysics and 2) preferable to no organized pursuit of metaphysics? The answer in both cases is, arguably, yes.

[12] See Lean 2021 for a discussion of the metaphysical import of such ontologies—especially in relation to naturalized metaphysics.

The first thing to consider is whether non-naturalized metaphysics could have the same payoff as Quinean applied ontology—that is, whether non-naturalized metaphysics could serve to clean up characterful scientific data. Suppose for the sake of argument that it could. It would have to serve that purpose either by design or by accident. But non-naturalized metaphysics has nothing directly to do with large scientific datasets. As a matter of definition, such metaphysics does not engage with science directly. So any metaphysics that set out with the explicit aim of cleaning up large scientific datasets wouldn't count as 'non-naturalized' in the first place. For it would be constrained by the content of the science to which it was designed to apply. Therefore, if the conceptual resources of a truly non-naturalized metaphysical system somehow helped to clean up some scientific dataset, it would have to be entirely incidental. By working directly with such datasets with the express aim of mitigating their characterful quality, Quine-style applied ontology would achieve the relevant payoff more consistently and predictably. And while the potential for non-naturalized metaphysics to produce such a payoff is merely hypothetical, the capacity of Quinean applied ontology to do so is demonstrable. So, relative to the payoff of cleaning up characterful scientific data, Quinean ontology is preferable to non-naturalized metaphysics.

Moreover, I doubt it would be terribly controversial to suggest that making scientific datasets more useful is something worth doing. One who recognizes the value therein would conclude that doing Quinean ontology is preferable to doing no metaphysics. The upshot is that I have found just what the antirealist naturalist needs: an approach to naturalized metaphysics that plausibly has an antirealist-compatible payoff, relative to which the narrow and broad success conditions can be met.

One might point out that, in this first case, the success conditions have only been met by retreating to the domain of pragmatic value. That may be so, but there is nothing illicit about it. We are seeking antirealist-compatible reasons to pursue naturalized metaphysics; those reasons need not be epistemic. However, I do think the antirealist naturalist can hope for more than what might be characterized as *merely* pragmatic value. While it is harder to show that naturalized metaphysics can have properly epistemic value where antirealism is held fixed, we will see that there are some interesting possibilities in that regard.

5.2 Science-Unifying Metaphysics

Ladyman and Ross defend an alternate conception of naturalized metaphysics in their rallying cry for the naturalization of metaphysics (2007, 2013). On their picture, naturalized metaphysics is an exercise in unifying scientific theses while privileging fundamental physics. This involves showing how two or more scientific theses explain more together than they do individually. Call this sort of naturalized metaphysics *science-unifying metaphysics*.

Realism operates in the background in the form of *structural realism*. According to structural realism, the structural, i.e. mathematical, content of our best scientific theories is carried over in limiting cases across theory change, and it is *that* content that realist commitment should track. According to Ladyman and Ross' eliminative ontic structural realism, the structural content describes the underlying structure of reality—which is composed of relations all the way down— approximately correctly.

Ladyman and Ross claim that, in wedding itself to successful scientific theories that make novel predictions and give "correct descriptions of the structure of the world" (2007: 92), science-unifying metaphysics is the only sort of metaphysics that qualifies as a "legitimate part of our collective attempt to model the structure of objective reality" (2007: 1). The science-unifying metaphysician can be "optimistic about bringing metaphysical hypotheses into closer conformity with objective reality" and thereby "contributing to objective knowledge" (2013: 109). So, like Quine, Ladyman and Ross think that the expected payoff of naturalized metaphysics is *true* metaphysics, or at least, metaphysics whose convergence on truth we can be optimistic about—and this expected payoff hinges on their realism.

If we replace the structural realist's optimism with antirealist pessimism, what—if any—value could there be in unifying scientific theses that one believes to be false? Given that unification is understood to increase explanatory scope (i.e. how much is explained) one might argue that there is epistemic value in relatively great explanatory scope, *even if the explanation on offer is in fact false.* On traditional conceptions of explanation—namely, factive ones—what I've just suggested is incoherent. However, the recent "non-factive turn" in epistemology (Turri 2018) develops non-factive conceptions of key epistemological notions. On such conceptions, being true is not a necessary condition of belonging to the relevant epistemological categories. In the case of explanation, one can have an explanation without it being the case that both explanandum and explanans are true (see for instance Bertrand 2022). This conception allows for the possibility of false explanations. The envisaged antirealist thinks that, insofar as science is substantially false, so too are the majority of its explanations. Yet the conceptual resources of the non-factive turn allow us to consider them explanations nonetheless. My suggestion here is that there could be something about explanation that is both independent of truth and epistemically valuable, in virtue of which explaining more is desirable.

For instance, the antirealist naturalist might appeal to Lipton's (2004) notion of the 'loveliness' of an explanation, i.e. the extent to which an explanation renders an explanandum intelligible. It seems conceptually possible that a false explanation could make an explanandum intelligible, just as, for instance, models and other idealizations can be heuristically valuable. A related approach would be to invoke a conception of non-factive understanding.[13] For instance, *understanding* can be thought of as the grasping of "a comprehensive set of interrelated propositions about [a] subject matter and how they relate to each other" (Sjölin Wirling 2021: 644). This does not require the truth of the propositions grasped, and it is an apparently epistemic value. Thus, the antirealist naturalist could argue that false explanations could be valuable in virtue of promoting non-factive understanding. Greater explanatory scope would then be valuable insofar as it would entail greater understanding.

In a similar spirit, Cartwright argues that *ceteris paribus* laws are "not even approximately true" (1983: 57), and yet they are explanatory. That is because they "organize, briefly and efficiently, the unwieldy, and perhaps unlearnable, mass of highly detailed knowledge that we have of the phenomena" (1983: 87). This organizing

[13] See Doyle et al. 2019, Elgin 2017, McSweeney 2023, Potochnik 2020, and Sjölin Wirling 2021.

power, she says, "has nothing to do with truth" (1983: 87). Organizing power seems at first glance pragmatically valuable, but Cartwright's reference to *learnability* suggests that it may also be considered epistemically valuable. The antirealist naturalist might extend this sort of view beyond the context of *ceteris paribus* laws to that of science-unifying metaphysics. They might argue that the false explanations of science-unifying metaphysics are valuable in virtue of their organizing power, and that greater explanatory scope is valuable because it entails greater organizing power.

There are many blanks to fill in, of course. How precisely do we further cash out intelligibility, understanding, and organizing power—and are they the same or different? Why are they valuable? What makes them *epistemic* values? Filling in these blanks clearly requires additional philosophical work. For my part, I wish only to flag this style of argument as a live possibility for the antirealist naturalist.

Regarding the narrow success condition, relative to the payoff of increased explanatory scope, science-unifying metaphysics arguably has an advantage over non-naturalized metaphysics. That is because science sets out to explain a wealth of data gathered through the observational and experimental practices that figure so prominently in scientific practice. In tying itself to the project of scientific explanation, science-unifying metaphysics ties itself to the explanatory aspirations of science and shares its vast explanatory scope. Non-naturalized metaphysics, by comparison, is not always clearly an explanatory enterprise, and when it is, its data are comparatively impoverished. They tend to be, primarily, everyday empirical appearances, common sense intuitions, the deliverances of thought experiments and other *a priori* modes of reasoning. So if greater explanatory scope is a boon, science-unifying metaphysics appears to be more valuable than non-naturalized metaphysics relative to it.

Yet one might wonder, even if naturalized metaphysics is preferable to non-naturalized metaphysics relative to explanatory scope, is it preferable from a broader lens that considers the aims and potential accomplishments of each? In particular, is it preferable to pursue naturalized metaphysics that aims to unify false science and thereby gain intelligibility, non-factive understanding, or organizing power—or to pursue non-naturalized metaphysics that aims at truth? The answer partly depends on how one weighs the relevant epistemic values and, granted, it's hard to top the value of truth. However, the answer also depends on how likely each inquiry is to reliably achieve its respective aims. I have argued elsewhere that because the constraints on the content of non-naturalized metaphysics are excessively permissive, non-naturalized metaphysics is unlikely to achieve truth or justification—and believing that it can or does reliably achieve those aims is a form of bad faith (Bryant 2020a). To establish the preferability of naturalized metaphysics relative to the lens of aims and accomplishments, the antirealist naturalist would need to show that naturalized metaphysics is comparatively more likely to achieve its aims reliably. While this remains to be shown, the bar for success is not particularly high.

Now turning to the broad success condition. Holding fixed antirealism, would possessing relatively great explanatory scope make science-unifying metaphysics preferable to no organized attempt at metaphysics? Here, the answer is conditional. It depends on whether a persuasive case can be made for the value of false explanations. I have not attempted to make that case here but have gestured toward some possible

avenues for further argumentation. The upshot is this: when naturalized metaphysics entails unifying scientific theses in the manner that Ladyman and Ross suggest, so long as one is willing to countenance non-factive accounts of explanation and of the value of explanatory unification, naturalized metaphysics has an antirealist-compatible, seemingly epistemic payoff relative to which the narrow and broad success conditions can be met.

5.3 Scientifically Informed Metaphysics

Perhaps the most common conception of naturalized metaphysics is one in which the metaphysician simply attends to science as she goes about her metaphysical theorizing. Chakravartty advances one such conception when he characterizes naturalized metaphysics as an exercise in making metaphysical claims and inferences that are "informed by, or sensitive to" the empirical aspects of science (2017: 76). He explains that for metaphysics to be 'sensitive to' or 'informed by' the empirical aspects of science is for those aspects to be a *basis* for and a *constraint* on metaphysical theorizing.[14] This, in turn, means that "the ground of empirical inquiry is the inspiration or motivation for certain metaphysical inferences… [and] the ground of empirical inquiry is being taken seriously as setting limits on the viable conclusions of those inferences" (2017: 84). While Chakravartty calls this project *scientific ontology*, to avoid any potential confusion with the Quinean project, I will call it *scientifically informed metaphysics.*

According to Chakravartty, the propositions of scientifically informed metaphysics characteristically carry lower epistemic risk than propositions with less empirical exposure, like those of non-naturalized metaphysics. This is one reason to prefer scientifically informed metaphysics to its rivals. Epistemic risk, Chakravartty says, is "a feature of propositions… that determines how confidently one is able to judge whether they are true or false; that is, whether and to what extent they are conducive to knowledge" (2017: 84). These are inversely correlated, such that the more confidently one can judge truth value, the lower the epistemic risk. Moreover, scientifically informed metaphysics does seek knowledge, Chakravartty says, and in any knowledge-seeking endeavour, "the less epistemic risk the better" (2017: 85). The expected payoff of scientifically informed metaphysics, then, is relatively low epistemic risk.

The mentions of truth, falsity, and knowledge might tempt one to conclude that realism is operating in the background here. But it need not be. Even if one thinks that knowledge is the aim of scientifically informed metaphysics, one might not think that the truth condition of knowledge is ever actually met. Moreover, notice that on Chakravartty's characterization, the inverse correlate of epistemic risk is not *confidence that p is true* but rather *confidence in one's judgment of the truth value of p.* We had better restrict that to *rational* confidence to rule out cases of unearned or unwarranted confidence. So if the antirealist is rationally confident that science is false, then they will be rationally confident that scientifically informed metaphysics is false—and thus sci-

[14] I have argued elsewhere that the constraining role of science with regard to naturalized metaphysics is epistemically significant and can explain the preferability of naturalized to non-naturalized metaphysics (Bryant 2021).

entifically informed metaphysics will carry relatively low epistemic risk notwithstanding its perceived falsity. This has the surprising consequence that having comparatively low epistemic risk is an antirealist-compatible payoff of scientifically informed metaphysics—or, more precisely, of its constitutive propositions.

Could the payoff of relatively low epistemic risk enable the satisfaction of the narrow and broad success conditions? First let us consider whether scientifically informed metaphysics is clearly preferable to non-naturalized metaphysics relative to considerations of epistemic risk. If we accept Chakravartty's conception of scientifically informed metaphysics and its relation to non-naturalized metaphysics, then it is clearly preferable in that regard. We saw that Chakravartty thinks the propositions of scientifically informed metaphysics lend themselves to more confident judgments of truth or falsity in virtue of their empirical exposure. Since such metaphysics is informed by the empirical aspects of science, that puts us in a relatively good position to judge whether its propositions are true or false.[15]

But how does all of this square with antirealism? The answer partly depends on what we think are the epistemically relevant features of *p's having low epistemic risk*—the features that explain why it is epistemically valuable. One might think that what is significant is that when *p* has low epistemic risk, we are in a relatively good *evidential* position relative to *p*. We are able to pronounce confidently on its truth value because we have lots of evidence pertaining to it. From that perspective, adding antirealism to the picture only helps matters. That is because, to the wealth of scientific evidence relevant to scientifically informed metaphysics, the antirealist adds additional evidence, such as evidence from the history of science. So one can argue that, relative to epistemic risk, scientifically informed metaphysics is preferable to non-naturalized metaphysics because it has more evidence that speaks to the truth or falsity of its claims.

All that remains is to consider whether, relative to epistemic risk, the antirealist has some reason to prefer scientifically informed metaphysics to no organized pursuit of metaphysics. Well, which is less epistemically risky: scientifically informed metaphysics or no metaphysics? At first glance, no organized pursuit of metaphysics carries *no* epistemic risk. Nothing ventured, nothing gained *or* lost. But on more careful consideration, we cannot compare the levels of epistemic risk assumed, respectively, by the scientifically informed metaphysician and the metaphysical quietist. That is because *p*'s epistemic risk corresponds, inversely, to one's degree of confidence in one's assessment of *p*'s truth value. The metaphysical quietist countenances no metaphysical propositions, and so makes no pronouncements upon truth or falsity in which to be confident or not. So the quietist gives us nothing to evaluate or compare in terms of epistemic risk. As a workaround, perhaps we could assess degrees of confidence not in metaphysical systems but in overall philosophical systems. But that will not work, because the antirealist believes aspects of her philosophical system (such as her epistemological principles) and merely accepts others (such as the propositions of

[15] That is not to say, however, that the naturalist is in an ideal position to make such judgments. Given underdetermination at various levels—of scientific interpretation by scientific theory, of scientific theory by empirical data, and by metaphysics by science—she is not in an ideal position. The claim is just that she is in a better position than her rivals.

scientifically informed metaphysics). At any rate, the upshot is that epistemic risk does not enable the antirealist naturalist to meet the broad success condition—at least not on Chakravartty's conception of epistemic risk. In sum, relative to the payoff of reduced epistemic risk, scientifically informed metaphysics arguably can meet the narrow success condition but can't meet the broad one.

*

I wish to discuss one final epistemic payoff of naturalized metaphysics, which is hinted at by the language of certain naturalists. For instance, Chakravartty says that the empirical aspects of science are a good ground for scientifically informed metaphysics because "empirical inquiry is our *best bet* for knowledge" (my emphasis, 2017: 85). In similar spirit, Maudlin writes that the metaphysician should interpret and elucidate physical theories because they "provide us with the *best handle we have* on what there is" (my emphasis, 2007: 1). Years prior to their remark on the worthlessness of false metaphysics (see §4 above), Ladyman and Ross commented:

> Since [scientific] knowledge can be incorporated into unified pictures, we can… have some justified metaphysics. Based as it is on incomplete science, this metaphysics probably is not true. However, if it is at least motivated by our most careful science at time t, then it is *the best metaphysics we can have at t*. (my emphasis, 2007: 2)

The common thread here is that science is our best form of inquiry about the world, and thus if we want to do metaphysics, then naturalized metaphysics will be the best form of metaphysics. These philosophers are working with different conceptions of naturalized metaphysics, and it is open to debate which conception is truly best and in which ways. For simplicity, I will continue to index the discussion to scientifically informed metaphysics, and I will suppose that it is the best metaphysics we can have at *t*.

None of this talk of bestness presupposes realism. Science can be epistemically best relative to the available alternatives without being true. Likewise, naturalized metaphysics can be best without being true. The way in which it is best—relative to knowledge, justification, understanding, explanation and prediction, or other epistemic goals—is open to precisification.[16] The point is that being best does not require successfully achieving whatever we take to be our epistemic aim—or that the epistemic aim be defined in terms of truth. Thus, being epistemically best is an antirealist-compatible expected payoff of scientifically informed metaphysics. With appropriate elucidation and argumentation, this could be a promising option for the antirealist naturalist.

[16] Presumably, the naturalist would say that naturalized metaphysics is 'best' *in the same sense* that science is best. For instance, they might say that science is our best shot at understanding the underlying nature of reality and that naturalized metaphysics is our best shot at understanding the same, say, at a greater level of abstraction. I assume that the operative senses of bestness would dovetail, because naturalists tend to think that the good-making features of science are inherited, to some extent, by naturalized metaphysics, and that is what typically explains the comparative desirability of naturalized metaphysics. But I won't foreclose *a priori* the possibility that the naturalist might find some way of arguing that science is best in one way and naturalized metaphysics in another. I thank Ylwa Sjölin Wirling for raising this issue.

Regarding the narrow and broad success conditions, the details would need to be filled in, but there is no in principle impediment to their satisfaction. If the claim that scientifically informed metaphysics is the best metaphysics we can have at *t* can be adequately spelled out and supported, then it would immediately follow that scientifically informed metaphysics is preferable to non-naturalized metaphysics. There could be a number of ways of doing this. One way might be to invoke Bayesianism and argue that the non-naturalistic metaphysician has limited evidence to conditionalize on and must therefore rely to an unacceptable extent on subjective priors. Bayesians sometimes argue that "prior opinion will tend to 'wash out' as believers acquire more and more information" (Joyce 2011: 445). With less data, subjective priors can exercise more influence. So scientifically informed metaphysics is arguably epistemically preferable relative to an epistemic policy that favours greater objectivity, understood in terms of the diminished role of subjective priors. The antirealist would then need to say why objectivity is epistemically valuable, independently of considerations of truth. This would be one way for the antirealist to flesh out the claim that naturalized metaphysics is the best metaphysics at *t*.

Regarding the preferability of scientifically informed metaphysics to no organized attempt at metaphysics, the idea that scientifically informed metaphysics is *the best metaphysics we can have at t* does not establish that it is worth doing. Neither does it rule it out. Rather, an independent case would need to be made for the value of metaphysics—and there are any number of argumentative directions that the proponent of scientifically informed metaphysics could go.

6. Conclusion

My aim was to show that the project of naturalizing metaphysics need not be accompanied by an underlying commitment to scientific realism. On the contrary, the naturalistic programme in metaphysics is compatible with even a strong form of antirealism that commits to the outright falsity of science. I identified two success conditions, narrow and broad, for the project of disentangling naturalized metaphysics from the assumption of realism. The antirealist must explain why, despite the falsity of science, naturalized metaphysics is preferable to non-naturalized metaphysics and to metaphysical quietism. I set out to show that it is possible for the antirealist to meet these conditions. I surveyed a number of conceptions of naturalized metaphysics and its potential payoffs in order to find avenues of argumentation that are open to the antirealist. The results were as follows:

(1) Quinean ontology can have the antirealist-compatible payoff of making large scientific datasets more useful. This pragmatic payoff arguably satisfies the narrow and broad success conditions, because Quinean ontology will achieve this aim more consistently and predictably than non-naturalized metaphysics and because making scientific datasets more useful is pretty clearly worthwhile.

(2) Science-unifying metaphysics can have the antirealist-compatible payoff of increasing the explanatory scope of science, so long as one is willing to countenance a non-factive account of explanation. In terms of explanatory scope, sci-

ence-unifying metaphysics is preferable to non-naturalized metaphysics because it shares the explanatory scope of science; it is preferable to metaphysical quietism if a case can be made for the value of non-factive explanations.

(3) Scientifically informed metaphysics can have the antirealist-compatible payoff of diminished epistemic risk, understood as the ability to pronounce confidently on the truth values of its propositions. The antirealist can argue that scientifically informed metaphysics is preferable to non-naturalized metaphysics from the perspective of epistemic risk because there is more evidence pertaining to its propositions. This payoff did not enable the antirealist to meet the broad success condition due to the inapplicability of the metric of epistemic risk to metaphysical quietism.

(4) Scientifically informed metaphysics can also have the antirealist-compatible payoff of being the best metaphysics available at *t*. Trivially, this would make it preferable to non-naturalized metaphysics; the challenge would be to substantiate the claim. I suggested that the antirealist might argue that scientifically informed metaphysics considers more data and therefore diminishes the influence of subjective priors in conditionalization. To meet the broad success condition and show that the best metaphysics we can have is preferable to metaphysical quietism, an independent case would need to be made for the value of doing metaphysics.

In all but one case (where metaphysical quietism was not risk evaluable), there were argumentative paths to satisfying the narrow and broad success conditions. Some are likely more attractive than others, but I leave those judgments to others. The details clearly need to be worked out in greater detail. My intent here was just to explore some of the antirealist's potential avenues of argumentation. It is telling just how many of them were revealed by such a small survey, holding fixed such a strong antirealist thesis. The avenues I have highlighted are hardly the only ones available: there are more modest varieties of antirealism, other conceptions of naturalized metaphysics, other antirealist-compatible payoffs, other conceptions of those payoffs, and other combinations thereof. Neither should one think that the success conditions must be met by privileging just one payoff or value; most kinds of inquiry will have more than one. The prospects of successfully wedding naturalized metaphysics to antirealism are, therefore, exceptionally promising.

References

Arp, R., Smith, B. and Spear, A. 2015, *Building Ontologies with Basic Formal Ontology*, Cambridge, MA: MIT Press.

Beebee, H. 2018, "The Presidential Address: Philosophical Scepticism and the Aims of Philosophy", *Proceedings of the Aristotelian Society*, 118, 1, 1–24.

Bertrand, M. 2022, "We Need Non-factive Metaphysical Explanation", *Erkenntnis*, 87, 991–1011.

Boyd, R. 1983, "On the Current Status of the Issue of Scientific Realism", *Erkenntnis*, 19, 1/3, 45–90.

Bryant, A. 2021, "Epistemic Infrastructure for a Scientific Metaphysic", *Grazer Philosophische Studien*, 98, 27–49.

Bryant, A. 2020a, "Keep the Chickens Cooped: The Epistemic Inadequacy of Free Range Metaphysics", *Synthese*, 197, 1867–1887.

Bryant, A. 2020b, "Naturalisms", *THINK*, 19, 56, 35–50.

Cartwright, N. 1983, *How the Laws of Physics Lie*, Oxford: Oxford University Press.

Chakravartty, A. 2017, *Scientific Ontology: Integrating Naturalized Metaphysics and Voluntarist Epistemology*, New York: Oxford University Press.

Chakravartty, A. 2013, "On the Prospects of Naturalized Metaphysics", in Ross, D., Ladyman, J. and Kincaid, H. (eds.), *Scientific Metaphysics*, Oxford: Oxford University Press, 27–50.

Chakravartty, A. 2007, *A Metaphysics for Scientific Realism: Knowing* the *Unobservable*, Cambridge: Cambridge University Press.

Cooper, A. and Green, C. 2016, "Embracing the Complexities of 'Big Data' in Archaeology: the Case of the English Landscape and Identities Project on JSTOR", *Journal of Archaeological Method and Theory*, 23, 1, 271–304.

Devitt, M. 1999, "A Naturalistic Defense of Realism", in Hales, S. (ed.), *Metaphysics: Contemporary Readings*, Belmont: Wadsworth Publishing, 90–104.

Devitt, M. 1997, *Realism and Truth*, 2nd Edition, Princeton: Princeton University Press.

Doyle, Y., Egan, S., Graham, N. and Khalifa, K. 2019, "Non-factive Understanding: A Statement and Defense", *Journal for General Philosophy of Science*, 50, 3, 345–365.

Elgin, C. 2017, *True Enough*, Cambridge, MA: MIT Press.

Ereshefsky, M. 2018, "Natural Kinds, Mind Independence, and Defeasibility", *Philosophy of Science*, 85, 5, 845–856.

Ereshefsky, M. 1998, "Species Pluralism and Anti-Realism", *Philosophy of Science*, 65, 1, 103–120.

Esfeld, M. 2009, "Hypothetical Metaphysics of Nature", in Heidelberger, M. and Schiemann, G. (eds.), *The Significance of the Hypothetical in the Natural Sciences*, Berlin: De Gruyter, 341–364.

Guay, A. and Pradeu, T. 2020, "Right Out of the Box: How to Situate Metaphysics of Science in Relation to Other Metaphysical Approaches", *Synthese*, 197, 5, 1847–1866.

Hawley, K. 2018, "Social Science as a Guide to Social Metaphysics?", *Journal for General Philosophy of Science*, 49, 187–198.

Hawley, K. 2006, "Science as a Guide to Metaphysics?", *Synthese*, 149, 3, 451–470.

Hunt, S.D. 2011, "Theory Status, Inductive Realism, and Approximate Truth: No Miracles, No Charades", *International Studies in the Philosophy of Science*, 25, 2, 159–178.

Joyce, J.M. 2011, "The Development of Subjective Bayesianism", in Gabbay, D.M., Hartmann, S., and Woods, J. (eds.), *Handbook of the History of Logic*, Amsterdam: North-Holland, 415–475.

Kant, I. 1783/1977, *Prolegomena to Any Future Metaphysics That Will Be Able to Come Forward as Science*, 2nd Edition, Translated by J. Ellington, Indianapolis: Hackett.

Ladyman, J. and Ross, D. 2013, "The World in the Data", in Ross, D., Ladyman, J. and Kincaid, H. (eds.), *Scientific Metaphysics*, Oxford: Oxford University Press, 108–150.

Ladyman, J., Ross, D., Spurrett, D. and Collier, J. 2007, *Every Thing Must Go: Metaphysics Naturalized*, Oxford: Oxford University Press.

Lean, O. 2021, "Are Bio-Ontologies Metaphysical Theories?", *Synthese*, 199, 11587–11608.

Lipton, P. 2004, *Inference to the Best Explanation*, 2nd Edition, London: Routledge.

Magnus, P.D. and Callender, C. 2004, "Realist Ennui and the Base Rate Fallacy", *Philosophy of Science*, 71, 320–38.

McKenzie, K. 2021, "Science-Guided Metaphysics", in Bliss, R. and Miller, J.T.M. (eds.), *The Routledge Handbook of Metametaphysics*, Abingdon: Routledge, 435–446.

McKenzie, K. 2020, "A Curse on Both Houses: Naturalistic versus A Priori Metaphysics and the Problem of Progress", *Res Philosophica*, Saint Louis University, 97, 1, 1–29.

McSweeney, M. 2023, "Metaphysics as Essentially Imaginative and Aiming at Understanding", *American Philosophical Quarterly*, 60, 1, 83-97.

Papineau, D. 2014, "The Poverty of Conceptual Analysis", in Haug, M. (ed.), *Philosophical Methodology: The Armchair or the* Laboratory, Abingdon: Routledge, 166–194.

Peirce, C.S. (1958), *Collected Papers of Charles Sanders Peirce*, Vol. 8, Edited by Burks, A.W., Cambridge, MA: Harvard University Press.

Potochnik, A., De Regt, H., Elgin, C. and Khalifa, K. 2020, "Idealization and Many Aims", *Philosophy of Science*, University of Chicago Press, 87, 5, 933–943.

Psillos, S. 1999, *Scientific Realism : How Science Tracks Truth*, London: Routledge.

Quine, W.V.O. 1981, *Theories and Things*, Cambridge, MA: Harvard University Press.

Quine, W.V.O. 1960/2013, *Word and Object*, New Edition, Cambridge, MA: MIT Press.

Rosen, G. 2020, "Metaphysics as a Fiction", in Armour-Garb, B. and Kroon, F. (eds.), *Fictionalism in Philosophy*, Oxford: Oxford University Press, 28–47.

Schrenk, M. 2017, *Metaphysics of Science: A Systematic and Historical Introduction*, New York: Routledge.

Sjölin Wirling, Y. 2021, "Non-uniformism and the Epistemology of Philosophically Interesting Modal Claims", *Grazer Philosophische Studien*, 98, 629–656.

Smith, B. 2014, "The Relevance of Philosophical Ontology to Information and Computer Science", in Hagengruber, R. and Riss, U. (eds.), *Philosophy, Computing and Information Science*, London: Pickering and Chatto, 75–83.

Smith, B. 2003, "Ontology", in Floridi, L. (ed.), *Blackwell Guide to Philosophy of Computing and Information*, Oxford: Blackwell, 155–166.

Smith, P. 1998, "Approximate Truth and Dynamical Theories", *British Journal for the Philosophy of Science*, University of Chicago Press, 49, 2, 253–277.

Turri, J. 2018, "The Non-Factive Turn in Epistemology: Some Hypotheses", in Mitova, V. (ed.), *The Factive Turn in Epistemology.* Cambridge: Cambridge University Press, 219-228.

Van Fraassen, B. 1980, *The Scientific Image*, Oxford: Oxford University Press.

Waters, C.K. 2019, "Presidential Address, PSA 2016: An Epistemology of Scientific Practice", *Philosophy of Science*, Cambridge University Press, 86, 4, 585–611.

Waters, C.K. 2018, "Ask Not 'What *Is* an Individual?'", in Bueno, O., Chen, R.L. and Fagan, M.B. (eds.), *Individuation across Experimental and Theoretical Sciences*, New York: Oxford University Press, 91–113.

Waters, C.K. 2017, "No General Structure", in Slater, M. and Yudell, Z. (eds.), *Metaphysics and the Philosophy of Science: New Essays*, Oxford: Oxford University Press, 81–108.

Waters, C.K. 2014, "Shifting Attention From Theory to Practice in Philosophy of Biology", in Galavotti, M.C. et al. (eds.), *New Directions in the Philosophy of Science*, Berlin: Springer International Publishing, 121–39.

Weston, T. 1992, "Approximate Truth and Scientific Realism", *Philosophy of Science*, Cambridge University Press, 59, 1, 53–74.

Weston, T. 1987, "Approximate Truth", *Journal of Philosophical Logic*, 16, 2, 203–227.

Williamson, T. 2013, "What is Naturalism?", in Haug, M. (ed.), *Philosophical Methodology: The Armchair or the Laboratory*, New York: Routledge, 29–31.

Worrall, J. 1989, "Structural Realism: The Best of Both Worlds?", *Dialectica*, 43, 1–2, 99–124.

Between Science and Logic: Securing the Legitimacy of Analytic Metaphysics

Andrea Strollo

University of Trieste

Abstract

Analytic metaphysics has been criticized for its dubious epistemological status. Today, anti-metaphysical sentiments often promote naturalized metaphysics as the only viable way to metaphysical theorizing. In this paper, I argue that analytic metaphysics (or at least a significant portion of it) has the same kind of legitimacy that naturalized metaphysics exhibits. I first point out that naturalized metaphysics is secured by the *de facto* legitimacy of natural science and its continuity with it. Then, I argue that analytic metaphysics can pursue a similar strategy by relying on the *de facto* legitimacy of logic. To achieve this result I propose to interpret analytic metaphysics as philosophy of logic.

Keywords: Analytic metaphysics, Naturalized metaphysics, Meta-metaphysics, Philosophy of logic, Epistemology of metaphysics.

1. Anti-Metaphysics

Skepticism and even aversion to metaphysics is a recurrent theme in philosophy.[1] Especially after the rise of modern science, metaphysics has been frequently frowned upon and dismissed as a relic of the past. Today, however, the relation between science and metaphysics is particularly complex. The reason is that metaphysical issues are connected to and often intertwine with foundational and theoretical problems of contemporary science. Of course, the kind of metaphysics involved in those debates is quite peculiar and distinguished from more traditional forms of metaphysical theorizing. It is an investigation deeply informed by science and developed in continuity with it, rather than a form of mostly *a priori* (or at least armchair)[2] speculation relying on a commonsensical image of reality. As a result, today we have two main strands of metaphysics rivaling

[1] Hume and logical positivism for example.
[2] See Nolan 2015.

Argumenta 10, 1 (2024): 35—53
ISSN 2465-2334

DOI 10.14275/2465-2334/20240.str
First published: dd November 2024

each other: a *naturalized* or *scientific metaphysics* on the one hand,[3] and a more traditional form of *speculative* metaphysics, often called *analytic metaphysics*, on the other hand.[4] Crucially, although analytic metaphysics usually pays lip service to naturalism and claims respect for science, it largely proceeds independently from it (see, for example, Soames 2015). Given such a different engagement with science, it comes as no surprise that, while a tolerant attitude toward naturalized metaphysics is widespread, the analytic approach is undergoing a renewed fire. Accordingly, opponents of metaphysics nowadays mostly target its speculative version, holding that if a metaphysical inquiry can be pursued, it can only be pursued in a naturalized form (Ladyman and Ross 2009). Indeed, also those sympathetic to analytic metaphysics often admit that the discipline looks epistemologically puzzling, as the wide and variegated debates in meta-metaphysics confirm (Wasserman, Manley and Chalmers 2009, Tahko 2015). While it is difficult to precisely define these two kinds of metaphysics, and classification of specific authors can be debatable,[5] the distinction is now customary, especially after the publication of Ladyman and Ross's *Every Thing Must Go*, in which an enthusiastic manifesto of naturalized metaphysics against analytic metaphysics is provided (See Strollo 2017 for systematic criticisms).

In this paper I claim that a skeptical attitude toward analytic metaphysics is misplaced. I argue that at least a significant portion of analytic metaphysics is as unproblematic as naturalized metaphysics and, in some measure, as science itself. In other words, I elaborate a conception of analytic metaphysics that does justice to a significant portion of it as actually practiced and is resistant to skeptical scruples at the same time. The plan of the paper is as follows. In the next section, I show why science is usually taken to have a more solid status and stress the discrepancy with respect to metaphysics. I claim that metaphysics, but not science, faces an apologetic challenge to secure its epistemic legitimacy. In section three, I argue that naturalized metaphysics can actually be justified by its reliance on science, from which it inherits a *de facto* legitimacy. Then, in section four, I consider analytic metaphysics. I propose a strategy to legitimize analytic metaphysics (or at least a significant portion of it) that replicates the relation between naturalized metaphysics and science pointing to the relation between analytic metaphysics and logic. I show that such an idea is naturally suggested by

[3] Represented typically by Ladyman and Ross (2009) and the works in Kincaid, Ladyman, Ross 2013.

[4] I use the label "analytic" metaphysics opposed to "naturalized" metaphysics to adhere to the common practice. However, I should note that naturalized metaphysics derives from the evolution of logical positivism and is characterized by most of the typical features of analytic philosophy, such as its argumentative nature, stress on clarity, reliance on formal tools, naturalism and respect for science, among others. Speaking as there was an opposition with the analytic tradition in philosophy is thus misleading. The expression "neo-scholastic metaphysics" might be a viable option, but, while I do not consider it derogatory (against the apparent intention of the proponents), it would still be quite inappropriate. Scholastic philosophy is primarily characterized by attempts at reconciling Christian faith and reason, a goal that is extraneous to contemporary analytic metaphysics as such. "*A priori"* metaphysics vs naturalized (or scientific) might also be problematic, since analytic metaphysics is not completely *a priori* (see Nolan 2015), and "naturalized" is not necessarily opposed to *a priori*.

[5] For example, Quine advocates a strong naturalization of philosophy and ontology, but his work hardly engages with detailed scientific results.

the history of the return of metaphysics in analytic philosophy and its recent evolution. In section five, I explain how this strategy does safeguard analytic metaphysics meeting the apologetic challenge. In the next section, six, I show that this approach is not revisionary, but it does justice to an actual trend. In section seven, I consider some specifications before concluding the paper.

2. The Descriptive and the Apologetic Challenges

The peculiarity of metaphysics with respect to science becomes fully apparent when the goals of their epistemologies are compared. In the case of science, the main goal of epistemology is explaining what knowledge in a certain scientific field consists of and how it is acquired. Epistemology takes the form of an investigation on a phenomenon that is not in question, namely scientific knowledge. Since scientific knowledge is actual, it is possible. The only question is how. What happens, then, if epistemologists are not able to provide such an account? From the point of view of the specific sciences, not much. Take mathematics. Given that mathematical knowledge is routinely achieved, mathematicians do not need to wait for permissions or indications from their fellow epistemologists. Lack of a suitable epistemology of mathematics may be unpleasant, but the consequences for the working mathematicians are not very serious. Of course, this does not exclude that epistemology might have deep implications, or that certain parts of science could even be criticized with philosophical arguments. The point is rather that such implications can hardly arrive to the point of discrediting the whole or even the majority of a well-established scientific discipline.

For metaphysics the situation is different. While mathematicians and scientists do not need to wait for epistemologists' permission to proceed, metaphysicians would highly benefit from a preemptive reassurance that they can achieve their theoretical goals. In other words, if the existence of scientific knowledge is just a matter of fact demonstrated by the success of science, and witnessed by factors such as the progress of the discipline, the consensus of their practitioners, its predictive power, shared standards, and so on, the possibility of knowledge in metaphysics should be secured by an epistemological defense.[6] Epistemology must show not only *how*, but also, and most importantly, *that* metaphysical knowledge is possible. Let us call the former the *descriptive challenge* ("show how knowledge in metaphysics is acquired"), and the latter the *apologetic challenge* ("show that knowledge in metaphysics can be acquired"). The threat of an apologetic challenge for metaphysics is what marks the epistemological difference with scientific fields of inquiry, and it is what puts metaphysics under fire. If the apologetic challenge was met, the situation would be similar to those of other fields. Providing an epistemology of metaphysics would be a task left to epistemologists, and not being a duty metaphysicians should be particularly worried about.

While the two tasks (accounting for *how* and showing *that* metaphysical knowledge is possible) can be distinguished, it might be thought that only the for-

[6] The same problem may also affect other fields of philosophy to different degrees. For some areas such as philosophy of language or philosophy of mind, however, the apologetic challenge could also be tamed in a way similar to that of naturalized metaphysics, by stressing the relation with contiguous sciences (like linguistics and cognitive neuroscience).

mer is relevant. Showing *how* metaphysical knowledge can be acquired would automatically solve the apologetic problem by showing *that* it can be acquired. This is why many supporters of analytic metaphysics have followed this route. Several strategies have been proposed so far, often pointing in opposite directions such as emphasizing the special role of intuitions or defending an anti-exceptionalist view according to which metaphysics is not a special form of inquiry. Another, tempting and often attempted move is that of stressing some methodological analogies with mathematics, like the reliance on *a priori* arguments.[7] I admit that similar ways of answering the apologetic challenge, if successful, would be effective and convenient. At the same time, however, I should stress that there is another way. As in the case of science, the two challenges can be met separately. For science, the apologetic challenge is neutralized from the beginning with *de facto* considerations, and the descriptive side about the *how* is just left to epistemologists. The case of mathematics is striking, since the absence of a satisfactory descriptive epistemology does not endanger the legitimacy of mathematical knowledge, even though the very existence of mathematical knowledge is deeply perplexing. In the next section, I argue that, similarly to science, the apologetic challenge has hardly any grip on naturalized metaphysics too.

3. The Legitimacy of Naturalized Metaphysics

Unlike analytic metaphysics, naturalized metaphysics is usually considered safe from epistemological and scientifically motivated worries. Although also naturalized metaphysics is sometimes criticised by hardcore empiricists like van Fraassen, its pedigree is not frequently questioned. The main reason, I think, is simple and can be put as follows. As noted in the previous section, whatever the right epistemology of natural science might be, the legitimacy of scientific knowledge is hardly questionable. Thinking otherwise would lead to a radical anti-intellectual stance, which seems at odds not just with metaphysics but with any theoretical enterprise, science included.[8] Given such a privileged status of science, naturalized metaphysics, which itself relies on science, can be easily secured. If science is, *de facto*, epistemologically safe, insofar as naturalized metaphysics is continuous with and possibly relevant to science, then also naturalized metaphysics must be *de facto* possible and legitimate.[9] Naturalized meta-

[7] See Williamson 2007, Baron 2018. Since I also appeal to a seeming mathematical discipline, it is perhaps worth noting that my strategy is different. I should also note that such a strategy is problematic. The legitimacy of mathematics stems from its undeniable success, not from its *a priori* methodology. Mathematics proves to be successful, regardless of, and perhaps even in spite of the methodology it employs. The success of metaphysics, by contrast, is what is precisely in question. Thus, if the example of mathematics might weaken a general methodological objection, it is not enough to secure the legitimacy of every *a priori* approach. Indeed, an armchair methodology might still be blamed as the main culprit of the epistemological bankruptcy of analytic metaphysics. After all, metaphysics does not only deal with merely abstract objects, but also with an external reality seemingly made of concrete entities and perceivable properties out of the range of a purely *a priori* study. Similar considerations also hold if applied mathematics, rather than pure mathematics is considered.

[8] Apart from radical skepticism.

[9] The status of philosophy of science is actually not uncontroversial, as shown by critical remarks of some prominent scientists (including Richard Feynman, Lawrence M.

physics can then obtain an indirect legitimization from its continuity with science. After all, a rough, but not unreasonable, way to view naturalized metaphysics is as an inquiry about what the world must be like if our best scientific theories are true,[10] so that the boundary between naturalized metaphysics and science is hard to trace, if traceable at all. Consequently, the legitimacy of scientific metaphysics stems from science itself.

The thesis of an indirect legitimacy of naturalized metaphysics can be reinforced by considering its history and evolution in the last century. The anti-metaphysical stance strongly supported by logical positivists proved to be hardly sustainable when the project revealed all its weaknesses. As a result, a resurgence of metaphysics slowly took its way in scientific circles themselves, as reported, for example, by James Ladyman (Ladyman 2012). The final outcome of this post-positivist evolution was complete and manifest by 1974, when John Watkins in the speech titled 'Metaphysics and the Advancement of Science', given at The British Society for the Philosophy of Science, claimed that: ''I have the impression that it is now almost universally agreed that metaphysical ideas are important in science as it is that mathematics is'' (Ladyman 2012). Notably, such a progressive rehabilitation of metaphysics in science has little to do with the parallel resurgence that occurred in analytic philosophy in the last decades. While stemming from a common source (namely the demise of logical empiricism) the different historical paths followed by the two kinds of metaphysics help explain the contemporary divide and rivalry between analytic and scientific metaphysicians. Those working in naturalized metaphysics mostly think of themselves as philosophers of science who contribute, more or less directly, to science itself. Naturalized metaphysics is an integral part of (philosophy of) science, confronting problems that are posed by particular scientific theories. To such scholars, analytic metaphysics is a different and alien discipline, originated in another environment with a different purpose and status.[11]

Krauss, Steven Weinberg). Some cautionary remarks on such criticisms, however, are in order. To be worrisome in this context, the attacks should be about the theoretical legitimacy of philosophy of science in its relevant forms, typically exemplified by recent philosophies of particular sciences. These, however, are rarely the target of those remarks (for example, Feynman's alleged claims were probably influenced by historicism and post-positivism, which was on the rise at that time). Moreover, what is often in question is not the theoretical legitimacy of philosophy of science, but its usefulness. Finally, the same authors sometimes venture into philosophy of science themselves, taking explicit positions on philosophical topics, as in Krauss 2013. This makes their voiced rejection of philosophy of science look more verbal than substantial. For a quantitative analysis of the impact of philosophy of science on science (confirming continuity and increasing relevance) see, for instance, Khelfaoui et al. 2021.

[10] The formulation, echoing Quine's view on ontology, is used (in Italian) by Corti & Fano (2020).

[11] When illustrating the story of the resurgence of metaphysics in the context of science, Ladyman presents, among others, the following crucial factors: the continuum between high theory and metaphysics (having to do with the impossibility of adequately specifying a pure observational basis for highly theoretical claims), the explicit engagement with metaphysical issues in science (for example Einstein defending scientific realism with reference to specific metaphysical views), the recognized surplus content of theoretical terms (according to which to explicate the meaning of theoretical terms more than relat-

This story is important for two reasons. First, it shows that it would be hard to delegitimize naturalized metaphysics without putting pressure also on natural science. The idea that naturalized metaphysics inherits a *de facto* legitimization from natural science is thus corroborated. Second, since analytic metaphysics does not directly engage with science and it is extraneous to such a story of reintegration into science, it cannot appeal to the same considerations to secure its epistemological status. Indeed, given its distance from science, analytic metaphysics looks theoretically suspect. Surprisingly as it may sound, however, I intend to secure also the legitimacy of analytic metaphysics and dissolve its apologetic challenge with a *de facto* argument, thereby laying aside the difficult task of providing a descriptively adequate epistemology.[12]

4. The Logic Door to the Resurgence of Analytic Metaphysics

A natural option to obtain a *de facto* justification of analytic metaphysics, different from the one I defend here, might be that of relying on an alleged continuity of analytic metaphysics with naturalized metaphysics. As long as analytic metaphysics is a continuation, at a more abstract level, of naturalized metaphysics, one could suggest that it also inherits the *de facto* legitimacy initially borrowed from science. Naturalized metaphysics would receive its legitimation from science, and then it would pass such a justification on to analytic metaphysics (for example, French and McKenzie 2012, French 2018, Vetter 2018). Although my current proposal does not need to rival this option, I suspect that such a strategy would not be enough. First, since the distance from science would be bigger for analytic metaphysics, the justification would lose strength. Naturalized metaphysics would still appear to be on a firmer foot. Second, analytic metaphysics does not engage with naturalized metaphysics like naturalized metaphysics does with scientific theories. Indeed, while occasional overlapping occurs, explicit engagement seems quite exceptional given the current division between the two communities of metaphysicians.[13] Third, the attitude toward science, from which the original *de facto* justification comes, is crucially different. Naturalized metaphysics is integrated into scientifically well-informed debates, according to the idea that since metaphysics complements science, it can be pursued in a scientific context. Analytic metaphysics, instead, hinges on the possibility of doing metaphysics even independently of science. The idea is that if science does not rule out metaphysics, it can be pursued even outside of a scientific context. Thus, even if analytic metaphysics were strictly continuous with naturalized metaphysics, the link of justification flowing from science seems cut. Given such difficulties, I turn to another strategy, for which analytic metaphysics and naturalized metaphysics are different, independently justified disciplines.

Since it engages with different projects, analytic metaphysics can hardly rely on natural science like naturalized metaphysics does. Nonetheless, a similar

ing observables is required), and holism about confirmation (for which metaphysics is part of the hard core of a research programme).

[12] I should specify that my goal is to secure at least a significant part of analytic metaphysics, not all analytic metaphysics. For ease of exposition, however, I mostly speak of analytic metaphysics in general.

[13] Exceptions are notable (for example, the work of authors like Claudio Calosi and Matteo Morganti, e.g. Calosi and Morganti 2016), but apparently not very widespread.

apologetic strategy can be adopted by replacing natural science with logic, and philosophical logic in particular. Since looking at the historical path is again helpful, I briefly rehearse such a history. I should stress, however, that my interest is not historical. I just want to find inspiration for a theoretical solution to the apologetic challenge by focusing on a particular trajectory of the resurgence of analytic metaphysics. Such a trajectory is a prominent one, but it certainly does not exhaust the complexity of the process.

The main steps of this process can be quickly summarized as follows (For example, Simons 2013). Firstly, it should be noted that while analytic philosophy typically opposed metaphysics in its early stages, the anti-metaphysical attitude was not dominant or universal. The founding fathers of analytic philosophy (Frege, Russell, Moore) all engaged with metaphysical problems and proposed metaphysical solutions, not just linguistic dissolutions, to them. The attitude changed with Wittgenstein, logical positivism and the philosophy of ordinary language. In these strands metaphysical problems were considered pseudo-problems arising from the violation of linguistic constraints. A careful linguistic analysis would have led either to genuine issues treatable by science or to their disappearance. It is from this phase that metaphysics later resurged. However, even during the rise and dominance of the linguistic turn not all analytic philosophers equally opposed metaphysical investigations. Two notable exceptions are found in Poland, with the logic school of Leśniewski and others, and, in the U.S.A. with the work of Gustav Bergmann and Donald Cary Williams. Later, in the '50, the metaphysical turmoil increased. On the one hand, metaphysical investigations became prominent in countries such as Australia, where a number of scholars, most notably David Armstrong, just embraced metaphysics. On the other hand, the work of important philosophers such as Strawson and Quine put an end to the general attitude of opposition to metaphysics. Quine's criticisms of the analytic/synthetic distinction in particular is usually considered as the turning point at which the dogmas of logical positivism became fully obsolete. From this point on, the door was open and analytic metaphysics could thrive again. Its resurgence was finally accelerated by the modal turn derived from development in modal logic, which is the crucial factor I want to focus on.

Although both naturalized and analytic metaphysics sprang from the same source (namely the demise of logical positivism) they soon took diverging paths. Once the tide of the so-called linguistic turn had passed,[14] naturalized metaphysics began its process of reintegration into scientific debates, as already hinted above. By contrast, a crucial factor in the analytic tradition, marking the full return of traditional speculative metaphysics as a central area of philosophical investigation, is notoriously connected with the works that fully established modal logic as a legitimate field of study. Kripke's semantics, together with the pioneering work of several other logicians such as Barcan Marcus and Hintikka, demonstrated that modal reasoning could have been regimented and precisely studied by formal means in a similar way to what classical logic did with respect to mathematical reasoning. The formally rigorous treatment vindicated the intelligibility of several traditional metaphysical notions (such as *de re* modality or even essentialism), on

[14] Note that naturalized metaphysics has been less affected by the influence of the linguistic turn, and, contrary to analytic metaphysics, linguistic analysis and considerations about natural language play no particular role in it.

the face of logical positivism and the last resistance of Quine. The ensuing reflection on modal logic gave rise to works where logic is deeply intertwined with metaphysical issues (consider, for instance, Quine 1953 and the papers in Linsky 1971). Indeed, the metaphysical significance of several questions raised in modal logic became clear and is nowadays standard. Textbook examples include the potential variance of domains in different possible worlds, the related validity of Barcan formulas, the problem of cross-world identity, the status of essentialism, and so on (See textbooks such as Fitting and Mendelsohn 1998 or Girle 2000). On the purely philosophical side, the approach proved extremely fertile, with modal and intensional analysis being applied to many philosophical problems. Such a modal turn had its notorious peak with David Lewis, who eventually put analytic metaphysics back at the center of the philosophical arena. Through the door of modal logic, traditional metaphysics came back.

While modal logic is the most notable and evident case, it is not the only formal study that entered the philosophical scene in the last decades. Another prominent example is formal mereology. Although the study of mereology and its formal versions dates way before the return of analytic metaphysics championed by Lewis, his modern study intensified in more recent times mostly because of his work.[15] Moreover, beside modal logic and mereology, the term 'philosophical logic' today indicates a host of different logics modeling philosophically relevant notions whose study is constantly growing. Easy examples are provided by logics that are syntactically and semantically similar to the systems for alethic modality (and sometimes covered under the term 'modal logic' in a broad sense), such as temporal logic, conditional logic, dynamic logic, deontic logic, and so on. From a historical perspective, the recent return of analytic metaphysics parallels and often interacts with such a development in philosophical logic. Works in the logic field fueled and promoted activity in the metaphysical camp, and formal work itself has often been driven by metaphysical urgencies.

To be historically accurate such a reconstruction should clearly include several details, however, the purpose of this quick historical sketch is just to remind a very familiar story about the correlation between the return of analytic metaphysics and the rise of modal and philosophical logic. Under the light of these historical impressions, a partnership between analytic metaphysics and logic suggests itself. It is to deepen this idea that I now turn.

5. Analytic Metaphysics as Philosophy of Logic

Following the historical suggestion, I claim that analytic metaphysics can obtain its legitimacy by leveraging on a discipline which is arguably as legitimate as natural science: logic. While, *prima facie*, logic can be roughly understood as the

[15] For example, and limiting attention to the last century, Leśniewski 1916, 1927-1931, Goodman and Leonard 1940. For a critical overview of contemporary mereology see Lando 2017. Lando actually argues that mereology is not logic since, for instance, formal principles are not enough to isolate its subject matter and intuitive constraints must be added. Lando nonetheless concedes that mereology exhibits, to some degree, generality and topic-neutrality, which also inspired traditional attempts to demarcate logic. He also concludes that "The formal features of parthood and of other cognate relations and operations are what philosophical mereology is about" (Lando 2017: 29). Overall, this seems to leave at least some room to implement the present strategy.

study of correct (deductive) reasoning,[16] to pursue the present strategy a more precise account is needed. In particular, what is needed is a view of logic meeting at least three constraints. First, it should classify as logic most, and possibly all, of the theories relevant for the project. Second, the account should vindicate the expected epistemological legitimacy of logic, on which analytic metaphysics is to be grounded. Third, to be general enough, it should avoid taking a precise stance on substantial issues in philosophy of logic. Note that the second and third constraints are not in tension as they might appear. The paper moves exactly from the idea of distinguishing descriptive and apologetic challenges, by stressing the *de facto* legitimacy exhibited by extrinsic and social factors, such as the progress of a discipline, the relative consensus among its practitioners, shared standards, and so on. Hence, to meet the last two constraints, it is enough to adopt an account of logic that captures a suitable collection of theories exhibiting a *de facto* legitimacy, revealed by similar factors, regardless of more substantial characterizations. To do that, what counts as logic can be determined by simply deferring to the relevant community of experts, namely logicians. In this sense, 'logic' is what a specific community of scholars recognizes as such by means of certain institutionalized practices.[17] In particular, since their judgment takes a prominent institutionalized form in the publication of specialized journals,[18] we can adopt a practice-based account according to which something counts as logic if it is in the range of such specialized journals, as witnessed by the record of their published papers.[19] It is easy to see that such an approach meets all three constraints above.[20] Indeed, since papers on modal logic, higher order logics, plural logic, and so on have been all routinely published in specialized logic journals, such theories count as logical theories whose epistemic legitimacy is sanctioned by the reliability of the community of its experts.[21, 22]

[16] This rough view of logic is not unproblematic, since, for example, both the normative aspect and the relation with reasoning could be challenged.

[17] Linnebo and Pettigrew, articulating a moderate form of naturalism, hold that "the opinions of scientists working in that discipline can suffice to establish that there exists a justification for some philosophically significant claim…" (Linnebo and Pettigrew 2011). Here the philosophically significant claim is whether a certain theory counts as logic.

[18] Including journals such as the *Journal of Symbolic Logic*, the *Review of Symbolic Logic*, *Annals of Pure and Applied Logic*, the *Journal of Philosophical Logic*, and so on.

[19] Alternatively, a theory might be considered a logic if it concerns the inferential principles governing some arbitrary notion, typically characterized by means of pervasive formalization. This view stems from the idea (Tarski 1983, Varzi 2002) that there is no real demarcation separating logical and non-logical expressions. Although this move would lead to a possibly worrying proliferation of logics (opening the door to disparate systems such as the logic of marriage or hope, as in Pan 2013), their epistemic legitimacy might still be defended in terms of the epistemology of inferential knowledge. This strategy, however, would force precise positions on substantial issues.

[20] The approach can also be intended as a *prima facie*, fallible, strategy that might be eventually replaced by a substantial one, when found. Nevertheless, such a putative account should still match the actual practice of logicians to a good extent.

[21] It could be objected that this account includes too much, since also papers on related topics, such as algebra or category theory, would be dubbed 'logic'. However, it should be kept in mind that what is needed here is not a demarcation that captures the real nature of logic, but one that corresponds to an epistemologically legitimate discipline while including enough theories that are typically subject to metaphysical speculations.

They are *de facto* legitimate, regardless of what the deeper nature of logic is and how its epistemology works.[23]

Once logic is so identified, we can return to analytic metaphysics. If naturalized metaphysics is interpreted as a proper portion of philosophy of science, investigating foundational and interpretational issues such as *what the world must be like if our scientific theories are true*, analytic metaphysics can be interpreted, and secured, in a similar way. It can obtain an indirect *de facto* justification by being interpreted as a portion of philosophy of logic, arguably continuous with logic, investigating foundational and interpretational issues such as *what the world must be like if our logical theories are true*. Note that the continuity with logic should be taken seriously. We have a continuum of various works with pieces more focused on philosophy at one end of the spectrum and others more focused on pure mathematics at the other. Between these two extremes, we have logic more broadly understood, whose precise boundaries with philosophy of logic and pure mathematics are often hard to trace, if traceable at all. Thus, although we might want to distinguish pure philosophy of logic from pure mathematical logic, it would be pretentious to neatly separate philosophy of logic and logic in general. The continuum is particularly clear if issues concerning truth or correctness are considered. Deciding whether, e.g., a certain axiom is true is a task that in many cases pertains to both logic and its philosophy.[24] Distinguishing between the two would be pointless (see also section 7 below on this). In this respect the situation of naturalized metaphysics is different. Although we have a continuum also between naturalized metaphysics and science, experimental testing plays a more significant role in theory choice in science. Hence, philosophy is bound to be more crucial to settle theoretical issues in logic than it is science. These considerations suggest that analytic metaphysics would be better identified with a portion of both philosophy of logic and logic, with only the likely exclusion of purely mathematical logic. It must nonetheless be a portion, because certain topics in philosophy of logic (like epistemological ones) might be outside the scope of metaphysical investigations, and some technical aspects of logic proper may not be of any particular metaphysical relevance. For the sake of simplicity, however, henceforth I speak of 'philosophy of logic' or '(philosophy) of logic' to stress that what is at stake is the part of the logic spectrum lying toward and including its philosophical end.

If such a view of analytic metaphysics is eventually adopted, the following reinterpretations suggest themselves: metaphysics of modality is to be reinterpret as philosophy of modal logic; metaphysics of properties as philosophy of higher

[22] The possible objection that, for example, formal mereology is not logic because it concerns a non-logical predicate would beg the question. What is needed is exactly a demarcation principle establishing what expressions are logical.

[23] Of course this is compatible with the idea that different areas of logic may exhibit *de facto* legitimacy in different degrees. For example, mature areas of research, such as modal logic, are more solid than relatively new fields, such as the logic of ground. Since the latter is not yet fully developed, the factors marking its legitimacy (progress, shared standards, relative consensus, and so on) are not fully established yet.

[24] The connection and continuity of logic with philosophy of logic can also be reinforced by noting the following traits, paralleling the case of naturalized metaphysics and natural science: The problem of content and demarcation of logical terms; The continuum between foundation of logic and its philosophy; The explicit engagement of logic with philosophical issues, and so on.

order and plural logics; metaphysics of identity as philosophy of the logic of identity; metaphysics of parthood as philosophy of (formal) mereology; metaphysics of grounding as philosophy of the logic of ground; metaphysics of dispositions as philosophy of the logic of powers; and so on and so forth.

While the historical connections already suggest that such reinterpretations are natural for a significant amount of contemporary work in analytic metaphysics, let me emphasize how this move can solve the apologetic challenge, before considering possible objections. Suppose that analytic metaphysics is accounted for in terms of a rational investigation of foundational and interpretational issues such as what reality must be like if our logical theories are true. Accordingly, analytic metaphysics would consist in a chapter of philosophy of logic. The apologetic challenge "Show that knowledge in analytic metaphysics can be acquired" becomes: "Show that knowledge in (philosophy of) logic can be acquired". Remember that the apologetic challenge is distinguished from the descriptive one of showing how logical knowledge is acquired. The descriptive challenge for logic is certainly non trivial, but one need not embark in that enterprise to show that logic and its philosophy are legitimate fields. A much simpler and more direct option is available. Indeed, while metaphysics has undergone fierce attacks, philosophy of logic and logic did not suffer any comparable, and perhaps any at all, criticism. Logical positivists themselves did not try to undermine the legitimacy of philosophy of logic, as they even contributed to it (for example, Carnap 1937). Why is philosophy of logic not a critical target like metaphysics? The main reason, I think, is that the legitimacy of logic is hardly questioned, and even hardly questionable. Logic exhibits a *de facto* epistemological legitimacy which is, analogously to natural sciences, revealed by features such as progress, relative consensus, shared standards, and so on.[25] In other words, the legitimacy of logic can be secured with *de facto* arguments. Since logical knowledge is actual, not much else is needed to secure its possibility, exactly as in the case of mathematics, physics or biology. The status of logic could even be reinforced further by pointing to the peculiarity of its specific subject matter broadly understood as correct deductive reasoning. Since deductive reasoning is a key component of every rational inquiry, a dismissal of logic seems mostly viable to radical skepticism. [26]

What about philosophy of logic rather than logic, though? The situation here is similar to that of philosophy of science and naturalized metaphysics. Once logical knowledge is secured, also philosophy of logic enjoys an indirect legitimization. As long as logic is legitimate, rational reflection on it must be legitimate too. Questioning the legitimacy of well conducted forms of philosophy of logic would put logic itself at risk. Indeed, several prominent figures in the history of logic have worked at the boundary of logic and philosophy, proving the continuity between the "two" camps. While today the intellectual division of labor between philosophers of logic and purely mathematical logicians may be deeper than in the years of Frege, Russell, or Brouwer, probably also as a result of modern hyper specialization, it would be hard to reject the legitimacy of phi-

[25] Even perhaps predictive power, as argued in Hjortland and Martin 2021.
[26] For simplicity I put logical nihilism aside, although the general point could be reframed in terms of validity of single inferences, which also the logical nihilist must accept. On logical nihilism see, for instance, Russell 2018, Cotnoir 2019. Dicher 2021 for criticism.

losophy of logic without also rejecting many logical projects stemming from it (think of intuitionistic logic, relevant logic, the study of paradoxes, or, more recently, the logic of ground).

The nice consequences of this situation for analytic metaphysics are straightforward. If philosophy of logic obtains legitimization from its strict relation and continuity with logic, and analytic metaphysics is interpreted as a portion of philosophy of logic, then it enjoys the same justification. By interpreting analytic metaphysics as (philosophy of) logic, the apologetic challenge is again met with *de facto* considerations. Like any other science, analytic metaphysicians can proceed in their research without waiting for epistemologists permission.

At this point it is worth noting that once analytic metaphysics is reduced to philosophy of logic, also its descriptive epistemology becomes parasitic of that of philosophy of logic. Understanding how metaphysical knowledge is obtained requires understanding how logical knowledge is obtained. The situation is similar for naturalized metaphysics, which, being continuous and subsidiary to natural science, also demands an epistemological account of science itself.[27] While the exact nature of such epistemologies is not important here (since the strategy is a *de facto* one), two remarks are worth making. One is that, so reframed, several potential objections to analytic metaphysics fade away. For example, the alleged problematic reliance on intuitions in analytic metaphysics becomes potentially harmless once viewed in terms of the role of intuitions about logic. Indeed, it might also turn out that logic does not require any special resort to *a priori* intuitions at all. According to logical anti-exceptionalism, "logic isn't special. Its theories are continuous with science; its method continuous with scientific method" (Hjortland 2017). If so, both analytic metaphysics and naturalized metaphysics would deal with sciences, although different ones. The second remark is that, according to the approach advocated in this paper, a host of metaphysical alternative views would present themselves in slightly different clothes. For example, the opposition between metaphysical realism and antirealism would be rephrased as realism or antirealism about logic. Accordingly, metaphysical disputes would not be lost but just reformulated as analogous disputes about logic.[28]

Before showing that the identification of analytic metaphysics and philosophy of logic is not just convenient but descriptively right, let me dispel some basic objections that could be moved against the viability of the suggested strategy. First of all, it could be objected that characterizing analytic metaphysics as investigating *what the world must be like if our logical theories are true* hardly makes sense, since metaphysics and natural science describes the world, but logic does not. Given such a discrepancy, it is helpless to try to get metaphysics out of logic. This objection, however, is easily neutralized. First, the idea that natural science is about the world is questionable, as shown by antirealist and instrumen-

[27] Note that since I am identifying analytic metaphysics with philosophy of logic, not just with logic, there is room for different epistemologies. Similarly, the epistemology of naturalized metaphysics is strictly related, but not necessarily identical to that of natural science. The issue is also complicated by the problem of how philosophy should be distinguished from other disciplines.

[28] That many options remain open is also a consequence of the fact that logic has been identified in practice-based terms, rather than by adopting a particular conception of the nature of logic.

talist conceptions. Moreover, scientific theories do not always (and sometimes hardly) wear their interpretations and connection with the manifest world on their sleeves, so that even if a scientific theory is taken to describe the world, the worldly picture emerging from it is often underdetermined (as the case of quantum mechanics demonstrates).[29] Second, also the claim that metaphysics is about the world is questionable. Metaphysics may be about our conceptual schemes rather than about an independent reality. Although such a view is probably not dominant nowadays, it is a possible conception nonetheless and it was supported, for example, by the early linguistic meta-philosophical views. Third, the claim that logic is not about the world is equally contentious. Several authors (such as Maddy or Sher)[30] explicitly disagree, and various forms of logical realism are frequently discussed (Sider 2011, McSweeney 2019, Tahko 2021). Surely, the metaphysical picture emerging from a logical theory (for example from the modal system S5) is often severely underdetermined (so that, for example, the choice between modal realism or modal fictionalism might not be simply dictated by the formalism). But, as remarked, this may be the case for scientific theories as well.

Another objection might stem from the fact that many different logics are available. For example, one can construct a plural logic and one can construct a second order logic, but how does this tell us what the world is like (at least with respect to properties)? The reply, however, is simple. Provided that the two logics have been saddled with a metaphysical interpretation, to answer the question we must decide what logic, if any, is the correct one. That there are many logical theories available does not immediately imply that all such logics are correct.[31] To decide whether reality is accounted for by the metaphysical picture delivered by plural logic or by the one delivered by second order logic we must decide which logic is right. Of course, theory choice is not easy, and determining what logic is correct is a complex and difficult task, but the proposal was never intended to make analytic metaphysics easy.[32]

6. 'Analytic Metaphysics as Philosophy of Logic' in Action

If analytic metaphysics is interpreted as a form of philosophy of logic, the apologetic challenge is met. This is already a strong reason to promote such an identification. But does the proposed strategy advocate a revisionary conception of metaphysics, or does it do justice to how analytic metaphysics is actually conducted? In this section I give some evidence suggesting that the proposed view is not only convenient but also descriptively adequate to a good extent. In particular, I show

[29] One could insist that if metaphysics is about the world and logic is about conceptual schemes, we do have a separation. In this case, however, on the one hand, metaphysics would be under the pressure of competing with naturalized metaphysics. On the other hand, philosophy of logic would still provide an alternative conception of metaphysics as mostly conceptual.

[30] Sher 1991, Maddy 2002.

[31] This is a standard specification, for example, in the debate on logical pluralism.

[32] The assessment will probably also involve metaphysical considerations. For example, a nominalist could criticize second order logic because it arguably supports a metaphysics of properties. Note that this interplay is vindicated by the present proposal and it is not problematic. Logic is intended to precede analytic metaphysics only in the epistemic order of justification, not under every respect.

that the account naturally aligns with a current and growing trend in analytic metaphysics, so that the present proposal just takes such a practice seriously.

First of all, I should call attention to a number of general similarities between analytic metaphysics and (philosophy of) logic, pointing toward a natural convergence of the two. However, for reasons of space, and since I already presented them elsewhere, I just quickly mention them to reinforce the overall appeal of such an identification (Strollo 2018). Similarities include the ambition to absolute generality, the apparent recalcitrance to empirical data, the role of linguistic competence and common sense as sources of evidence, the role of paradoxes, the role of language and reasoning, and the mutual correspondence between several meta-theoretical disputes (such as the possibly merely verbal nature of disagreement: see Hirsch 2010). Since such traits are hardly so systematically shared with other fields, logic and analytic metaphysics present themselves like disciplines with a similar and peculiar profile. But there is more.

As already noted, today many metaphysical issues are paired with corresponding philosophical logics. The divide between metaphysics and philosophy of logic, for example, fades away in many works on contemporary mereology. A quick look at the papers collected in Baxter and Cotnoir's *Composition as Identity* provide several instances of this approach (Baxter and Cotnoir 2014). Would it be unreasonable to consider Turner's paper[33] (just to randomly pick one) as a piece of philosophy of logic, and philosophy of formal mereology in particular? Hardly so. Indeed, this seems a natural way of presenting its content. At the bare minimum, metaphysics and philosophy of logic overlap there. Or take the recent interest in fundamentality. Research in the logic of ground directly stems from and intertwines with metaphysical issues. Again, in such cases it would be pointless to tell discussions on the philosophy of the logic of ground apart from discussions on the metaphysics of grounding. Take Fine's "The pure logic of ground", deRosset's "On weak ground" or Poggiolesi's "On defining the notion of complete and immediate formal grounding" (Fine 2012, deRosset 2014, Poggiolesi 2018). Discussing whether they should count as papers in the logical camp rather than the metaphysical field is pointless. Apart from the superficial feature of how many formulas a paper may host,[34] they are both logically and metaphysically relevant at the same time. Similar cases could be proposed for many other topics such as the possibility of absolute generality or plural quantification (Torza 2015, Florio and Linnebo 2021). If an objector, complaining about the lack of systematicity of the above examples, raised the concern that they could be the mere result of cherry picking, it should be clear that the high number of pickable cherries supports the present thesis nonetheless.

There is, however, even more than this widespread alignment, frequent overlapping and interaction. The methodology of merging metaphysics and logic together has been explicitly adopted by prominent philosophers. Direct support for the identification of analytic metaphysics with (philosophy of) logic is indeed manifest in some recent works. The clearest and most obvious case is

[33] Turner 2014 discusses a formal regimentation of Baxter's view of identity where Leibniz' law is dropped.

[34] Similarly, the philosopher of physics David Albert submits a paper to a physics journal rather than to a philosophy journal if the paper contains more than two equations (https://www.youtube.com/watch?v=UNpLfXOfzZ8&ab_channel=BigThink, min 3.53).

Williamson's *Modal Logic as Metaphysics*, where, already in the title, Williamson is upfront in the kind of project he engages in (see Williamson 2013 already in the preface). But the same methodology is also adopted in other works, for example those about higher order logic and the metaphysics of properties, like: Bacon, Hawthorne and Uzquiano, "Higher-order free logic and the Prior-Kaplan paradox"; Fritz and Goodman "Higher order Contingentism Part 1"; or Trueman, *Properties and Propositions, the metaphysics of Higher order logic.*[35]

Given such a scenario, I suggest that the development of philosophical logic, paralleling and often intertwining with metaphysical debates, is now crystalizing in a specific methodology which relies more and more on logical methods. The idea that analytic metaphysics is a form of philosophy of logic naturally emerges from this growing trend. Hence, even independently from the epistemological merits I already emphasized, the proposal presents itself as descriptively correct to some extent, fitting a widespread contemporary practice. That a significant portion of analytic metaphysics is conducted as a form of philosophy of logic is, first of all, a fact that should be registered. The proposed interpretation is thus not intended to be revisionary, but to take on board a trend that already exists and independently grows in contemporary analytic metaphysics. Therefore, in some of its prominent contemporary forms, analytic metaphysics is already epistemically unproblematic.

7. Limits and Specifications of the Proposal

Assume that my proposal works and analytic metaphysics is reinterpreted as (philosophy of) logic. Is such a view able to vindicate *all* analytic metaphysics? I have no ambition to answer 'yes' to this question. Before discussing potentially recalcitrant cases, however, it is important to say something about the role of formalization and mathematical systems in the present view. Although philosophy of logic can be, and typically is conducted after a formal system is fully developed, philosophical considerations are often crucial both to prepare the ground for and while a formal theory is being elaborated.[36] At the same time, although formalization is important and valuable, informal philosophy concerning notions displayed in informal reasoning is philosophy of logic enough. It thus follows that one should not object to my proposal by pointing to pieces of metaphysical speculation that do not explicitly rely or engage with formal systems.

Even with such specifications in force, the view of metaphysics as philosophy of logic seems unable to do justice to all analytic metaphysics. Take, for example, the debate on the nature of time or the one on the metaphysics of artifacts (see Carrara and Olivero 2021 for a critical overview). In what sense are such debates disputes in philosophy of logic? Hence, one could object that there are important parts of contemporary analytic metaphysics that are neglected by the present proposal. My basic reply to that is: "yes, but...". *Yes*, I admit that the proposal might have a limited range and be unable to do justice to all metaphysics. *But,* at the same time, the portion it vindicates is significant nonetheless. Indeed, since the ambition to vindicate all metaphysics in one move would be too high a task, if the proposal fits at least a significant part, it retains much of its

[35] Fritz and Goodman 2016; Bacon, Hawthorne, Uzquiano 2016; Trueman 2020.
[36] Of course a merely mathematical study might be motivated independently of consideration about truth.

value. Moreover, two additional remarks are relevant. First, also in recalcitrant cases some space for logical considerations is available. For example, some issues about time can be recast in logical terms by means of temporal logic, and some problems about artifacts can be connected to the logic of identity.[37] Second, the interpretation of metaphysics as philosophy of logic is not the only strategy able to tame epistemological worries. Naturalized metaphysics is another option. The logical proposal put forward here is not intended to replace naturalized metaphysics but to team it. Accordingly, the overall metaphysics of time could be considered as the result of integrating the naturalized metaphysics of time with reflections on the philosophy of temporal logics. A similar labor division in metaphysics between notions more or less apt to a logical treatment is again mirrored in actual practice. Discussions on time, causation, and natural laws, invite, if not require serious engagement with natural science and lead the metaphysician under the realm of naturalized metaphysics. By contrast, traditional discussions on notions such as identity, grounding, parthood, properties, modality, seem in principle immune to empirical results and lead the metaphysician to logical regimentations (see Bryant 2020).

Possibly, even once combined with naturalized metaphysics, not all analytic metaphysical inquiries would be covered, so that other approaches might be needed. However, even in this case, a large amount of metaphysical work would have been already secured. Indeed, it might also be suggested that metaphysical reflections escaping logical and scientific treatments are just the kind of general and philosophical reflections that must struggle in unexplored territories, where epistemological safety could never be forthcoming. That is where metaphysics fades into general philosophical speculation. The fertility of such epistemically risky philosophical inquiries is a topic for another discussion, but such theorizing is often a necessary prerequisite to develop firmer studies. Such debates are the preliminary steps to eventually develop specific sciences or logics with their associated, and epistemologically safe, metaphysical sides.

8. Conclusion

In this paper I argued that the epistemological legitimacy of a significant portion of analytic metaphysics can be provided by interpreting it as (philosophy of) logic. Such an identification allows an indirect *de facto* justification, similar to that of other well established fields of inquiry. In particular, the status of analytic metaphysics becomes similar to that of its rival: naturalized metaphysics. Notably, such a conception vindicates a recent growing trend in analytic metaphysics, where metaphysics is actually conducted as (philosophy of) logic. As currently practiced, analytic metaphysics is in large part already safe. Analytic metaphysicians should then continue their work without worrying about defending the intellectual legitimacy of their study.

A particular side benefit of this proposal is that it tames the rivalry between analytic and naturalized metaphysics. The two metaphysical approaches can be taken to compete in addressing the same questions only in a very general and vague sense, since they actually focus on different notions calling for different

[37] For example, the question of whether the future is determined, could be reformulated as the issue of whether excluded middle holds for future events. For artifacts and identity see, for example, Carrara 2009.

methodologies.[38] That no opposition is really there could even be made explicit by speaking directly of philosophy of science and philosophy of logic, instead of using the vexed term 'metaphysics'. Probably, even the fiercest opponent of analytic metaphysics does not raise an eyebrow if a metaphysical paper is presented as a work in philosophy of logic. Once analytic metaphysics is labeled as 'philosophy of logic' scruples against it seem to vanish. While I do not suggest dropping the term, current aversion to 'metaphysics' might be more the result of old and outdated biases triggered by a word, rather than an authentic opposition to the actual contemporary practice.

References

Bacon, A., Hawthorne, J. and Uzquiano, G. 2016, "Higher-Order Free Logic and the Prior-Kaplan Paradox", *Canadian Journal of Philosophy*, 46, 4-5, 493-541.

Baron, S. 2018, "A Formal Apology for Metaphysics", *Ergo*, 5, 39, 1030-1060.

Baxter, D.L.M. and Cotnoir, A.J. (eds.) 2014, "Composition as Identity", Oxford: Oxford University Press.

Bryant, A. 2020, "Keep the Chickens Cooped: The Epistemic Inadequacy of Free Range Metaphysics", *Synthese*, 197, 5, 1867-1887.

Calosi, C. and Morganti, M. 2016, "Humean Supervenience, Composition as Identity and Quantum Wholes", *Erkenntnis*, 81, 6, 1173-1194.

Carnap, R. 1937, *The Logical Syntax of Language*, London: Kegan Paul.

Carrara, M. and Olivero, I. 2021, "On the Semantics of Artifactual Kind Terms", *Philosophy Compass*, e12778, https://doi.org/10.1111/phc3.12778

Carrara, M. 2009, "Relative Identity and the Number of Artifacts", *Techné: Research in Philosophy and Technology*, 13, 2, 108-122.

Corti, A. and Fano, V. 2020, "La Metafisica è Morta! : Lunga Vita alla Metafisica!", *Rivista di Filosofia Neo-Scolastica*, 911-941.

Cotnoir, A. 2019, "Logical Nihilism", in, Kellen, N., Pedersen, N.J.L.L. and Wyatt, J. (eds.), *Pluralisms in truth and logic*, Cham: Palgrave Macmillan, 301-329.

deRosset, L. 2014, "On Weak Ground", *Review of Symbolic Logic*, 7, 4, 713-744.

Dicher, B. 2021, "Requiem for Logical Nihilism, or: Logical Nihilism Annihilated", *Synthese*, 198, 8, 7073-7096.

Fine, K. 2012, "The Pure Logic of Ground", *Review of Symbolic Logic*, 5, 1, 1-25.

Fitting, M.C. and Mendelsohn, R.L. 1998, *First-Order Modal Logic*, Dordrecht: Kluwer.

Florio, S. and Linnebo, Ø. 2021, *The Many and the One: A Philosophical Study of Plural Logic*, Oxford: Oxford University Press.

French, S. and McKenzie, K. 2012, "Thinking Outside the Toolbox: Towards a More Productive Engagement Between Metaphysics and Philosophy of Physics", *European Journal of Analytic Philosophy*, 8, 1, 42–59.

French, S. 2018, "Toying with the Toolbox: How Metaphysics Can Still Make a Contribution", *J Gen Philos Sci*, 49, 211–230

[38] See Paul 2012 for a defense of the idea that the methods are basically the same.

Fritz, P. and Goodman, J. 2016, "Higher-Order Contingentism, Part 1: Closure and Generation", *Journal of Philosophical Logic*, 45, 6, 645-695.

Girle, R. 2000, *Modal Logics and Philosophy*, Mcgill-Queen's University Press.

Goodman, N. and Leonard, H. S. 1940, "The Calculus of Individuals and its Uses", *Journal of Symbolic Logic*, 5, 2, 45-55.

Hirsch, E. 2010, *Quantifier Variance and Realism: Essays in Metaontology*, Oxford: Oxford University Press.

Hjortland, O. 2017, "Anti-Exceptionalism about Logic", *Philosophical Studies*, 174, 631–658.

Hjortland, O. and Martin, B. 2021, "Logical Predictivism", *Journal of Philosophical Logic*, 50, 2, 285-318.

Khelfaoui, M., Gingras, Y., Lemoine, M. and Pradeu T. 2021, "The Visibility of Philosophy of Science in the Sciences, 1980–2018", *Synthese*, 199, 3-4, 1-31.

Kincaid, H., Ladyman, J. and Ross, D. (eds.) 2013, "Scientific Metaphysics", Oxford: Oxford University Press.

Krauss, L.M. 2013, *A Universe from Nothing. Why there is something rather than nothing*, Free Press.

Ladyman, J. 2012, "Science, Metaphysics and Method" *Philosophical Studies*, 160, 1, 31-51.

Ladyman, J. and Ross, D. 2009, *Every Thing Must Go: Metaphysics Naturalized*, Oxford: Oxford University Press.

Lando, G. 2017, *Mereology: A Philosophical Introduction*, London: Bloomsbury.

Leśniewski, S. 1916, "Podstawy Ogólnej Teorii Mnogości I", [Foundations of the General Theory of Sets, I], *Prace Polskiego Koła Naukowego w Moskwie, Sekcya matematyczno-przyrodnicza*, No.2., Moscow: Popławski.

Leśniewski, S. 1927-1931, O podstawach matematyki [On the Foundations of Mathematics], I–V.*Przegląd Filozoficzny*, 30 (1927), 164–206; 31 (1928), 261–291; 32 (1929), 60–101; 33 (1930), 77–105; 34 (1931), 142–170 (1927–1931).

Linnebo, Ø. and Pettigrew, R. 2011, "Category Theory as an Autonomous Foundation", *Philosophia Mathematica*, 19, 3, 227–254.

Linsky, L. 1971, *Reference and Modality*, London: Oxford University Press.

Paul, L.A. 2012, "Metaphysics as Modeling: The Handmaiden's Tale", *Philosophical Studies*, 160, 1, 1-29.

Maddy, P. 2002, "A Naturalistic Look at Logic", *Proceedings and Addresses of the American Philosophical Association*, 76, 2, 61-90.

McSweeney, M.M. 2019, "Logical Realism and the Metaphysics of Logic", *Philosophy Compass*, 14, 1, e12563.

Nolan, D. 2015, "The A Posteriori Armchair", *Australasian Journal of Philosophy*, 93, 2, 211-231.

Pan, T. 2013, "The Logical Structure of Hope", *Croatian Journal of Philosophy*, 13, 3, 457-462.

Poggiolesi, F. 2018, "On Constructing a Logic for the Notion of Complete and Immediate Formal Grounding", *Synthese*, 195, 3, 1231-1254.

Quine, W.V.O. 1953, "Three Grades of Modal Involvement", *Proceedings of the XIth International Congress of Philosophy, Vol. 11*, Amsterdam: North-Holland, 65–81.

Russell, G. 2018, "Logical Nihilism: Could There Be No Logic?", *Philosophical Issues*, 28, 1, 308-324.

Sher, G. 1991, *The Bounds of Logic: A Generalized Viewpoint*, Cambridge, MA: MIT Press.

Sider, T. 2011, *Writing the Book of the World*, Oxford: Oxford University Press.

Simons, P. 2013, "Metaphysics in Analytic Philosophy", in *The Oxford Handbook of the History of Analytic Philosophy*, Oxford: Oxford University Press, 709-728.

Soames, S. 2015, "David Lewis's Place in Analytic Philosophy", in Loewer, B. and Schaffer, J. (eds.), *A Companion to David Lewis*, John Wiley & Sons, 80-98.

Strollo, A. 2017, "Analytic Metaphysics Should Not Go", *Philosophical Inquiries*, 5, 2, 33-53, ISSN (print) 2281-8618-ETS, ISSN (on-line): 2282-0248.

Strollo, A. 2018, "Metaphysics as Logic", *Rivista di Estetica*, n.s., n. 69, LVII, 7-20, ISSN: 0035-6212.

Tahko, T. 2015, *An Introduction to Metametaphysics*, Cambridge: Cambridge University Press.

Tahko, T. 2021, "A Survey of Logical Realism", *Synthese*, 198, 5, 4775-4790.

Tarski, A., 1983, "On the Concept of Logical Consequence", in J. Corcoran (ed.), *Logic, Semantics and Metamathematics, 2nd edition*, Hackett: Indianapolis, 409-420.

Torza, A. (ed.) 2015, "Quantifiers, Quantifiers, and Quantifiers: Themes in Logic, Metaphysics, and Language", Synthese Library, Vol. 373), Springer.

Trueman, R. 2020, *Properties and Propositions: The Metaphysics of Higher-Order Logic*, Cambridge: Cambridge University Press.

Turner, J. 2014, "Donald Baxter's Composition as Identity", in Baxter, D. and Cotnoir, A. (eds.), *Composition as Identity*, Oxford: Oxford University Press.

Varzi, A. C. 2002, "On Logical Relativity", *Philosophical Issues*, 12, 197–219.

Vetter, B. 2018, "Digging Deeper: Why Metaphysics is More Than a Toolbox", *J Gen Philos Sci*, 49, 231–241, https://doi.org/10.1007/s10838-017-9387-7

Wasserman, R., Manley, D. and Chalmers, D. (eds.) 2009, "Metametaphysics: New Essays on the Foundations of Ontology", Oxford: Oxford University Press.

Williamson, T. 2007, *The Philosophy of Philosophy*, Wiley-Blackwell.

Williamson, T. 2013, *Modal Logic as Metaphysics*, Oxford: Oxford University Press.

Metaphysics as a Science: A Sketch of an Overview

Lauri Snellman

University of Helsinki

Abstract

This article sketches a pragmatist method for metaphysics. Bottom-up or descriptive metaphysics describes the domains of quantification, essences and the categories of a linguistic activity by describing the linguistic activities of encountering reality and seeking and finding objects and relationships. Constructive or top-down metaphysics constructs alternative conceptual schemes, which can be used as world-view backgrounds to construct scientific paradigms and theories. Metaphysical theories are then assessed by comparing the research traditions that arise when the theories are used as conceptual schemes. The pragmatic circle can be generalized into a world-view circle of forming a conceptual scheme, articulating the scheme and drawing interpretations, and assessing and modifying the world-view. Different metaphysical conceptual schemes can be contrasted through a dialogue between languages, which allows a comparison of how different metaphysical frameworks can recognize reality and offer good models for being qua being.

Keywords: Metaphysics, Language-games, World-views, Pragmatism, Conceptual schemes.

1. Introduction

Metaphysics has been questioned since the 18th century Enlightenment and its foundational projects (see, e.g., KrV, H). Similar questions about the scientific status of metaphysics have been raised in recent debates (see Ladyman 2007, Morganti and Tahko 2017, Snellman 2023). This article offers a sketch of metaphysical methodology by building connections between language-games, quantification, world-views and frameworks for scientific research. These connections then offer an approach that leads to bottom-up descriptive metaphysics and constructive or top-down metaphysics as framework construction. Different metaphysical systems are connected with Kuhnian world-views or frameworks, which are then compared dialectically.

Argumenta 10, 1 (2024): 55—70
ISSN 2465-2334

DOI 10.14275/2465-2334/20240.sne
First published: dd November 2024

Metaphysics can proceed bottom-up as a description of quantification and its categories. Hans-Johann Glock (2012) has described four approaches to metaphysics in the late 20th century: V.W.O. Quine's (1953) description of the domains of quantification, and P.F. Strawson's (1959) descriptive metaphysics, direct-reference essentialism and truthmaker theory. I argue that descriptive metaphysics can offer a metatheory for quantification, as the concept of being ("there is") is located in language-games of seeking and finding (see Hintikka 1973: EP 2, Snellman 2023). Language-games also function as categories, as they offer the possibilities of description and reidentification of their objects and hence their typical properties and essences (Garver 1994). The description of language-games and their activities of seeking and finding can then categorize the entities in their domains metaphysically. Moreover, this does not reduce metaphysics to an intralinguistic activity, as our linguistic activities build on the facts and relationships of the world (see Dickson 1995).

Metaphysics can also function in a top-down manner by offering alternative world-views and conceptual schemes to interpret experience and to help us deal with reality. The concept of conceptual schemes builds on Thomas Kuhn's (1969) and Ludwig Wittgenstein's (OC) work. A world-view offers both conceptual rules for assessing arguments and also research and experimentation practices for looking at the world from a particular angle. This also entails that all experience is theory-laden and there is no theory-neutral way of characterizing observations. These interpretations of experience by looking at the world from a given angle then lead to Gestalt-perception (see PI part 2, xi, Snellman 2023). A metaphysical system like atomism or Aristotelianism can then provide a world-view for defining a conceptual system to interpret experience and to guide research practices. Tuomas Tahko and Matteo Morganti (2017) offer an account of empirically testing metaphysics. Metaphysicians first articulate a general conceptual scheme or a metaphysical theory. The theory has abstract terms like categories, properties, causation or substances, and is then applied as a world-view level metatheory for formulating research programs or paradigms. The theories of these paradigms and research programs are then tested against the empirical results. This amounts to an indirect test of the metaphysical theory as well. A metaphysical theory then functions as a general conceptual scheme or a world-view, which is operationalized through paradigms. Paradigms then lead to theories, which give us models for interpreting phenomena, recognizing their underlying realities and seeing the phenomena as something. These levels of interpretation can be contrasted with levels of strategies of action: policy, operationalization, campaign strategies and tactics (see Ackerman and Kruegler 1994).

Tahko and Morganti's model however involves contrasting world-views and their languages. Incommensurable languages can be compared: the metaphysical circle of metaphysical theory → articulation and use as a background for science → testing of theories and the related circle of world-view → research problems → anomalies can be seen as generalizations of C.S. Peirce's (EP 1, 186-209) pragmatic circle abduction → deduction → induction. J.G. Hamann (N I, 29-31) also points out that the comparison of world-views requires contrasts between different conceptual schemes like Leibnizianism, Newtonianism and Cartesianism, leads to comparisons between their interpretative approaches as well as an account of their empirical results. This however involves a dialogue between differing conceptual systems, and testing a world-view can be approached through dialectical conversations between world-views. These dialogues can show that one system is better than another if its ways of encountering the world cannot solve its own

anomalies, recognize the successes of its competitors, or lead to obfuscation about a reality (Taylor 1995, MacIntyre 1988).

2. Language-Games and the Categories of Quantification

Metaphysics is a science of being qua being, or the attributes that concern being as such (Simons 1995). The concept of being is however a linguistically mediated concept in our language, and an approach to the logic or properties of being has to be approached through an account of the linguistic activities for the concept and the relationships in which the concept is embedded (ZH 7, 161-183). This claim needs some unpacking. First, all concepts are constituted by language-use: linguistic mediation is understood here in a fairly strong sense. From this it follows that giving an account of the concept of being (or being qua being) requires describing the contexts of use for the term "exist". C.S. Peirce (EP 2, 168) and Jaakko Hintikka (1973) have identified the activities of seeking and finding as the background of the concept of being.

The concept of being can then be approached through the study of language-games. Wittgenstein uses the concept of a language-game to emphasize that language is used as a part of an activity. He gives examples like asking questions and giving answers, and receiving battle reports and issuing orders. Language-use thus involves both words and activities. It takes place in the world and includes the world's systems. The meaning of an expression is then its use in these linguistic activities (PI 1-42, esp. 19, 23, 42). The metaphor of games for language also emphasizes that language-use is structured by rules that form its structure. Both the use of the words "let's play a game of chess" and the game of chess have rules, and the activities of playing chess connect the two sets of rules (PI 197; Snellman 2023; Glock 1997: 150-155, 193-198).

Wittgenstein's philosophical method is the grammatical description of the rules and practices of language-use. Newton Garver (1994: 217-235) argues that Wittgenstein's concept of grammar generalizes the linguistic concept of a grammar. Linguistic grammar concerns the rules of use for syntactic elements like letters and expressions, but philosophical grammar concerns the use of speech-acts, or actions where language is used. We can also develop philosophical grammar by using Hamann's distinction between elements and institutions. Expressions like "Pepsi" and "Let's play a game of chess!" are the elements of language. They have a role in a language-game or a linguistic practice, and one draws distinctions between them by distinguishing what possibilities of use or discourse-possibilities they offer. To use Garver's example, Arabic does not distinguish between the sounds "Pepsi" and "Bepsi", but "Bepsi" is ungrammatical in English because English spelling distinguishes B and P. The rules of a language-game are the institutions of the language. They are the social, linguistic and logical patterns of repeated use that determine whether an expression makes sense and how expressions are used in communicative roles and to attain communicative goals (see Snellman 2023: Ch. 4.1; Glock 1997: 193-198).

Language-games then form systems that are composed of the elements of objects and speech-acts, and the institutions of rules. Hamann (ZH 7, 169-170, see also Mainzer 2004) argues that systems analysis must distinguish between the elements while tracing the relationships and institutions interrelating them. The analysis moreover reveals the laws of language-use, and its underlying practices and realities contained in the language-game. A descriptive metaphysics of quantification can

then be given by examining the language-games for using the words "there is", "all", "some" and "none". Wittgenstein locates the rules for the expression "there is" in the activities of encountering reality and interacting with objects:

> Children do not learn that books exist, that armchairs exist, etc. etc.,—they learn to fetch books, sit in armchairs, etc. etc.
> Later, questions about the existence of things do of course arise. "Is there such a thing as a unicorn?" and so on. But such a question is possible only because as a rule no corresponding question presents itself. For how does one know how to set about satisfying oneself of the existence of unicorns? How did one learn the method for determining whether something exists or not? (OC 476).

The rules of everyday language-games then give a meaning for the terms "Books exist" and "Armchairs exist", because one can encounter a book by taking hold of it and an armchair by sitting in it. The bodily practices and criteria for encountering an object then give a meaning to the expression "there is", or ∃. C.S. Peirce and Jaakko Hintikka elaborate on this by developing game-theoretic accounts of these language-games for seeking and finding. The sentence "Some woman is adored by all Catholics" is true, because the utterer of the sentence can point to the virgin Mary and the sentence will then be true whichever Catholic (such as Pope Francis) the interpreter picks to falsify the sentence (EP 2, 168). Similarly Hintikka (1973) argues that the sentence "There are transuranium elements" is true, because one can produce them in a nuclear reactor. The rules for Peirce's and Hintikka's games for seeking and finding G(ϕ) can be given:

(1) The players are the Utterer and the Interpreter.
(2) The objects are the objects of the model M and their relationships (M, I).
(3) The game G(ϕ) in model M begins with the sentence ϕ and the interpretation {}.
(4) If $\phi = \neg\psi$, the Utterer and the Interpreter exchange turns and winning conditions, and the game continues from ψ.
(5) If $\phi = \psi \wedge \chi$, the Interpreter chooses ψ or χ, and the game continues from the subformula chosen.
(6) If $\phi = \psi \vee \chi$, the Utterer chooses ψ or χ, and the game continues from the subformula chosen.
(7) If $\phi = \exists x_n \psi x_n$ and the interpretation is s, the Utterer chooses a∈M, and the game continues from ψx_n and the assignment s∪$\{(x_n, a)\}$.
(8) If $\phi = \forall x_n \psi x_n$ and the interpretation is s, the Interpreter chooses a∈M, and the game continues from ψx_n and the assignment s∪$\{(x_n, a)\}$.
(9) If ϕ is atomic and the assignment is s, the Utterer wins iff the Interpreter loses iff ϕ is true in M on the assignment s.
(10) The sentence ϕ is true iff the Utterer has a winning strategy in the game G(ϕ). The sentence ϕ is false iff the Interpreter has a winning strategy in the game G(ϕ) (Pietarinen and Snellman 2006: 79).

Describing the language-games of seeking and finding then offers a basis for bottom-up or descriptive metaphysics. Strawson (1959: 15-86) and Glock (2012) argue that descriptive metaphysics involves the description of our conceptual scheme. Here it involves the description of the use of "there is". These descriptions also function as a background for Quinean descriptions for the values of quantification—i.e., the objects that are involved in the language-game and pointed out in it. The identification of objects then takes place in language-games

and according to its rules. Strawson argues that there are two necessary conditions for encountering and identifying objects. First, the objects must be located within a common grid of identifying reference, so that different speakers can refer to the same object. He gives the coordinate system for space (x,y,z,t) for visual identification, and the coordinate system (loudness, timbre, pitch, t) for an auditory world of sounds and voices. Second, objects must be reidentifiable across time and possible scenarios in order to be located in a grid of reference. We identify objects by locating them in a story of interactions, because it is stories that provide the character and characteristic properties of an object (see MacIntyre 1981, Smolin 2015). Physical objects are reidentified according to their causal roles and powers, and persons are reidentified through the characters they display in and through their actions (Snellman 2023).

The description of language-games for seeking and finding can then provide the identity-criteria for objects that in turn gives the essences and grounds for categorizing the objects of quantification: "Essence is expressed by grammar. […] Grammar tells what kind of object anything is. (Theology as grammar.)" (PI 371-373, see also ZH 7, 169). This Wittgensteinian and Hamannian slogan gives us a clue, how to develop a descriptive metaphysics out of the rules for language-games of seeking and finding. Grammatical description of language-games can help point out both the grids of possible properties, grids of identification and principles of reidentification in language-games. Garver describes how language-games can function as categories in the Aristotelian sense, as Aristotelian categories distinguish between different uses of "is" according to the various possible speech-acts associated with these senses (Garver 1994: 61-72). For example, "Is Viiru more of a cat than Tassu?" does not make sense because cats are substances, but "Is a fire engine redder than the red sun?" makes sense because red is a predicate or a property. Similarly, one can describe the practices of seeking and finding objects and pointing out their properties in order to get their possible property spaces and principles of reidentification (Snellman 2023: Ch. 4.3). Categories are then logical types of identity criteria for seeking and finding, and also types of objects that are typologized by these rules.

Wittgenstein distinguishes between looking at the blue colour of a vase and tracing its outline. There is a different bodily mediated sensuous practice or sensorimotor practice for pointing out colours and another for pointing out vases (PI 33-34, Noë 2004). These various habits then can be used to answer questions such as "What is the colour of the vase?" with "It's yellow" or "It's green", so yellow and green are the possible properties of the vase. Similarly, one can ask "What is the shape of the vase?" and have the possible answers "It's round" and "It's a cube", so roundness and cubeness are possible vase shapes. One can also ask questions about the location of the vase and its causal roles: "That's a nice vase. Where did you buy it?" or "Did you wash the vase? Where did you put it? Could you have put it in the cupboard?" We get a connection between questions and answers, activities of seeking and finding and properties, and identification grids and possibilities.

Question	Sensuous basic intuitions	Space of alternatives
Discourse possibilities	Possible values for aspect picked out	States of affairs

Moreover, possible answers to the questions about the purchase and location of the vase locate it in causal stories, which point to its location across time and at different possible locations. The storylines allow for reidentification across time and possible situations. Wittgenstein gives a similar grammatical description of mental states:

> Continuation of the classification of psychological concepts.
> Emotions. Common to them: genuine duration, a course. (Rage flares up, abates, and vanishes, and likewise joy, depression, and fear.)
> Distinction from sensations: they are not localized (nor yet diffuse!) […]
> Consider the following question: Can a pain be thought of, say, with the quality of rheumatic pain, but *un*localized? Can one *imagine* this?
> If you begin to think this over, you see how much you would like to change the knowledge of the place of pain into a characteristic of *what* is *felt*, into a characteristic of a sense datum, of the private object I have before my mind (Z 488, 498; quoted in Garver 1994: 70-71).

One can then categorize mental states according to how they are experienced. Their reidentification conditions are determined by their courses in time or paths of possible development in our lives, as they flare up and gradually cool down when our relationships to their objects change. One can also point to a pain in a leg, so that a pain is localized in the body. One can then characterize the category of emotions with the grid (Qualitative feeling at t, Expressions at t, Strength at t, Object at t) and reidentify them by pointing out their role in our lives by embedding them in a life story (see Snellman 2023: 4.3).

Wittgenstein also offers the concept of Übersicht to characterize his method of doing philosophy: one can define a simple language-game (e.g., PI 2) and use it as a point of comparison by isomorphically projecting it onto more complex language-games. Similarly, one can also view categories of logical types of identity criteria, which also characterize objects according to their natural types of continuity. Moreover, the term "category" also suggests that we can use mathematical category theory (see Smith 2016, Leinster 2014) to project logical types of rules onto our activities of seeking and finding and thus categorize the objects that are the objects of these activities. We can take E.J. Lowe's (1998: Ch.8) example of categorial criteria for change: the splitting of an uranium atom into a lead atom creates a new object, because the chemical element changes. The change of a tadpole into a frog and a caterpillar into a butterfly are lifecycle changes, because the DNA stays the same (see Snellman 2023: 4.3).

We thus have a rule of identification for animals: "All larvae turn into adult animals", or larva → adult. This logical rule is followed in non-metaphysical language-games by identifying lifecycle changes in frogs and caterpillars. An interpreter of nature or a researcher points to a caterpillar = Bfly (Larva) or a tadpole = Frog (Larva), and follows how they grow into a butterfly = Bfly (Adult) or to a frog = Frog (Adult) according to their real tendencies. Then there is a natural contrast or natural transformation between the cases of rule-following in the activities of applying the rule larva → adult in studying frogs or adults. Moreover, these comparisons are natural as they are fixed by the genetically fixed tendencies caterpillar → butterfly and tadpole → frog. The rules for a category thus point out logical types of activities of seeking and finding. The categorial rules also capture intrinsic necessities of DNA changes by making the contrasts made in applying the rule natural relative to the DNA change (PI 372), as the following commutes:

$$\begin{array}{ccc} \text{caterpillar = Bfly (Larva)} & \xrightarrow{DNA} & \text{butterfly = Bfly (Adult)} \\ \downarrow \exists \textit{rule contrast} & & \exists \textit{rule contrast} \downarrow \\ \text{tedpole = Frog (Larva)} & \xrightarrow{DNA} & \text{frog = Frog (Adult)} \end{array}$$

Language-games thus give the grounds for categorization, because categories are both logical types of activities of seeking and finding, and types of objects that can thus be described according to their types of properties and continuities. The focus on activities of seeking and finding and on metaphysical theories as charting models for "super-concepts" (PI 197) that can be embedded onto empirical activities also goes together as a view of metaphysical alternatives as high-level policies of looking at the world, because Gestalts and activities of seeking and finding go hand in hand.

3. Metaphysics, World-Views and the Starting-Points of Science

There is also a top-down approach to metaphysics that develops conceptual schemes for use as starting-points for scientific research. Morganti and Tahko (2017) have developed a "moderately naturalistic" approach to metaphysics. They argue that metaphysics and science have different methods but partially overlapping subjects: the abstract conceptual structures are applied as starting-points for scientific research and the theories are then tested against experience. One can next assess metaphysical theories by their fruits in a pragmatist manner (See Ochs 2004). I read Tahko and Morganti's view through a theory of frameworks in order to locate metaphysical alternatives like atomism, Aristotelianism and Spinozism as general conceptual schemes of a world-view.

Wittgenstein's *On Certainty* (OC) and Thomas Kuhn's *The Structure of Scientific Revolutions* (Kuhn 1969) are key books for the tradition of frameworks. Wittgenstein argues that the soundness and plausibility of arguments is always assessed against the background of an entire framework of propositions that function as rules in our language-games. For example, the sentence "This is a hand" is taken for granted, because it functions as a rule for seeking and finding hands and other material objects (see Hintikka 1973: 71). Learning a language-game means learning these framework propositions, so their use as standards is built into their role in the game. Kuhn's concept of a paradigm similarly explores how frameworks of scientific research (laws, examples of problem-solving, metaphysical commitments, values) structure experimental activity and the experimental activities of seeking and finding in science. A paradigm-shift and the associated shift of metaphysical commitments then leads to new Gestalt-perception of reality: burning is seen-as phlogiston escape in a phlogiston theory but it is seen-as oxidization in an oxygen theory. Paradigms moreover shift through scientific revolutions. A paradigm becomes established when it can solve key open problems with its laws and metaphysical commitments. It then offers a model for interpreting phenomena by applying the resources of the framework (laws, examples of problem-solving, metaphysical commitments, values) to solve open problems like puzzles. One paradigm is replaced by another one if it starts to encounter anomalies or open problems that it is not able to solve through its resources, and a competing paradigm can solve them.

One can take a logical point of view of the world-view commitments of a language-game, Gestalts and world-view circles. There is a strong link between Gestalt-perceptions and activities of seeking and finding. Wittgenstein (PI part 2, xi) gives the example of the puzzle-picture of a face formed by an outline of tree-branches. The picture can be seen as trees or as a face by different sensorimotor practices that embody different activities of seeking and finding. One can trace the organization of tree-trunks and see the picture as trees. One can spot the face

in the picture by tracing the outline or structure of the face, and thus see the picture as a face. Locating a picture or a phenomenon in a context moreover establishes analogies or metaphors that determine the sensuous practices of seeking and finding. The letter H can be seen as shoddy, legalese or childish by imagining drawing it shoddily, lawyers writing it, or children learning to write it. A Gestalt-perception is a thought flashing through sight, because the sensorimotor activities of seeking and finding are already proto-conceptual recognition activities in a context (see Snellman 2023: Ch. 4.2; Noë 2004: Ch.6).

Kuhn (1969) defines a paradigmatic circle of paradigm → solving open problems → anomaly → scientific revolution. The paradigmatic circle can however be seen as a world-view circle: forming a world-view → drawing interpretations → assessing and modifying world-views (see Polanyi 1959: 264-267; Naugle 2004: 310-321). The world-view circle is however a generalization of Peirce's pragmatic circle: abduction → deduction → induction. Peircean abduction means guessing the best or most natural explanation for a phenomenon, while deduction means drawing logical conclusions about the hypothesis and induction means testing the conclusions statistically. (EP 1, 186-209, EP 2, 443-445.) The exploration and testing of world-views can then be viewed through a pragmatic logic. The connections between Gestalt-perception and seeking and finding also means that exploring new ways of seeking and finding can be used to define new ways of interpreting empirical phenomena and looking at the world. They can lead to new empirical results and new ways of conceptualizing and categorizing existing results. Categorial principles and language-games rules like "This is a hand" and "All larvae grow into adult animals" can moreover be embedded onto our empirical practices of seeking and finding, so that they can be seen as a kind of abstract framework or a high-level strategy for interpreting experience.

Top-down or constructive metaphysics thus offers abstract principles or general conceptual schemes, which can be used to define new scientific paradigms and practices of seeking and finding. Metaphysical theories can help us make sense of the world in our practices and can be compared by assessing the associated world-views. Morganti and Tahko (2017) offer the following model:

(1) Metaphysicians create a general conceptual model of being qua being or the nature of some part of reality. Metaphysicians analyse the model, elaborate it and derive logical consequences of it.
(2) Metaphysical theories offer alternatives for scientific theorizing. Metaphysical theories are used as world-view- and paradigm-level backgrounds for scientific theory formation. For example, materialist atomic theories or the idea of infinitely divisible "gunk" can be used as world-view level models when forming physical theories.
(3) Metaphysical theories prove to be good or bad according to whether the paradigms and scientific theories operationalizing them manage to interpret empirical phenomena. Metaphysical interpretations are assessed with concepts like simplicity, coherence, applicability and other theoretical virtues.
(4) The use of metaphysics in forming world-view level presuppositions of scientific theories gives the abstract categorial terms (substance, relation, law of nature, property, identity, relation…) an empirical interpretation. The practice of testing hypotheses also locates the theoretical virtues of metaphysical theories in empirical interpretative practices.

We can also use Roy Bhaskar's (2008: 183-184) view of the levels of scientific research and contrast it with the levels of strategy from conflict studies (Ackerman

and Kruegler 1994: 45-48). Scientific research proceeds from a general conceptual scheme, which corresponds to policy-level strategies for viewing the world. General conceptual schemes lead to paradigms, which operationalize them by indicating how the conceptual resources of a conceptual scheme are to be mobilized to achieve its interpretative goals. Theories then offer maps or models for scientific expeditions of understanding phenomena, and they also define the campaign strategies of seeking and finding objects in a given phenomenon (see Ziman 2000: 126-132). Research practices like arguments and experimental manipulations are tactics, because they implement the strategy of interpretation provided by a theory.

General conceptual schemes like atomism are grand strategies or policies for viewing the world. A general conceptual scheme includes a network of concepts that functions as a high-level map for understanding and navigating in the world. It also offers guidelines for interpreting and explaining the world at a general level, as these concepts have their logic and associated strategies of possible application and explanation. A general conceptual scheme also has interpretative goals and often also aims at meeting practical needs in human life. It can then be given as (conceptual system, interpretative resources, goals). Newtonian mechanistic materialism, which includes atoms, voids and forces as fundamental concepts, offers an example of a conceptual scheme. Its explanations may only appeal to spatial and kinematic factors (mechanism). They must explain complex wholes in terms of their simple parts (reductionism) and fix the future based on the current state (determinism). Moreover, mechanical materialism attempted to explain the entire world by reducing everything to the movements of atoms in a void (Kallio 1996, Burtt 2015).

Paradigms like Newton's model of the solar system operationalize conceptual schemes. They define standard scientific operating procedures and values for turning the general models of a general conceptual scheme into a network of theories for interpreting phenomena: (general models, theory matrix, standard interpretative practices). (Ziman 2000: 192-198.) Alternatively, Kuhn (1969: Afterword) defines them as a matrix (laws, examples of problem-solving, metaphysical commitments, values). Newton's model of a solar system places the sun at the centre, and gravity causes planets to orbit it. The model uses Newton's law of gravity ($F = \frac{Gm_1m_2}{d^2}$). It operationalizes the mechanistic world-view, because the Sun and the planets have a place and a momentum that determine the forces in the system, and all forces are vector sums of their components. The explanation of planetary orbits is a paradigm case for explanation in Newton's model. All planets fall towards the Sun but their momentum is along their orbit, so the planets circle the Sun like a ball swirling at the edge of a string. The values of Newtonian science also privilege mathematical explanation, as dependencies are to be first expressed as mathematical dependencies and then tested empirically (Kallio 1996, Burtt 2015).

Ziman (2000: 123-132, 192-198) argues that paradigms offer a point of departure for scientific campaigns and expeditions, which aim at understanding phenomena by building theories about them. Theories and models define the strategies of these scientific campaigns, as they allow us to seek and find their objects in phenomena through interpretative activities. He also compares theories with maps and models, and models with metaphors. Theories are maps, because both theories and maps represent a functional structure in reality through use, and

these representations are for a given purpose. Theories and maps are both models, or symbolic systems representing a real one. A model uses symbols to point out the parts of a system, and its functional interrelationships according to its interactions (ZH 7, 169-170). The isomorphism of a model and the functions of a system then allow us to see the system as the model, because the isomorphism between the symbols and the phenomena give us a way of sensuously seeking and finding the functional parts and relationships of the phenomenon through theory-laden experience.

Take the example of a metaphor between DNA and codes. The metaphor of reading a file, sending it to a printer and then reading the printout can be used as a model for chemical DNA reading in a cell nucleus, MRNA transfer onto ribosomes, and protein production. This process allows us to identify (i.e., seek and find) codes in the functioning of molecules and to understand their roles in the relationships of a cell. Arguments, analogies, manipulations and experiments of the scientific interpretation are then the tactics of a scientific expedition (Ziman 2000: 147-151; Snellman 2023).

The role of top-down or constructive metaphysics can then be characterized by reading Morganti and Tahko's (2017) proposals for the scientific assessment of metaphysics and the levels of interpretative strategies through the world-view circle. The function of constructive or top-down metaphysics is to define a metaphysical theory or a world-view which then functions as a general conceptual scheme, or as a kind of policy or higher-level strategy for looking at the world. Analytic metaphysics can also draw out the logical consequences of these conceptual schemes in order to articulate their conceptual maps of reality, explanatory strategies and goals. The role of metaphysics then corresponds to the world-view formation stage of the world-view pragmatic circle (Polanyi 1959: 264-267).

These higher-level interpretations are used as a background for scientific theorizing when they are operationalized through paradigms and research programs. The paradigms also define networks of theories and possible practices of interpretation, which lead to looking at phenomena from a new angle or having a new Gestalt-perception of them. Since Gestalt-perceptions however are associated with the sensorimotor practices of seeking and finding objects, the category system of the conceptual scheme and the analogies offered by a paradigm lead to new Gestalts by defining a new ontology for theories as well. The paradigm-formation and theoretical interpretation phase also corresponds to the interpretative stage of the world-view circle (Snellman 2023, Kuhn 1969).

The role of metaphysics as formulating a background framework for paradigm- and theory-formation however calls into question Morganti and Tahko's (2017) straightforward appeal to theoretical values and abductivist methodology. Kuhn famously argues that different world-views are incommensurable, and they ascribe different meanings to theoretical virtues like simplicity and coherence (Kuhn 1969: Afterword; see also Polanyi 1959: esp. 145-171). Then the dependence of both the interpretation of theoretical values and of empirical results on a background framework leads to a puzzle: how is the testing and comparison of incommensurable world-views possible? Since linguistic activities give the background for experience and argument, the testing of theories involves a comparison of their languages. The assessment, criticism, and modification of world-views then has to involve comparing different frameworks to assess whether they are good practices for looking at reality.

4. Comparison of World-Views and Metaphysical Conceptual Schemes

Metaphysics thus deals with the conceptual schemes and grand strategies of viewing the world. Descriptive or bottom-up metaphysics attempts to characterize the language-games of quantification and encountering reality, the objects encountered and the logical types of rules and objects for categorizing them. Constructive top-down metaphysics develops conceptual schemes and rules for categories, which are then operationalized through scientific paradigms and define new ways of looking at phenomena, new Gestalts, and new activities of seeking and finding (see OC, H 214-216).

The question of scientific metaphysics is then intertwined with the question of world-views: how can different world-views be contrasted and compared? In the philosophy of science, the question has often been put in terms of incommensurability: how can we contrast different conceptual systems when they have by definition different conceptual logics and lead to different perceptions of the world? (See Kuhn 1969; Naugle 2004; Taylor 1995: Ch.3). I use Hamann's account of the comparison of incommensurable languages to generalize Peirce's pragmatic circle by describing, how one can test and compare world-views by contrasting their respective pragmatic circles in a dialogue of world-views. Hamann took up the issue of contrasting conceptual schemes as early as 1759:

> Everybody understands his language and not those of others; Descartes has understood his reason, Leibniz his, and Newton his. Do they understand themselves better through mutual conversation (untereinander)? We must learn their languages, in order to analyze their concepts; we must test their materials; we must investigate the designs of their doctrinal constructions, their grounds, their ends and the conclusions. This must not be according to their promises and presuppositions that they burden us with by offering them as axioms, empirical facts and conclusions (N I, 30-31).

Hamann then takes up the incommensurable conceptual schemes of Enlightenment thinkers, emphasizing that researchers using incommensurable languages can understand each other and also gain a better understanding of their own conceptual schemes by learning the languages of others and contrasting them with their own conceptual schemes. There are two different ways of characterizing conceptual schemes. The first uses a given conceptual scheme to translate the concepts of another language Y into one's own X, or analyse the concepts of another with a synthesis of one's own concepts (ZH 7, 175, Davidson 1984). The other describes the activities of language-use: since the concepts are located in language-games, one can describe the whole activity by, for example, giving an overview of it or rules for learning it (see Taylor 1985, 256-282, Hintikka 1997, Preface). Moreover, the axioms and materials correspond to general conceptual schemes, the empirical facts correspond to practices of drawing interpretations from the world-view and the conclusions are something to be assessed through contrast. The world-view circle of conceptual scheme → drawing interpretations → modifying and assessing a world-view then arises, but assessment takes place by comparison of multiple world-views.

Hamann also discusses the conflict of languages in a letter to Jacobi (ZH 7, 175; Bayer 2012: 156-170 = 2002: 1-21). One language X calls a phenomenon p

"faith" and related claims "true", but Y calls it a "delusion" and labels the claims false, so X and Y offer rival categories to reinterpret the same phenomena. The languages X and Y are underpinned by different world-views and underlying practices, or forms of life (PI 19), and the different world-views rest on these differing ways of acting in the world. Both X and Y aim at interpreting the concepts of others in terms of their own manner, but the dialogue is not one of static translation into a given metalanguage as in Davidson (1984). Instead, there is a constant tug-of-war between the conceptual schemes, because the interpretations conflict and both X and Y can learn from each other:

(1) The speaker of language-game X learns language-game Y and analyses the expression y of Y with the expressions of X: "y" is true iff x, x',…, "y" is used iff x, x'…
(2) The speaker of a language-game X learns language-game Y and encounters an expression y with the rules and use U_y which does not have a corresponding concept in X. X is modified to include the expression x with the rules and use U_x s.t. x and y have the same use conditions.
(3) As in 1,2 but with language-games X,Y and expressions x,y interchanged to reflect changing roles.

The language-enrichment move is one possibility that makes Hamann's scheme stronger than Davidson's. There is another possibility of using both X and Y as pointers to a larger metalanguage or a language-game Z, which can form a metatheory or a synthesis for both X and Y and includes both as limited subgames. Peirce (EP 2, 411-418) describes finding a solution to a maths problem as creating a new strategy of problem-solving or seeking and finding solutions by using current knowledge as clues. Polanyi (1959: esp. 71-76) similarly describes how a rat learns to run a maze: she gains a true understanding of the situation by formulating a mental map, which also functions as a strategy when making turns in a maze. The forming of new interpretation Z then offers a mental map or a new language-game for encountering the realities revealed by X and Y. Z is formed by taking the existing problems, the facts we encounter by trying to solve them in X and Y and the functioning of X and Y in the encounter as clues. Z then reinterprets and locates both the facts of X and Y as part of the wider map or conceptual scheme it offers, and the habits of X and Y in the language-game Z. Z can then function as a metatheory in the Davidsonian sense of translation-rules 1 and 3.

Language-games X and Y can moreover be contrasted by describing their structures as games. Hintikka (1997: Preface) argues that we can talk about the meaning of our languages because meanings are embedded in language-use and we can describe our practices of use. Taylor similarly argues that we can formulate truth- or use-conditions like rules 1-4 only by describing a language-game and its expressive functions as a whole. For example, the language-game of the builders can be described by giving its relationships (PI 2; Snellman 2023: Ch. 4.1):

(1) The players are A and B.
(2) The objects of the game are slabs, girders, pillars and cubes.
(3) The word-signs of the game are "Slab!", "Girder!", "Pillar!" and "Cube!".
(4) The context of the game is building a house. Therefore, A wins iff B wins iff B brings the material that A calls for, e.g., a slab for "Slab!" and the endpoint is e.g. (("Slab!", Slab), ("Pillar!", Pillar)…).
(5) The actions c_n of the game are the speech acts of shouting the word-signs and bringing materials.

(6) A plays at the start of the game, and when B has delivered a building-block. The actions a_n of A are shouting the word-signs of the game.
(7) B plays when A has shouted a word-sign. The actions b_n of B are bringing building-blocks to A.

Languages can then be described by describing their practices, or by using one as metalanguage to analyse the other and vice versa. They also can mutually enrich each other, either by adding concepts from the other or being a basis for a synthesis. This leads to the question of how conflicts between languages (Bayer 2012: 156-170; ZH 7, 175) are to be resolved.

MacIntyre (1988: 349-369) offers an account of comparing different traditions or world-views, which is at the same time a Hamannian conflict-of-languages model and a Peircean pragmatist view. An enquirer starts from her own tradition X and she can learn the language of Y, as in the Hamannian model. MacIntyre argues that the next step in the comparison between X and Y is to assess their strategies for dealing with the world by seeing how well they can encounter phenomena in the world by interpreting and categorizing them with their conceptual resources. Both traditions X and Y have their own epistemologies, because they have their framework rules for interpreting experience and arguments. These epistemologies or standard scientific procedures and conceptual resources then open up different ways of identification, classification and characterization of the reality that is made manifest in our activities. One then gets an account of testing world-views by looking at their activities of seeking and finding. A practice is adequate to reality or true iff it is not defeated by a future discrepancy with the revealed reality. Falsity then is failure of a representation shown by anomalies and dialectical questioning. MacIntyre's view of truth then resembles Peirce's in that we cannot know that our representations will prove correct in the future and truth means that our strategies and practices for interpretation are not defeated in the long run (see Pietarinen and Snellman 2006; EP 2, 339-341).

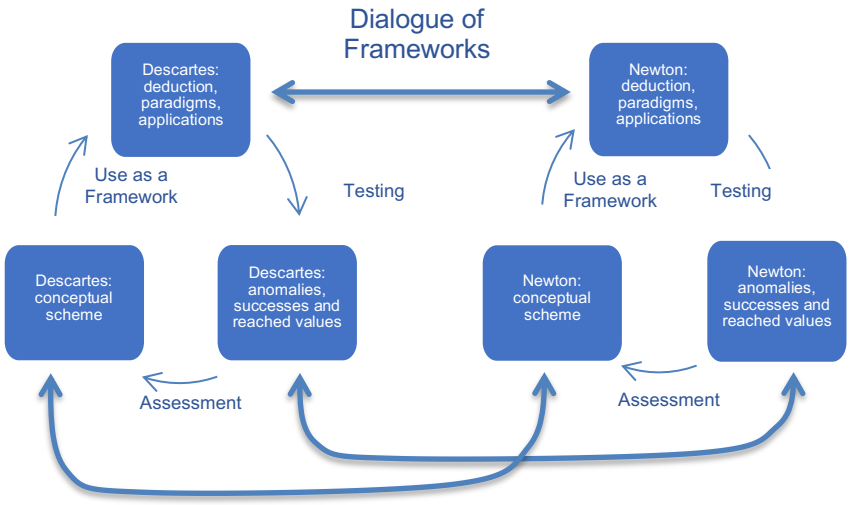
MacIntyre's pragmatist account of truth forms the basis for a comparison and testing of world-views. An enquirer views the world through the prism of the language-games X and Y, and sets out to find anomalies for their interpretation strategies. Now if X can point out some anomaly y Y cannot solve and solve it from X's resources or vice versa, X is shown to be stronger than Y and vice versa. MacIntyre considers the situation where X can not only solve Y's anomalies y, but it can also prove that Y does not have the resources to solve them and explain why Y's resources are insufficient. This however amounts to a falsification of Y in a broad Peirce-Hintikka sense. If X can show that the strategies of Y = (conceptual system, interpretative resources, goals) are not sufficient for pointing out and interpreting some phenomenon and are instead defeated, then Y is false because it has no successful interpretative strategy for recognizing the phenomenon in question. Then the insufficiency of the categories of Y is shown by using X to show that Y has no strategy to recognize the reality y out of its conceptual resources. Taylor gives further cases of comparison or testing in which a framework can be better than its competitors by recognizing some reality, value or problem:

(1) A and B are checked against a body of facts. The theory explaining more facts wins. (Popper, Peirce)
(2) A and B solve problems in parallel. A looks for anomalies in B and vice versa. If A can point out that B cannot explain some anomaly, then A wins and vice versa. (MacIntyre)

(3) A and B develop ways of dealing with reality according to their different goals. A is shown to be better if B cannot recognize the success of A out of B's resources. For instance, Aristotelian astronomers could not look into Galileo's telescope or explain the success of modern science.
(4) A takes an element from B but is better able to orient itself towards human good by rearranging the elements of B and leaving some out. E.g., banning judicial torture led to more humane punishments.
(5) The transition A → B directly removes some error, contradiction, confusion or allows one to point out some blind spot or obfuscation. E.g., recognizing one's anger leads one to take others into account and to read their actions better (Taylor 1995: Ch. 3).

Taylor's cases then depend on the success of our interpretative strategies in recognizing reality and orienting us in the space of values (see Taylor 1989, Hein 1983). The first case can recognize facts. The second case was discussed, but in the third case Aristotelian astronomy (A) cannot account for Galileian astronomy's (B) strategies for encountering reality and ways of achieving human cognitive values. In the fourth case, a ban on judicial torture (A) can better recognize and attain the good of human dignity already recognized by early modern court practice (B), although judicial criteria have changed in the move from A to B. Recognizing an error in A both improves our ability to come to terms with reality and chart the problems of B. Testing interpretations and comparing world-views then involves contrasting them and examining, if they can point out realities with their categories and conceptual resources, and if they can realize values arising out of the human condition.

These comparisons of incommensurable world-views in fact amount to comparisons between two different pragmatic circles: both A and B have their world-view circles (forming a world-view → drawing interpretations → assessing interpretations, testing and modifying world-views). The conceptual schemes of A and B are then contrasted through the dialogue of incommensurable world-views. Their abilities to form interpretations and solutions are contrasted and charted by comparing their conceptual resources and explanatory strategies. Their practical utility, value-conformity and power are explored both by pointing out realities, facts, and by anomalies in dialogue with reality. The multiple interconnected pragmatic circles running in parallel like an Enigma machine are then compared with the resources provided by their dialogue:



5. Conclusion

Descriptive metaphysics gives an account of objects by describing quantificational language-games for seeking and finding. Top-down or constructive metaphysical theorizing offers a framework for empirical investigation. It is interpreted by using it to define new practices for recognizing objects and relationships through phenomena, thus developing new ways of seeking and finding them. An underlying link between bottom-up descriptive and top-down constructive metaphysics is the role of activities of seeking and finding as the background for the concept of being and for Gestalts and world-views as well. Metaphysics thus articulates conceptual schemes and world-views for our language-games and for use in investigating and encountering the world.

The world-views explored by metaphysics are contrasted by their ability to recognize objects and to develop strategies for interpreting empirical phenomena. The contrast takes place in a dialogue between traditions, which defines a set of interrelated pragmatic circles for world-views. World-views then offer starting-points, are operationalized by paradigms and lead to theories and Gestalts. The resources of a world-view are then tested both in the pragmatic circle and through the contrasts between world-view circles in dialogue and dialectical questioning. Metaphysics is thus both bottom-up descriptive and top-down interpretative, assessed by contrasting metaphysical and quantificational systems. Metaphysics can then be a science in a broad Peircean sense, involving pragmatic circles of developing interpretations, deducing possible approaches and theories and then testing them in a dialogue of seeking and finding objects and relationships in phenomena.

References

Ackerman, P. and Kruegler, C. 1994, *Strategic Nonviolent Conflict*, Westport: Praeger.

Bayer, O. 2012, *A Contemporary in Dissent: Johann Georg Hamann as a Radical Enlightener*, Grand Rapids: Eerdmans.

Bayer, O. 2002, *Vernunft ist Sprache: Hamanns Metakritik Kants*, Stuttgart: Frommann-Holzboog.

Bhaskar, R. 2008, *A Realist Theory of Science*, London: Verso.

Burtt, E.A. 2016, *The Metaphysical Foundations of Modern Science*, Angelico Press.

Davidson, D. 1984, *Inquiries into Truth and Interpretation*, Oxford: Clarendon.

Dickson, G.G. 1995, *Johann Georg Hamann's Relational Metacriticism*, Berlin: de Gruyter.

Garver, N. 1994, *This Complicated Form of Life*, La Salle: Open Court.

Glock, H.J. 2012, "Strawson's Descriptive Metaphysics", in Haaparanta, L. and Koskinen, H. (eds.), *Categories of Being*, Oxford: Oxford University Press.

Glock, H.J. 1996, *A Wittgenstein Dictionary*, Oxford: Blackwell.

Hamann, J.G. (1955-1979), *Briefwechsel 1-7,* Edited by Walther Ziesemer and Arthur Henkel, Frankfurt am Main: Insel Verlag. (ZH 1-7)

Hamann, J.G. 1947-1951, *Sämtliche Werke 1-3*, Edited by Josef Nadler, Wien: Verlag Herder. (N I-III)

Hamann, J.G. (2007), *Writings on Philosophy and Language*, Edited by Kenneth Haynes, Cambridge: Cambridge University Press. (H)

Hein, H. 1983, "Hamann und Wittgenstein. Aufklärungskritik als Reflexion über die Sprache", in Gajek, B. (ed.), *Acta des zweiten internationalen Hamann-Colloquiums in Marburg/Lahn*, Marburg: Elwert, 21-57.

Hintikka, J. 1973, *Logic, Language-Games and Information*, Oxford: Clarendon Press.

Kallio, T. 1996, *Kvanttilainen todellisuus*, Helsinki: Gaudeamus.

Kant, I. 1998, *Kritik der Reinen Vernunft*, Hamburg: Felix Meiner Verlag. (KrV)

Kuhn, T. 1969, *The Structure of Scientific Revolutions*, Chicago: Chicago University Press.

Lowe, E.J. 1998, *The Possibility of Metaphysics*, Oxford: Clarendon Press.

Ladyman, J., Ross, D., Spurrett, D. and Collier, J. 2007, *Every Thing Must Go*, Oxford: Oxford University Press.

Leinster, T. 2014, *Basic Category Theory*, Cambridge: Cambridge University Press.

MacIntyre, A. 1981, *After Virtue*, London: Duckworth.

MacIntyre, A. 1988, *Whose Justice? Which Rationality?*, London: Duckworth.

Mainzer, T. 2004, "System: An Introduction to Systems Science", in Floridi, L. (ed.), *The Blackwell Guide to the Philosophy of Computing and Information*, Malden: Blackwell, 2004, 28-39.

Naugle, D.K. 2002, *Worldview*, Grand Rapids: Wm. Eerdmans.

Noë, A. 2004, *Action in Perception*, Cambridge, MA: MIT Press, 2004.

Peirce, C.S. (1996-1998), *The Essential Peirce*, vol. 1-2, Edited by The Peirce Edition Project, Bloomington: Indiana University Press. (EP)

Pietarinen, A.V. and Snellman, L. 2006, "On Peirce's Late Proof of Pragmaticism", in *Truth and Games: Essays in Honour of Gabriel Sandu*, Helsinki: Suomen Filosofinen Yhdistys, 275-288.

Polanyi, M. 1958, *Personal Knowledge*, Chicago: University of Chicago Press. (PK)

Quine, V.W.O. 1953, "On What There Is", in *From A Logical Point of View*, Cambridge, MA: MIT Press.

Simons, P. 1995, "Metaphysics: Definitions and Divisions" in Dancy, J. and Sosa, E. (eds.), *A Companion to Metaphysics*, Oxford: Blackwell, 310-312.

Smolin, L. 2017, *Three Roads to Quantum Gravity*, New York: Basic Books.

Snellman, L. 2023, *Evil and Intelligibility*, Leiden: Brill.

Strawson, P.F. 1959, *Individuals*, London: Methuen.

Taylor, C. 1995, *Philosophical Arguments*, Cambridge, MA: MIT Press.

Taylor, C. 1989, *Sources of the Self*, Cambridge: Cambridge University Press.

Wittgenstein, L. (1971), *On Certainty*, Edited by G.E.M Anscombe and G.H. von Wright, Translated by G.E.M Anscombe and D. Paul, Oxford: Blackwell. (OC)

Wittgenstein, L. 1971, *Philosophical* Investigations, Translated by G.E.M Anscombe, Oxford: Blackwell. (PI)

Ziman, J. 2000, *Real Science*, Cambridge: Cambridge University Press.

Laws of Metaphysics for Essentialists

Tuomas E. Tahko

University of Bristol

Abstract

A recent methodological approach at the interface of metaphysics and philosophy of science suggests that just like causal laws govern causation, there needs to be something in metaphysics that governs metaphysical relations. Such *laws of metaphysics* would be counterfactual-supporting general principles that account for the explanatory force of metaphysical explanations. There are various suggestions about how such principles could be understood. They could be based on what Kelly Trogdon calls *grounding-mechanical explanations*, where the role that grounding mechanisms play in certain metaphysical explanations mirrors the role that causal mechanisms play in certain scientific explanations. Another approach, by Gideon Rosen, takes it that there are *essentialist* principles or laws that tell us about what grounds what. Finally, Jonathan Schaffer defends an approach that he considers to be neutral regarding grounding or essences. In this paper I will assess these suggestions and argue that for those willing to invoke a non-modal notion of essence, there is a more promising route available: metaphysical and scientific explanations may be unified in terms of general essences. Accordingly, essentialists may be better viewed as *outlaws* when it comes to laws of metaphysics.

Keywords: Grounding, Essence, Metaphysical Explanation, Scientific Explanation, Dependence, Metaphysical Laws.

1. Unifying Scientific and Metaphysical Explanation

This paper discusses two interesting, related questions at the interface of metaphysics and philosophy of science. They are both linked to the idea that there is an important analogy—or more than just an analogy—between scientific explanations that involve causal laws or laws of nature (I use these notions synonymously), and metaphysical explanations that involve laws of metaphysics. Laws of metaphysics could be understood as counterfactual-supporting general principles that are responsible for the explanatory force of non-causal, metaphysical explanations. Here is a simple example, which assumes that set membership captures a distinctly metaphysical relation: 'if Socrates exists (or existed), then the singleton set of Socrates, {Socrates}, exists' (cf. Fine 1994, Schaffer 2018). And here is an analogous example of a counterfactual-supporting principle in the realm

Argumenta 10, 1 (2024): 71—88
ISSN 2465-2334

DOI 10.14275/2465-2334/20240.tah
First published: dd November 2024

of laws of nature: 'If a positively charged particle were to come in the vicinity of a negatively charged particle, these particles would attract each other'. The two questions to be discussed are:

(1) Is the proposed analogy between scientific and metaphysical explanation substantive and helpful?[1]

(2) Can we unify scientific and metaphysical explanation?

If the answer to the first question is affirmative, then this could provide a route towards a positive answer to the second question.

Metaphysical explanation itself is now commonly discussed under the label *grounding*. There are several suggestions in the literature as to what the relationship between scientific explanations involving causation and metaphysical explanations involving grounding is supposed to be. A strong motivation to develop theories about this connection is related to the *unity of explanation*, the thought that our explanatory endeavours in the sciences and in philosophy are importantly similar, if not identical. Here are a few representative quotations from recent work in this area:[2]

> [T]here is a far-reaching structural analogy between causation and grounding. Just as earlier states of the universe typically give rise to later ones by causing them, metaphysically more fundamental facts give rise to less fundamental ones by grounding them. Certain general metaphysical principles, which I will call 'laws of metaphysics', play essentially the same role in grounding as natural laws do in causation (Kment 2014: 5).

> The unificatory role of explanation clearly calls for explanations to involve generalizations, which serve to subsume a given case under a more general pattern. But it is also worth noting that the generalizations involved cannot merely happen to hold in our world, but must also be *non-accidental* generalizations which are counterfactually robust. And so the unificatory role of explanation requires the presence of counterfactual-supporting general principles, to serve as stable patterns (Schaffer 2018: 7).

> […] just as there is a type of scientific explanation that appeals to causal mechanisms—causal-mechanical explanation—there is a type of metaphysical explanation that appeals to grounding mechanisms—grounding-mechanical explanation (Trogdon 2018: 1290).

Each of these approaches is different and I cannot discuss all the details here, but I take it that they share an important hope, namely, the hope to unify (at least a subset of) scientific and metaphysical explanations. In each case, this hope is strongly supported by an analogy between certain aspects of metaphysical and scientific explanation, specifically, an analogy between grounding and causation. This suggests an affirmative answer to question (1).

There has also been a significant critical reaction to this claim of unity between scientific/causal and metaphysical explanation, and especially to the

[1] Thanks to an anonymous reviewer for suggesting this formulation of the question.

[2] Other important work discussing the relationship between scientific and metaphysical explanation, as well as grounding and causation, includes Bennett 2017, Bernstein 2016, Fine 2012, Glazier 2016, Koslicki 2016, Kovacs 2017, 2020, Rosen 2017, Schaffer 2016, Wilsch 2015, J. Wilson 2014, 2016, and A. Wilson 2018.

analogy between grounding and causation (e.g., Bernstein 2016, Koslicki 2016, and J. Wilson 2014, 2016). One motivation behind this reaction is scepticism about grounding more generally. I am sympathetic to the arguments suggesting that we ought to go more fine-grained and distinguish between different metaphysical dependence relations, or 'small-g grounding relations', such as composition, functional realization, and set membership, instead of trying to account for all of these in terms of a unified notion of grounding (Wilson 2014: 539). This would appear to suggest a negative answer to question (1), although it's not clear whether this is the direct intention of all those who have criticised the grounding-causation link.

However, even if one favours a variety of metaphysical dependence relations instead of a singular 'big-G grounding relation', this does not necessarily entail a negative answer to question (2). I suggest that we can unify scientific and metaphysical explanation despite the challenges posed by a more fine-grained approach to metaphysical dependence relations. Accordingly, I wish to defend a positive answer to question (2), albeit motivated differently from the one developed on the basis of a positive answer to (1). Instead of a direct analogy between grounding and causation, I will seek a unified account of scientific and metaphysical explanation via *essentialist explanation*—the notion is familiar from Martin Glazier (2017) with this very title. While Glazier argues that some metaphysical explanations that involve essences cannot be understood in terms of ground (without any dedicated attention to scientific explanation), I will argue that at least some scientific explanations are best understood as involving essences (while remaining neutral about whether or not they can also be understood in terms of ground).[3]

2. Explanation Tracks Dependence

A key assumption of the framework that I wish to adopt is the idea that *any* kind of explanation must be linked to dependence relations. Specifically, what gives explanations their explanatory power is some relation or relations of dependence that obtain between the *explanandum* and the *explanans*. Roughly, this allows us to distinguish between 'worldly' or metaphysical, and representational or epistemic content. This is a rather traditional view, which can be found, for instance, in Jaegwon Kim's account of metaphysical explanation:

> My main proposal, then, is this: *explanations track dependence relations*. The relation that "grounds" the relation between explanans, *G*, and its explanatory conclusion, *E*, is that of dependence; namely, *G* is an explanans of *E* just in case *e*, the event being explained, depends on *g*, the event invoked as explaining it (Kim 1994: 68).

Kim is not using 'grounds' in the technical sense invoked in the contemporary grounding literature (because this use had not yet been introduced), but the view he entertains seems to be straight-forwardly compatible with the 'tracking' or 'backing' view of metaphysical explanation that is receiving attention in the grounding literature (e.g., Audi 2012: 119–120, Schaffer 2012: 124, Trogdon 2013: 103–104, Thompson 2016: 44, Maurin 2019, Sjölin Wirling 2020, and Skiles and Trogdon 2021).

[3] I am sympathetic to the thought that we can give a reductive account of ground in terms of essence, but I will not pursue this line here.

The tracking view of metaphysical explanation enjoys relatively wide support, but the notion of 'ground' can be used to express both metaphysical and epistemic content. Sometimes the issue is put in terms of *unionism* and *separatism* (e.g., Raven 2015: 326). Unionism is the view that grounding is a type of metaphysical explanation and hence explanatory in its own right, whereas separatism distinguishes grounding and (metaphysical) explanation. On the latter view, ground and metaphysical explanation may be separated in such a way that ground is the metaphysical part and metaphysical explanation is the epistemic part, as it were. But the two aspects are linked via the idea that grounding relations back metaphysical explanation. The reason why this issue is particularly relevant in the present context is that this is thought to be analogous to the case of causation, i.e., causal explanations are backed by the causal relations in the world.

My own sympathies are primarily with separatism, broadly speaking: it provides a natural distinction between the metaphysical content, i.e., a worldly relation or relations of grounding or dependence, and the epistemic content, i.e., metaphysical explanation as a form of mind-dependent understanding. We can make a similar distinction in the case of scientific explanation and the causal (or similar) relations that back those explanations. In fact, this is one sense in which these explanations could be considered analogous.

The key upshot is that since 'laws of metaphysics' involve metaphysical explanations and all explanations track dependencies, there must be some dependencies underlying these 'laws' or whatever does the relevant explanatory work.

3. Laws of Metaphysics?

In this section I will first consider Jonathan Schaffer's (2018) take on the laws of metaphysics, before suggesting an alternative understanding of them in terms of essence, with reference to Gideon Rosen's account.

Schaffer attempts to put forward an understanding of laws of metaphysics which is neutral with regard to grounding or essences (although he does appear to also commit to the idea that metaphysical explanation is backed by grounding relations). To be a 'law' is here understood minimally, a law is a counterfactual-supporting general principle. Schaffer's case for the laws of metaphysics is simple: if there are metaphysical explanations, they require laws of metaphysics—counterfactual-supporting general principles—in order to have explanatory force. One argument that Schaffer considers in favour of this idea is that there is a unificatory role of explanation and this role calls for explanations to involve counterfactually robust generalizations, i.e., laws of metaphysics. He also puts forward an argument from causal explanation and from paradigm cases, but all three of his arguments are interconnected. I will frame my discussion of Schaffer's proposal in terms of the following three issues:

(1) If the account is neutral with regard to grounding and essence, then what makes metaphysical explanations *metaphysical*? In other words, what is supposed to be distinctively metaphysical about the laws of metaphysics?

(2) Even if the unificatory role of metaphysical explanation is important, why should we need laws of metaphysics to uphold this role? For those of us willing to invoke essences, there is a straightforward route to unification, or so I will argue, via the involvement of (robust, genuine, counterfactually stable) *general* or *natural kind essences*.

(3) The suggested distinctly metaphysical principles involved in 'paradigm cases' of laws of metaphysics, such as *set formation*, can be equally well (or better) accounted for in terms of general essences, which Schaffer eschews.

The suggested upshot of my analysis is that laws of metaphysics collapse to general essentialist principles. Let us look at each of these three issues in a little more detail.

3.1. What Makes Metaphysical Explanations *Metaphysical*?

Schaffer's challenge is to demonstrate that there are metaphysical explanations without resorting to any distinctively 'metaphysical' machinery such as grounding or essences. I resist this challenge and propose an explicit commitment to essentialist 'machinery'. But why should we attempt to be neutral about this machinery in the first place? Schaffer's motivation for offering a minimal or neutral account is presumably to avoid the complications that more specific proposals face and to show the general applicability of the notion of a law of metaphysics. Schaffer (2018: 2) lists some candidate cases of the relevant *non-causal explanatory connections*, which are not particularly surprising: they rely on specific metaphysical principles concerning things like truthmaking, the determinate/determinable distinction, the truth-conditions of disjunctions, set membership, and so on. By now, most readers are surely familiar with such paradigm cases of 'because' that are typically discussed in the grounding literature, so I will not spend time in presenting these cases. The important point is that any explanations of this type have what Schaffer calls a 'metaphysical flavor', and he specifies: these cases 'have the feel of concerning the constitutive generation of a dependent outcome' (2018: 3).

This an important point and it is related to the discussion in the previous section: what is responsible for the 'metaphysical flavour' is some dependence relation that 'backs' the relevant metaphysical explanation. Schaffer (2018: 12) would seem to agree on this point, as he also cites Kim's famous account of explanation. Now, as Schaffer acknowledges, this much is compatible with a type of *grounding pluralism*, such as Jessica Wilson's (2014) 'small-g' grounding relations (e.g., composition and set membership) and presumably also Kathrin Koslicki's (2015) approach. Schaffer thinks that the grounding pluralist as well can accept his entire argument for laws of metaphysics, which suggests that there must be something that unifies the 'small-g' grounding relations as well. But if that's the case, then the whole point about laws of metaphysics seems to be entirely *terminological*: if the existence of worldly, non-causal dependence relations that back explanation is postulated, then laws of metaphysics do not do any additional work here, much like Wilson's original case against 'big-G' 'Grounding' suggests in the case of grounding. In Wilson's case, the point is that we do not need to postulate a novel 'Grounding' relation that is operative in the various cases of metaphysical dependence, because we already have the 'small-g' grounding relations, i.e., the specific metaphysical dependence relations. In the present context, connecting these specific dependence relations with laws of metaphysics does not tell us anything about how to understand the relevant dependence relations themselves or what, if anything, unifies these dependence relations as the ones that back metaphysical explanations. So, I really don't think that this is going to be enough for any serious proponent of laws of metaphysics who hopes to unify

explanation—recall that this was supposed to be one of the key motivations for postulating laws of metaphysics.

Schaffer's account is supposed to be neutral with regard to grounding or essences, but he does think of laws of metaphysics in terms of grounding, and he would say that: 'a law of metaphysics is a counterfactual-supporting general principle about what grounds what' (2018: 6). So, he can perhaps salvage the account from this objection, but then it won't be neutral anymore. This is not a problem in its own right, but does mean that one of Schaffer's original motivations for postulating metaphysical laws seems to be undermined. The problem is that there are competing accounts of what, in general, supports counterfactual generalisations, which brings us to (2).

3.2. Unification via General Essences

I agree with Schaffer that the unificatory role of metaphysical explanation is important, just like it is important to unify scientific explanation. The thought here is simple: we should strive to find the lowest common denominator, since our explanatory endeavours can be simplified if two distinct phenomena share the same or similar basis. But why should we need laws of metaphysics to do this? My own view is that *general essences*, such as natural kind essences (as opposed to *individual essences*), can do the job here.[4] It is worth mentioning that there are also essence-based accounts of laws of metaphysics, such as Rosen's, where it lies in the *nature* (or essence) of the grounded fact to be grounded in a certain way (2017: 285). So, on Rosen's account, it is something about the nature of the grounding relation that does the unifying:

> The plausible claim is that just as it lies in the nature of [$p \vee q$] to require either [p] or [q] as a ground, so it lies in the nature of [[p] grounds [$p \vee q$]]—and in particular, in the nature of the grounding relation itself—that facts of *this* sort need to be grounded in [p] together with an essentialist principle saying what grounds what. In a resonant slogan: It lies in the nature of metaphysical ground that particular grounding facts are always grounded in the grounds plus grounding laws (Rosen 2017: 285).

Contra Schaffer, Rosen contends that the relevant counterfactual-supporting general principle about what grounds what is an *essentialist* principle. But one might nevertheless think that Schaffer's grounding-based approach and Rosen's essentialist approach toward laws of metaphysics are on a par since they both rely on some further ontological elements to determine 'what grounds what' (despite Schaffer's attempts to remain neutral). However, I think that there is a type of category mistake looming in both suggestions. In fact, Rosen (2017: 284) even responds to such an accusation of a category mistake, concerning the idea that a law (of metaphysics) could figure along with [p] as part of the ground for [$p \vee q$]. Rosen insists that a general grounding law, say, about the nature of disjunction, can indeed be part of the grounds.

[4] A general essence explains why an entity is of this rather than that kind, but does not distinguish entities of the same kind, that is, all members of a given natural kind would share the same natural kind essence. Abstract objects like sets can also have general essences.

My worry is slightly different though, which is why it applies to both Schaffer and Rosen: why should we require any further principle—a law of metaphysics regarding 'what grounds what'—to secure the dependence between the *explanandum* and the *explanans*? One reason to be wary is that introducing a further fact about 'what grounds what' into this equation would itself seem to require an explanation, threatening infinite regress. But if we follow the simple idea that explanation tracks dependence, we have already given the whole story by the time we have identified what the relevant dependence relation and its relata are. In the example at hand, this appears to be relatively simple: the relata are [p] and [$p \vee q$] and the relation is presumably logical consequence (or logical dependence): if [p] is true then [$p \vee q$] is true. It is true that we can say of this relation that it holds in virtue of the nature of disjunction and in this sense that nature or essence contributes to the overall explanation. But there is no reason to think that the full explanation requires any additional 'grounding law' or law of metaphysics over and above the laws of logic or logical necessities which are true in virtue of the natures of all logical entities (cf. Fine 1994: 9–10).[5] So, on this view, the *modal force* and counterfactual robustness of generalisations involving logical constants like disjunction can be traced to the essences of these entities. More precisely, these *kinds* of entities, namely logical constants, have a *general essence* which gives rise to logical necessities.

Admittedly, Rosen's view need not differ very radically from the account I am proposing here. He does hold, like I do, that the answer to the question of why [p] grounds [$p \vee q$] must be that: 'it lies in the nature of disjunction that disjunctions are grounded in their true disjuncts' (Rosen 2017: 291). Moreover, he thinks that this explanation is an ultimate explanation in the sense that Glazier (2017: 2878) specifies, namely, that's where the explanation ends.[6] But consider Rosen's concluding passage:

> In many cases, if you want to know what grounds some particular fact [Fa], the answer is that [Fa] obtains in virtue of prior particular facts [$\varphi(a)$] together with a general law to the effect that whatever φs is thereby F (Rosen 2017: 289).

Now, the question that we need raise here is: what grounds that general law that whatever φs is thereby F—or better: what gives this general law its modal force (thereby making it a law)? In my view, the answer must be given in terms of the essences of the participating entities, e.g., it is part of the essence of entities of a given natural kind that they behave in a certain way. But once we have established this, we have no need to refer to a general, metaphysical law. Accordingly, it might be best to describe the essentialist approach that I favour as an *outlaw*, or a 'lawless' position (cf. also Mumford 2005).

So, I do think that it is a mistake to succumb to talk about 'laws of metaphysics', 'grounding laws' or 'general laws' in this connection or indeed to talk about the nature of metaphysical ground itself. For all we need here is the relatively

[5] There are further questions about the nature of logical consequence. For an interesting take on logical consequence and ground, see Schnieder 2018.

[6] Compare this to the debate about whether there are any laws of nature in the dispositional essentialist and powers literature: Stephen Mumford (e.g., 2005) argues in favour of 'lawlessness', i.e., the idea that powers do all the work that laws are usually postulated for, whereas Alexander Bird (e.g., 2007) defends the idea that once we have all the powers, we get the laws for 'free'. (Thanks to Toby Friend for suggesting this.)

familiar picture about essence as a basis of modal truths (as specified, e.g., in Fine 1994, Lowe 2008, and Tahko 2023a), applied to the case of metaphysical explanation understood as tracking dependence relations. This leads us to (3), which concerns other 'paradigm cases' of laws of metaphysics.

3.3. Paradigm Cases of Metaphysical Laws

Let's consider the case of *set formation*, which is indeed a very paradigmatic case. Set formation is, for Schaffer, one of the clearest cases of a law of metaphysics:

> [I]n order to explain the existence of {Socrates} from the existence of Socrates, the principle of set formation is needed to give the connection. Without set formation, the existence of Socrates and the existence of {Socrates} are just two facts with no special connection, much less the kind of asymmetric dependence that backs explanation (Schaffer 2018: 13).

Well, this is true as far as it goes, but set formation (which Schaffer limits to the context of a hierarchical conception of sets, such as the one embedded in Zermelo–Fraenkel set theory) is a very specific operation and I struggle to see what it has in common, say, with the case of disjunction discussed above, or the case of determinable/determinates. Yet, if laws of metaphysics are supposed to unify explanation, then one might think that they should together form a unified basis—similarly, many accounts of the metaphysics of laws of nature seek to find a unified basis for laws, e.g., based on powers or dispositional properties. The grounding pluralist would here point out that there are several distinct dependence relations in effect in these cases, so trying to find a single relation that unifies the cases is doomed. With some reservations, I am inclined to agree. However, building on the previous discussion regarding disjunction, we have a rather easy solution available. The solution is that just like logical constants can be regarded to have a general essence, so can sets. Indeed, any entity, be it abstract or concrete, has a general essence, which expresses the identity and existence conditions of the type of entity in question (see Tahko 2018, 2023a for further discussion). This line of thought follows an essentialist picture that is familiar, e.g., from E.J. Lowe's:

> Consider the following thing, for instance: the set of planets whose orbits lie within that of Jupiter. What kind of thing is that? Well, of course, it is a set, and as such an abstract entity that depends essentially for its existence and identity on the things that are its members—namely, Mercury, Venus, Earth, and Mars. Part of what it is to be a set is to be something that depends in these ways upon certain other things—the things that are its members. Someone who did not grasp that fact would not understand what a set is (Lowe 2008: 37).

More specifically, as I have argued elsewhere (Tahko 2018: sec 2.2.2), it is plausible that on the type of hierarchical conception of sets that we are here operating with, the set-theoretical hierarchy has an implicit modal character which is expressed by the general essence of sets. This modal character is in fact already present in the above quote from Lowe, as he specifies that sets *essentially depend* for their existence and identity on their members. Now, if this conception captures the general essence of sets, then in order to explain the existence of {Socrates} from the existence of Socrates we only need to understand that {Socrates} *is a set* and hence it essentially depends on Socrates for its existence and identity. In other

words, the general essence of sets imposes modal constraints and determines the relevant asymmetric dependence that backs explanation in cases involving sets.

Here we have the makings for a unified account of metaphysical explanation without any extra laws about 'what ground what': we simply need to recognize the role of general essences in establishing the relevant modal elements that secure the dependence and hence counterfactual robustness between the *explanandum* and the *explanans*. I suppose that one may call these essentialist truths 'laws of metaphysics' (or 'essentialist laws', as Rosen 2017: 291 seems to do). But I do not think that this is ideal since they do not have the structure of laws as we usually understand them. Admittedly, sometimes it is suggested that statements like 'all electrons have unit negative charge' express laws, but my reaction to this is very similar: these are truths about the general essences of entities and their modal implications.[7]

At the outset, I promised a unified account of metaphysical and scientific explanation, and we are not there yet. So, let us now move to some more scientifically-motivated cases and see if the same picture can be applied in that context.

4. Grounding Mechanisms and Scientific Explanation

Even if the reader is happy to follow me to the realm of essentialist explanation, it may appear that it must come with the cost of abandoning any hope of unity between scientific and metaphysical explanation. After all, the helpful analogy between these types of explanation was supposed to be based precisely on laws of metaphysics that correspond to causal laws and I have suggested that we do not need to appeal to laws of metaphysics to secure metaphysical explanation. I would now like to take a closer look at this analogy between scientific and metaphysical explanation in order to see if we can make some progress.

One promising route for laying out the analogy (or more than just an analogy) between scientific and metaphysical explanation is to consider cases of scientific explanation that appeal to causal mechanisms, as suggested by Trogdon (2018). The idea is that there are grounding explanations that are analogous to causal-mechanical explanations in science. These would be metaphysical explanations that appeal to *grounding mechanisms* or as Trogon calls them, *grounding-mechanical explanations*. So, the role that grounding mechanisms play in certain metaphysical explanations mirrors the role that causal mechanisms play in certain scientific explanations. Trogdon (2018: 1290) pitches this approach as different from Schaffer's and Alastair Wilson's, who both suggest that just like causal relationships, grounding relationships as well can be represented by directed graphs.

Trogdon also discusses cases such as set formation and the determinate-determinable relation and takes it that these are metaphysical determination relations, and that it is an essential truth about these relations that they stand in the relevant grounding relationships (e.g., it is part of what it is to be set formation that the existence of the members of a set ground the existence of the set). But we have already discussed cases of this type, so let us focus on the more original part of Trogdon's proposal. This concerns cases where 'the corresponding grounding facts aren't enough on their own to ground what they ground—they're mere

[7] I am uncertain about how exactly this lines up with Kit Fine's (2015) views about the unified foundations for essence and ground, but it seems to me that what I propose is not too far apart from the Finean picture. (Thanks to Sam Kimpton-Nye for highlighting this potential connection.)

partial grounds' (Trogdon 2018: 1291). Trogdon gives three candidate relations that involve grounding-mechanical explanation: constitution, functional realization, and mereological realization. I will focus on the last of these, partly because of Trogdon's choice of example, which makes for some interesting discussion.

Here is Trogdon's example in more detail:

> Mereological realization: part of what it is to be mereological realization is that if the Ps (e.g. certain molecular properties) stand in this relation to Q (e.g. the property of being hard) on an occasion such that the xs have the Ps, y has Q, and the xs compose y, then the fact that the xs compose y and have the Ps is among some plurality of facts that grounds the fact that y has Q (e.g. the fact that the xs compose y and have thus-and-so molecular properties is among some plurality of facts that grounds the fact that y is hard) (Trogdon 2018: 1292).

A little later, Trogdon (*ibid.*, 1297) applies this case to a cut diamond's hardness and proposes that the fact that a diamond is hard is partially grounded in the fact that its constituent carbon atoms are bonded and spatially arranged in a specific way. This grounding connection can then be modelled in terms of a grounding mechanism involving mereological realization (as in Gillett 2007) and the idea that causal powers (such as the diamond's hardness) are constituted by other causal powers. The resulting model of the relevant grounding relations is simple enough (Trogdon 2018: 1298). The diamond is composed of carbon atoms, which have certain properties, such as being bonded and spatially arranged in a specific way. These properties constitute the grounding fact and bestow causal powers to the diamond's constituent carbon atoms. The two crucial assumptions here are the following:

(1) The property of being hard is a constituent of the grounded fact (that the diamond is hard), and it is individuated by the causal powers that it bestows to the diamond.

(2) The causal powers of the carbon atoms consist of the causal powers bestowed to the diamond.

In purely philosophical terms, it is perhaps a controversial assumption that we can individuate properties like being hard in terms of the causal powers that they bestow to the thing that they are properties of (a view going back at least to Shoemaker 1980). But we can set this philosophical concern aside, because there is a more interesting issue underlying this example. This issue concerns the property of *being hard* more generally.

4.1. The Case of Hardness

Hardness is an interesting property. It can be measured by a scratching test, so a material's hardness can literally be measured in terms of its resistance to scratching by another material. Hence, in this case the property of hardness is effectively individuated in terms of the causal power to *resist* scratching. However, it also seems that this is not what hardness really *is*, i.e., it is not just the power to resist scratching—hardness can manifest in other ways as well, such as by resisting compression (or indeed not manifest at all), so it is at least multiply realizable in this sense. This may lead one to think that, say, the hardness of a diamond should really be conceived of in terms of its carbon microstructure, i.e., whatever realizes its hardness. Why should we think that hardness is anything over and above the

causal powers of the carbon microstructure? In other words, to what extent, if at all, should hardness be conceived of as a *real property* with causal powers, distinct from the powers that the carbon atoms in a specific configuration possess? This is an issue that the mereological realization model as presented above does not seem to directly address. Accordingly, the case calls for further analysis. I will suggest that it fits the pattern of a typical essentialist explanation, which does not require any further laws or a general principle in addition to the relevant general essences. But before we get there, we need to consider some further scientific detail.

As it happens, Carl Gillett (2016: 65–9) has also discussed the case of diamonds and carbon atoms as an example of compositional explanation. The case is precisely that of the diamond's hardness causing a scratch in a medium, which is glass in Gillett's example. Gillett's framework is very rich and complicated, and I cannot discuss it here in detail, but he does have something interesting to say about the question I have just raised, namely, the individuation of the relevant property of hardness and the causal powers that are bestowed to the diamond. Here is what Gillett (2016: 69) proposes: '[H]ardness and diamonds, and carbon atoms and their properties/relations, are each partially individuated by the processes that result from them.' I take it that Gillett here means 'ontologically individuated', rather than just epistemically individuated.[8] So, on Gillett's line of thought, it would seem that hardness is partially individuated by the diamond's ability to scratch glass. But in order to give a full account of what hardness really is, we will presumably have to see what other work it can do as well, and what other processes it can be involved in. However, it would clearly be hopeless to try to give a comprehensive list. Instead, I would like to borrow Mark Wilson's (2006: Ch. 6) detailed analysis of hardness and its history. He also provides a splendid diagram (Wilson 2006: 338; I will not attempt a reconstruction here) of the vast variety of different tests for hardness, of which the scratch test as applied to diamonds is merely one of many examples. This poses a further challenge for the analysis of hardness: given that it comes in a variety of very different guises, is there any plausible way to unify the phenomenon?

Even without discussing the various examples of hardness tests in any detail, we can quickly see that if we wish to (partially) individuate the property of hardness in terms of the processes that it is involved in, we will be at it for a very long time. Worse, it is not at all clear that the resulting property of hardness can be sensibly thought to be a singular property or power at all. This suggests that we would seem to need a very long list of general principles or laws of metaphysics to account for hardness, which may be taken to speak against their generality in the first place. To take one example, when we talk about the 'hardness' of certain types of plastic, it turns out that a Brinell-type 'squeeze and release' test often applied to metals will not be very useful, since plastics also have viscoelastic properties that cause the size of the indentation resulting from the test to decrease over time. These issues can have rather extreme results: 'If we followed the usual standards for the hardness of a steel, ordinary tire rubber would prove to be rather "harder" than cold-worked steel' (Wilson 2006: 339). The upshot is that we may not be able to individuate hardness, even partly, in terms of the processes that it

[8] Gillett is careful to distinguish between 'internal' and 'ultimate' ontology: 'Work in ultimate ontology seeks to articulate what entities there are in the world, including the relations between them. In contrast, internal ontology simply seeks to articulate the ontological posits of certain scientific products (Gillett 2020: 33).

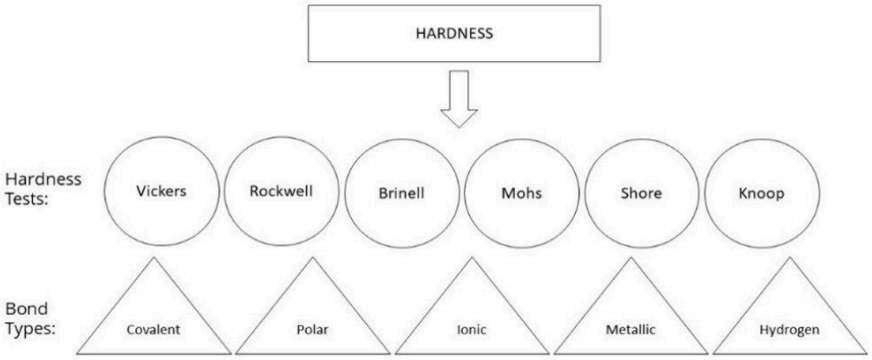
is involved in; there are just too many, and too varied, processes. So, what is *hardness*? This passage from Sidney Avner's *Introduction to Physical Metallurgy*, (also quoted in Wilson 2006: 341–2) is telling:

> The property of "hardness" is difficult to define except in relation to the particular test used to determine its value. It should be observed that a hardness number or value cannot be utilized directly in design, as can [yield value], since hardness numbers have no intrinsic significance. Hardness is not a fundamental property of a material but is related to the elastic and plastic properties. The hardness value obtained in a particular test serves only as a comparison between materials or treatments (Avner 1974: 24).

The message is clear: hardness is not a fundamental property if it is a *property* at all. This by itself does not undermine the attempt to give a grounding-mechanical explanation of, say, the diamond's hardness, since it was suggested that the diamond's hardness is grounded in the properties of carbon atoms arranged in a specific way. But if we now say that, correspondingly, the hardness of a plastic or a metal will be grounded in the properties of their constituent atoms arranged in a certain way and propose that this is a unified grounding-mechanical explanation of the property of hardness, then I think that we have gone astray. For one thing, the constituent atoms of these other materials are arranged in very *different* ways and have different bonds that underlie the relevant properties of the material. Moreover, the tests that we use to measure their hardness are also different. I do not believe that there is a useful analogy between metaphysical explanation and causal explanation on offer here; certainly not on the basis of this example. In other words, we have not yet found anything sufficiently *general* in order to put forward an analysis of hardness that would be in line with typical examples of metaphysical explanation.

However, I do think that Gillett and Trogdon are both onto something important. Even if the property of hardness turns out to be multiply based in a very messy way or indeed 'wildly disjunctive' (cf. Kim 1992: 10), it does not of course mean that the relevant causal powers would not be grounded in *something*. In other words, even if the reductive base of hardness is disjunctive, there may still be a unified account of hardness available. Yet, we are certainly not going to find a unified account of hardness at the level of carbon (or other) microstructure. We should be looking deeper. What I have in mind is that just like in the case of the earlier examples drawn from more abstract contexts, such as the case of sets, we are going to need to find the relevant existence and identity conditions—the general essence—of the kind of entity in question. So, once again it turns out that the story about the underlying general principles cannot be given without reference to some further metaphysical machinery, i.e., general essences. But since hardness appears to be shared by a vast range of different macrophysical objects, the only hope for a unifying the phenomenon would have to be something that is shared by the different realizations of this macrophysical property. What could this possibly be, and can we really find an essentialist explanation here?

Fortunately, we have learned quite a bit more about the chemistry and physics of hardness since Avner's 1974 book (see, e.g., Gilman 2009). One thing seems clear: hardness is a property that is only associated with collectives of atoms and molecules. Just like properties such as transparency or diffraction are properties that only collectives of, say, water molecules have (see Tahko 2021: 62). We already noted that different varieties of hardness are also realized by a variety of different chemical bonding mechanisms: covalent bonds, ionic bonds, polar bonds, metallic bonds, hydrogen bonds. So, we need to go even deeper to find anything in common. This is exactly what the research from the last few decades has accomplished (see, e.g., Gao et al. 2003, and Šimůnek and Vackář 2006). Fortunately, it is not in fact very difficult to find something in common for all these different cases of bonding, for they all involve the *electromagnetic force*, which can be conceived as the manifestation of the property of *electric charge*. Ultimately, it is the electromagnetic force that holds atoms and molecules together, so in this sense it is also responsible for any 'repulsion' that a macrophysical material manifests in the case of a Brinell test, a scratching test, or indeed any manipulation of a material that we might employ as a test for hardness. We can describe electromagnetic interaction via Coulomb's Law and The Lorentz Force Law, which summarises the effect of the electric force and the magnetic force. The technical details are beyond the scope of the present paper, but a simple illustration might help:



$$(1)\quad H(\mathrm{GPa}) = 350[(N_e)^{2/3} e^{-1.191 f_i}]/d^{2.5}$$

$$(2)\quad H = \frac{C}{\Omega} n \left[\prod_{i,j=1}^{n} N_{ij} S_{ij} \right]^{1/n} e^{-\sigma f_e},$$

$$f_e = 1 - \left[k \left(\prod_{i=1}^{k} e_i \right)^{1/k} \Big/ \sum_{i=1}^{k} e_i \right]^2$$

Figure 1: Hardness tests, bond types, and first principles calculations.

Figure 1 outlines six different hardness tests and five different chemical bond types. I make no attempt to match all of these, but to illustrate, the Vickers and Knoop hardness tests use a diamond indenter in the shape of a pyramid; the Vickers test can be used for all metals whereas the Knoop test is often used for brittle materials.

Depending on the material, different bond types come into play. In *Figure 1*, equation (1) is a calculation representing the hardness of an overly covalent crystal, originating in Gao et al. 2003 and picked up by Šimůnek and Vackář (2006), where N_e is the electron density expressed in the number of valence electrons per cubic angstrom, d is the bond length in angstroms, and f_i is the ionicity of the chemical bond in a specific crystal. Equation (2), from Šimůnek and Vackář 2006, is a generalised equation to calculate the hardness of more complex crystals than binary compounds. In (2), we see a system with n different binary systems described by bond strengths S_{ij} derived from the energies e_i, e_j, where N_{ij} is the number of the binary system ij, and k corresponds to the number of different atoms in the system. These recent developments are important because the experimental hardness tests are in fact fairly inaccurate:

> In principle, hardness should be related to crystal orientation. However, during the indentation, the force of the diamond wedge is diverted sideways, so the sample is subjected to a combination of stresses—compression, shear, and tension in various directions. Consequently, the anisotropic effects are reduced. Additionally, the strength of shear or tension of a sample is highly dependent on the presence of defects in the sample. As a result, experimental values of hardness can vary by more than 10% for the same sample (Šimůnek and Vackář 2006: 1).

We do not need to go into more technical detail than this. What is important is that the *first principle calculations* that equations (1) and (2) are based on represent a method to calculate physical properties directly from basic physical quantities such as mass and charge, Coulomb force of an electron, and so on. So, hardness is indeed not a fundamental property of materials. But it is, ultimately, based on bond strengths and other measurable properties (and the laws that govern them), of which electric charge is the most obvious candidate for a fundamental property.

While this explanation doesn't necessarily undermine a grounding-mechanistic account, it's clear that the source of the explanation is not available just 'one level down' from hardness. Rather, all we have here—all we need—is the fundamental property of electric charge possessed by (presumably) fundamental natural kinds such as fermions. This is precisely what we should expect on the essentialist line: we have successfully reduced the various dis-unified higher-level explanations to fundamental natural kinds whose general essences ultimately constrain all the phenomena that we typically capture under the label of 'hardness'. Let us now take a step back and look at the broader picture and its applicability.

4.2. Reductionist-Essentialist Explanation

The plausibility of the grounding-mechanistic account depends on whether or not it is compatible with the account that is now starting to emerge, call it 'reductionist-essentialist' explanation. Much more work remains to be done for us to be able to calculate a given type of hardness for a given material, but there is already ample evidence that this can be done, and the first principles calculations mentioned above also appear to be more accurate than any of the mechanical hardness tests developed. There are several ways that all this can be spelled out and of course the jury is still out there regarding some aspects of the fundamental forces that are involved in this story. But one, albeit crudely simplified, way to go would be to say that it is the dispositional essence of charge that is ultimately responsible

for the disposition of hard materials to resist scratching or whatever test we might invent for hardness.

The upshot of this type of account is that we can indeed unify scientific and metaphysical explanation because laws of nature and 'laws of metaphysics' may both be analysed in the same way (since the dispositional essentialist explains laws of nature in terms of essential properties). There are many proponents of the traditional dispositional essentialist view (e.g., Bird 2007), but in contemporary literature on dispositional essentialism some further variations have emerged. In particular, there are those who argue that (natural) properties like *charge* ground various dispositions, which may also open the door to versions of dispositionalism that do not rely on essences (Coates 2020, Tugby 2021, 2022, and Kimpton-Nye 2021).

However, my preferred strategy obviously relies on general essences, so let me attempt to formulate reductionist-essentialist explanation in more general terms, where we are interested in the behaviour of a given concrete entity *a* of kind *K*:

(I) Target of explanation: entity *a* of kind *K* has defining feature (or property/behaviour) *F*.

(II) Observation (empirical): having *F* is dependent on sub-feature (e.g., structure, another property or set of properties) *G*.

(III) General explanation: it is part of the *general essence* of entities of kind *K* that they depend on *G* for their existence.

(IV) Particular explanation: *G* necessitates *F*, so *a* has *F* because it is of kind *K*, i.e., has the particular general essence that members of *K* have.

The case of hardness can be made to fit this picture fairly easily: a given diamond is hard because its constituent carbon atoms are bonded and spatially arranged in a specific way and (let us assume) it is part of the general essence of diamonds that their constituent carbon atoms are thus bonded. So, this particular diamond is hard because the structure of its constituent carbon atoms necessitates the hardness of all diamonds. All the explanatory work is done by the kind membership (i.e., general essence of the kind) and the relevant dependence relation. It is worth noting that this dependence relation is plausibly 'internal', i.e., it holds necessarily given the existence of its relata—so it is not an additional 'element of being' or indeed a law of metaphysics.[9]

Can we find other good examples besides the case of hardness? Yes, but as with other cases of purportedly *reductive* explanation, like reductionist-essentialist explanation clearly aims to be, it can be laborious to provide sufficient scientific detail—this is stage (ii) of the general pattern presented above. Elsewhere Tahko 2023b), I have examined another case from physics, concerning the predicted stability of superheavy elements, i.e., elements with an atomic number greater than 103. The case of the yet to be synthesised element with atomic number 126, *unbihexium* is of particular interest. However, the fact that no samples of the element exist pose an interesting challenge: where does the empirical information required for stage (ii) come from?

The answer involves taking a close look at what Eugene Wigner coined the 'magic numbers': 2, 8, 20, 28, 50, 82, and 126. The numbers are based on

[9] For further discussion on relevant ontological dependence relations of this type, see Tahko and Lowe 2020.

combinations of protons and neutrons which appear to produce higher stability of the atomic nucleus (these are combinations of protons or neutrons arranged into complete shells within the nucleus). Now, fitting it into the above pattern, we might say that if the target of explanation is the predicted stability of element 126, then the relevant observation is that certain combinations of protons and neutrons produce a higher stability and we can predict this in the case of element 126 because it shares this structural feature with the already observed cases, e.g., calcium (Z = 20), which has two 'magical' isotopes, with neutron numbers 20 and 28. This gives us the general explanation: it is part of the general essence of atomic nuclei that their stability depends on a structure of binding energies and energy levels, giving rise to further dependencies involving the shell model of the nucleus. Accordingly, the structure of the shells influences the energy levels and ultimately determines the stability of the nucleus. There is obviously plenty more scientific detail that can be given about this case as well (see, e.g., Chapman 2020), but this brief overview should suffice to show that other candidate examples that fit the general pattern proposed above can be found.

5. Conclusion

The overall upshot of the paper is that we need not resort to talk of laws of metaphysics, even though metaphysical explanation can be regarded as a genuine form of explanation. Moreover, we can unify this metaphysical form of explanation and scientific explanation because they share the same ultimate basis, which on my preferred view are the general essences of the entities that these explanations concern. Alas, it is not my goal here to pursue these details. Instead, I conclude here, having provided what I promised at the outset: a (sketch of a) unified account of scientific and metaphysical explanation in terms of general (natural kind) essences (for further details, see Tahko 2021).[10]

References

Audi, P. 2012, "A Clarification and Defense of the Notion of Grounding", in Correia, F. and Schnieder, B. (eds.), *Metaphysical Grounding: Understanding the Structure of Reality*, Cambridge: Cambridge University Press, 101–21.

Avner, S. 1974, *Introduction to Physical Metallurgy*, New York: McGraw-Hill.

Bennett, K. 2017, *Making Things Up*, Oxford: Oxford University Press.

Bernstein, S. 2016, "Grounding is not Causation", *Philosophical Perspectives 30: Metaphysics*, 21–38.

Bird, A. 2007, *Nature's Metaphysics: Laws and Properties*, Oxford: Oxford University Press.

Chapman, K. 2020, "The Transuranic Elements and the Island of Stability", *Philosophical Transactions of the Royal Society A*, 378, 20190535.

Coates, A. 2020, "Making Sense of Powerful Qualities", *Synthese*, 198, 9, 8347-8363.

Dasgupta, S. 2017, "Constitutive Explanation", *Philosophical Issues 27: Metaphysics*, 74–97.

[10] I would like to thank Francesca Bellazzi, Toby Friend, Sam Kimpton-Nye, and Will Morgan.

Fine, K. 1994, "Essence and Modality", *Philosophical Perspectives: Logic and Language*, 8, 1–16.

Fine, K. 2012, "Guide to Ground", in Correia, F. and Schnieder, B. (eds.), *Metaphysical Grounding: Understanding the Structure of Reality*, Cambridge: Cambridge University Press, 37–80.

Fine, K. 2015, "Unified Foundations for Essence and Ground", *Journal of the American Philosophical Association*, 1, 2, 296–311.

Gao, F., He, J., Wu, E., Liu, S., Yu, D., Li, D., Zhang, S., Tian, Y. 2003, "Hardness of Covalent Crystals", *Physical Review Letters*, 91, 015502.

Gillett, C. 2007, "Understanding the New Reductionism: The Metaphysics of Science and Compositional Reduction", *Journal of Philosophy*, 104, 193–216.

Gillett, C. 2016, *Reduction and Emergence in Science and Philosophy*, Cambridge: Cambridge University Press.

Gillett, C. 2020, "Why Constitutive Mechanistic Explanation Cannot Be Causal", *American Philosophical Quarterly*, 57, 1, 31–50.

Gilman, J.J. 2009, *Chemistry and Physics of Mechanical Hardness*, Hoboken: Wiley.

Glazier, M. 2016, "Laws and the Completeness of the Fundamental", in Jago, M. (ed.), *Reality Making*, Oxford: Oxford University Press, 11–37.

Kim, J. 1992, "Multiple Realization and the Metaphysics of Reduction", *Philosophy and Phenomenological Research*, 52, 1–26.

Kim, J. 1994, "Explanatory Knowledge and Metaphysical Dependence", *Philosophical Issues 5: Truth and Rationality*, 51–69.

Kimpton-Nye, S. 2021, "Reconsidering the Dispositional Essentialist Canon", *Philosophical Studies*, 178, 3421–3441.

Kment, B. 2014, *Modality and Explanatory Reasoning*, Oxford: Oxford University Press.

Koslicki, K. 2015, "The Coarse-Grainedness of Grounding", *Oxford Studies in Metaphysics*, 9, 306–44.

Koslicki, K. 2016, "Where Grounding and Causation Part Ways: Comments on Schaffer", *Philosophical Studies*, 173, 101–12.

Kovacs, D.M. 2017, "Grounding and the Argument from Explanatoriness", *Philosophical Studies*, 174, 2927–2952.

Kovacs, D.M. 2020, "Metaphysically Explanatory Unification", *Philosophical Studies*, 177, 1659–1683.

Lowe, E.J. 2008, "Two Notions of Being: Entity and Essence", *Royal Institute of Philosophy Supplements*, 83, 62, 23–48.

Maurin, A.S. 2019, "Grounding and Metaphysical Explanation: It's Complicated", *Philosophical Studies*, 176, 1573–1594.

Mumford, S. 2005, "Laws and Lawlessness", *Synthese,* 144, 397–413.

Raven, M. 2015, "Ground", *Philosophy Compass*, 10, 5, 322–33.

Rosen, G. 2017, "Ground by Law", *Philosophical Issues 27: Metaphysics*, 279–301.

Schaffer, J. 2016, "Grounding in the Image of Causation", *Philosophical Studies*, 173, 49–100.

Schaffer, J. 2018, "Laws for Metaphysical Explanation", *Royal Institute of Philosophy Supplement*, 82, 1–22.

Schnieder, B. 2018, "On Ground and Consequence", *Synthese*, 198, 6, 1335–1363.

Shoemaker, S. 1980, "Causality and Properties", in van Inwagen, P. (ed.), *Time and Cause: Essays Presented to Richard Taylor*, Dordrecht: Reidel, 109–35.

Šimůnek, A. and Vackář, J. 2006, "Hardness of Covalent and Ionic Crystals: First-Principle Calculations", *Physical Review Letters*, 96, 085501.

Sjölin Wirling, Y. 2020, "Is Backing Grounding?", *Ratio*, 33, 129–137.

Skiles, A. and Trogdon, K. 2021, "Should Explanation Be a Guide to Ground?", *Philosophical Studies*, 178, 4083–4098.

Tahko, T.E. 2018, "The Epistemology of Essence", in Carruth, A., Gibb, S.C. and Heil, J. (eds.), *Ontology, Modality, Mind: Themes from the Metaphysics of E. J. Lowe*, Oxford University Press, 93–110.

Tahko, T.E. 2021, *Unity of Science, Elements in Philosophy of Science*, Cambridge University Press.

Tahko, T.E. 2023a, "Possibility Precedes Actuality", *Erkenntnis*, 88, 3583–3603.

Tahko, T.E. 2023b, "The Modal Basis of Scientific Modelling", *Synthese*, 201, article number 75.

Tahko, T.E. and Lowe, E.J. 2020, "Ontological Dependence", *The Stanford Encyclopedia of Philosophy* (Fall 2020 Edition), https://plato.stanford.edu/archives/fall2020/entries/dependence-ontological/

Thompson, N. 2016, "Metaphysical Interdependence", in Jago, M. (ed.), *Reality Making*, Oxford: Oxford University Press, 38–56.

Trogdon, K. 2013, "An Introduction to Grounding", in Hoeltje, M., Schnieder, B. and Steinberg, A. (eds.), *Varieties of Dependence: Ontological Dependence, Grounding, Supervenience, Response-Dependence (Basic Philosophical Concepts)*, München: Philosophia Verlag, 97–12.

Trogdon, K. 2018, "Grounding-Mechanical Explanation", *Philosophical Studies*, 175, 1289–309.

Tugby, M, 2021, "Grounding Theories of Powers", *Synthese*, 198, 12, 1118–11216.

Tugby, M. 2022, "Dispositional Realism without Dispositional Essences", *Synthese*, 200, 3, 1–27.

Wilsch, T. 2015, "The Nomological Account of Ground", *Philosophical Studies*, 172, 3293–312.

Wilson, A. 2018, "Metaphysical Causation", *Noûs*, 52, 4, 723–51.

Wilson, J. 2014, "No Work for a Theory of Grounding", *Inquiry*, 57, 535–79.

Wilson, J. 2016, "The Priority and Unity Arguments for Grounding", in Aizawa, K. and Gillett, C. (eds.), *Scientific Composition and Metaphysical Ground*, London: Palgrave-MacMillan, 171–204.

Wilson, M. 2006, *Wandering Significance: An Essay on Conceptual Behavior*, Oxford: Oxford University Press.

Understanding with Epistemic Possibilities: The Epistemic Aim and Value of Metaphysics

Ylwa Sjölin Wirling

University of Gothenburg

Abstract

According to a recent proposal, the epistemic aim of metaphysics as a discipline is to chart the different viable theories of metaphysical objects of inquiry (e.g. causation, persistence). This paper elaborates on and seeks to improve on that proposal in two related ways. First, drawing on an analogy with how-possibly explanation in science, I argue that we can usefully understand this aim of metaphysics as the charting of epistemically possible answers to metaphysical questions. Second, I argue that in order to account for the epistemic goodness of this aim, one should appeal to the epistemic value it has in virtue of providing resources for non-factive understanding of the objects of metaphysical inquiry.

Keywords: Epistemic possibility, Epistemic value, How-possibly explanation, Metaphilosophy, Non-factive understanding, Understanding.

1. Introduction

This paper takes off from two claims about metaphysics as a collective, epistemic endeavour. First, the familiar observation that metaphysics as a discipline is plagued by systematic, persistent disagreement between researchers who we take to be equally competent, applying the same methods, and who are all among the experts on the topic. I'll refer to this as Unresolved Dispute. Second, the decision to take seriously the fact that some instances of metaphysical inquiry and its products (e.g. metaphysical accounts or theories) are assessed *positively* by its own lights—i.e. in line with the norms and standards of epistemic assessment that apparently govern the discipline. I'll refer to this assumption as Successful Metaphysics.

Given what we may call the "standard view" of metaphysics' epistemic aim, the observation Unresolved Dispute and the assumption Successful Metaphysics are in tension. According to the standard view, the epistemic aim of

Argumenta 10, 1 (2024): 89—106
ISSN 2465-2334

DOI 10.14275/2465-2334/20240.wir
First published: 30 November 2024

metaphysics is to produce theories that provide knowledge of true answers to metaphysical questions, and the epistemic success of metaphysics is to be judged in relation to this aim. But Unresolved Dispute indicates that this aim is not being furthered by actual metaphysical inquiry. To those who wish to take seriously Successful Metaphysics, this suggests that the standard view is mistaken, and the norms and standards that govern metaphysics must flow from some different epistemic aim.

In this paper, I focus on a recent attempt to *reconceive* of metaphysics' epistemic aim in order to accommodate both Unresolved Dispute and Successful Metaphysics. According to this proposal, the aim of metaphysics is to chart the various tenable accounts of metaphysics' objects of inquiry (e.g. causation, modality, and so on). The aim of this paper is to develop and complement this proposal. In particular, I will raise what I call the Value Question for this "equilibrist" proposal and then sketch a two-part answer to that question. I will first suggest that we should understand this aim in terms of epistemic possibility: the aim is to construct and chart epistemically possible answers to metaphysical questions. Then I will argue that metaphysics, when understood in this way—along with other epistemic activities that have a similar character and function, including art interpretation and certain practices of how-possibly modelling in science—is epistemically valuable in virtue of providing resources for what I call *non-factive understanding* of the objects of (in this case, metaphysical) inquiry.

2. Background: Problems with the Standard View

It will be useful to begin by taking a look at what we may call the standard view on metaphysics' aim, and the problems that some philosophers see with it. I should note that the purpose of this section is not to argue conclusively that the standard view is untenable, but merely to show why one might be motivated to pursue an alternative view like equilibrism.

I take the standard view to be a claim about the aim of metaphysics as a discipline. What do I mean by that? The aim of a discipline is the central, or primary epistemic aim around which the discipline is structured, and in terms of which we can understand its epistemic norms, which practitioners are required to comply with *qua* metaphysicians, and the epistemic assessments and evaluations that are made in the course of metaphysical inquiry. It is in relation to this aim that the general state or shape of the discipline is judged. That is, whether metaphysics makes (enough) epistemic progress, or is in good epistemic shape, depends on whether actual metaphysical practice and its products relate appropriately to the discipline's epistemic aim.[1] The aim needs to be epistemic because metaphysics is supposed to be an *epistemic* activity, a form of inquiry.

We should recognise that the aim of metaphysics as a discipline may come apart from the aims that motivate individual metaphysicians to pursue their research. For instance, one may pursue metaphysics with the aim of achieving money or fame, but this does not make money or fame the aims of the discipline in question. This is of course not unique to metaphysics, but goes for other epis-

[1] Of course, it is not required that *all* instances of actual metaphysical practice, or all that comes out of it, are so related. Compare: not all instances of medical research, and not all results it produces, are in good standing by the discipline's own lights, but this doesn't licence rejection of the whole discipline as being in bad epistemic shape.

temic disciplines as well, and it applies also to the *epistemic* aims that individuals may have when pursing a particular type of inquiry. Consider Sheila who goes into medical research with the sole motivation of securing knowledge for herself of how ovarian cancer can be cured. She is content to terminate inquiry as soon as she has discovered the answer, whether or not her results are scientifically acceptable, or whether anyone else ever comes to know about them. This arguably does not make Sheila's knowledge the epistemic aim of (this branch of) medical research.[2]

Relatedly, since the general epistemic shape of the discipline is judged in relation to its central epistemic aim, we should also recognise that the discipline may produce or instantiate some epistemic goods without being in epistemically good shape. E.g. the fact that astrology has produced some knowledge (of various false theories describing how movements of celestial bodies influence human affairs, say) does not make astrology a discipline in good epistemic shape. Conversely, being in epistemically bad shape as a discipline does not imply being entirely devoid of epistemic value.

According to what many refer to as the "standard view" of philosophy, the aim of philosophy as a discipline—in the sense just outlined— is to uncover or attain knowledge of true answers to philosophical questions (see e.g. Brennan 2010; Chalmers 2015; Kornblith 2013; Stoljar 2017).[3] So, the epistemic quality of the discipline is judged by how philosophical practices and their products relate to this aim. Applied to metaphysics more specifically, the aim of metaphysics is, on the standard view, to uncover or attain knowledge of true answers to metaphysical questions (Kriegel 2013: 1; Paul 2012: 4). Metaphysical questions, as I understand them here, concern the underlying nature of the real world studied in science and everyday inquiry. The features of reality that metaphysical inquiry targets are normally prior to, more fundamental and more general, than those studied by the empirical sciences (Paul 2012: 5-6). They include the nature of modality, causation, property instantiation, mereological composition, change, and so on.

I will take the standard view to hold that metaphysics aims at knowledge that is publicly accessible, or somehow shared or collective. This means I will not count as proponents of the standard view the philosophers who claim that it is enough for philosophy to be in good epistemic shape that a few individuals secure knowledge of the true answers to philosophical questions (see e.g. Cappelen 2017; Keller 2017). Note that the corresponding view would also be implausible for other large-scale, collective inquiries such as biology, economics, or medicine.[4] This is not to say that some individual's knowledge of e.g. the true nature of causation is not epistemically valuable. Again, a discipline's producing some items of epistemic value is not sufficient to render it in epistemically good shape as a discipline.

[2] See Nado 2019 for an argument to the effect that knowledge is typically not the aim of professional inquiry such as science or journalism.

[3] Throughout this paper I focus exclusively on metaphysics, but much of what I will go on to say might apply equally well to philosophy more generally. But this is not a generalisation I will defend or make a point of here.

[4] See Dellsén, Lawler and Norton 2022 for recent discussion of conceptions of progress in science and how they relate to the notion of progress in philosophy.

The idea that knowledge is the aim of inquiry has a plausible ring to it. So, why would one seek out an alternative to this idea that the collective epistemic endeavour of metaphysics aims at some form of collective or publicly available knowledge?

One motivation comes from what I call Unresolved Dispute, namely the fact that metaphysics is plagued by systematic, persistent disagreement between researchers who we take to be equally competent, applying the same methods, and who are all among the experts on the topic. This is so even if we can all agree with Frances (2017) that philosophers agree on a great number of claims about reasons or arguments, e.g. of the form "problem *p* constitute a serious challenge to a theory *t"* or "*f* is a powerful reason in favour of *t".* Our very best research in metaphysics has failed to result in anything like convergence on what the truth is with respect to e.g. the nature of modality, causation, properties, and so on. Wildly different and mutually incompatible theories are more or less equally supported in the sense centrally relevant to the discipline, and are considered live options that metaphysicians may legitimately explore and defend.

Many prominent metaphysicians have, when pondering the subject, recognised that this situation is not going away: metaphysical disputes are not just unresolved, but in some sense irresolvable. Armstrong points out that the best one can do in metaphysics is to attempt to produce "visions (hopefully coherent) of the fundamental structure of the world, a vision that will compete with other visions", but that it is folly to expect to settle which one of these visions is the correct one (2010: 1). In a famous passage, Lewis similarly notes that

> when all is said and done, and all the tricky arguments and distinctions and counterexamples have been discovered, presumably we will still face the question which prices are worth paying, which theories are on balance credible, which are the unacceptably counterintuitive consequences and which are the acceptably counterintuitive ones. On this question we may still differ. And if all is indeed said and done, there will be no hope of discovering still further arguments to settle our differences (1983b: x).

That is, Unresolved Dispute is a well-known state of affairs—in metaphysics as in many other areas of philosophy.[5] As Dellsén et al. (forthcoming) correctly points out, whether Unresolved Dispute suggests that philosophy, or metaphysics in particular, is in bad epistemic shape depends on what one takes progress in the discipline to consist in—on what the aim of discipline is. But many philosophers have, *assuming the standard view,* argued that Unresolved Dispute suggests

[5] Stoljar (2017) claims that there is actually a lot of agreement on central issues in philosophy—that is, he denies Unresolved Dispute. His vindication of philosophical progress in the standard sense relies on his insistence that philosophical problems have a particular structure, an assumption that can certainly be questioned. Frances (2017) also presents an impressive list of substantial claims that philosophers have established and agree upon. For the purposes of this paper I don't need to deny that there is such agreement (also in metaphysics), and I think the value of those results can be accounted for by the account I will go on present later in this paper. But it is nonetheless true that wildly different theories of the same phenomenon are considered equally viable, and there is little reason to think we will be able to adjudicate between them.

that philosophy is in a bad epistemic shape (e.g. Goldberg 2009; Fumerton 2010; Brennan 2010; Kornblith 2013; Beebee 2018). I will not go into the details of these arguments here, and I do not suggest that they cannot be resisted or circumvented. Suffice it to say that the basic idea is that the persistent disagreement between metaphysicians on the answers to the questions that the discipline allegedly seeks to answer knowledgably, is at the very least a very strong reason to think that metaphysics has failed to produce much shared, collectively available knowledge of answers to those questions.

This has led several authors to a very pessimistic view of metaphysics. But there are reasons to resist this negative assessment of the discipline. Metaphysics, much like any other discipline of professional inquiry, is regulated by various epistemic norms for what researchers are obliged to do, and epistemic standards to which items such as theories, hypotheses, or claims are held. Among these norms and standards are practices for assessing theories and norms according to which theories with certain properties are to be considered better supported than others. It is reasonable to think that these norms and standards derive from the central epistemic aim of the discipline—that they are supposed to regulate inquiry in way that enables it to approximate or progress towards its aim.

In line with these norms and standards, several—but far from all—instances of metaphysical inquiry and its products, the metaphysical theories, are positively assessed. Metaphysicians regard the products of their epistemic practices—the metaphysical theories or accounts that are produced, scrutinised, refined, and sustained throughout the processes which constitutes metaphysical research—as thereby having received epistemic support in the sense(s) relevant to the discipline. Those who fail to live up to the standards are discarded along the way. But as already noted, the standards that regulate metaphysical inquiry consistently allow for mutually incompatible and wildly different answers to the same question to be positively assessed in the sense centrally relevant to the discipline. If the aim of the discipline, from which the norms and standards are supposed to flow, is knowledge—as the standard view has it—then we may conclude that these norms and standards are woefully insufficient. But we *could* instead decide to take the norms and standards seriously, as in fact managing to regulate metaphysical inquiry in a way that enables it to promote its aim. Then, however, we will need to reconceive of the aim of metaphysics as a discipline, finding one that *is* plausibly promoted by inquiry in line with the standards and norms that consistently fail to include tools for deciding which of a number of wildly different alternative answers is the correct one. A promising alternative account of metaphysics' aim is the *equilibrist account,* which the rest of this paper focuses on.[6]

[6] A common move is to argue that philosophers may still, in the face of systematic disagreement, be rational in believing their philosophical theories to be true (Kelly 2016; King 2012; Rotondo 2015) or else in holding some other belief-like propositional attitude towards their preferred theories (Barnett 2019; Goldberg 2013). This may or may not amount to a change in the view of philosophy's aim, as a discipline. As Beebee (2018) points out, even if individual rational beliefs/attitudes can be salvaged, and some of these turn out to be true, that does little to save epistemic face for metaphysics as a collective discipline given the standard view. This is not to deny that such states have epistemic

3. The Equilibrist Aim

A recent proponent of this alternative is Helen Beebee (2018). Drawing on methodological remarks by David Lewis, she proposes that we give up the standard view and instead see the aim of metaphysics as the endeavour to "find out what equilibria there are that can withstand examination" (Beebee 2018: 16; Lewis 1983b: x).[7] That is, the aim of metaphysical inquiry is not to produce theories that make available knowledge of whether Humean supervenience is true, under what conditions some parts compose a whole, or what the nature of property instantiation really is, but to chart the plurality of tenable answers to these questions. Gideon Rosen (2020) sketches a similar position which he calls fictionalism about metaphysics. For a fictionalist (or "agnostic", to use Rosen's alternative term), metaphysical inquiry is not a search for metaphysical truths but an "exercise in model-building" with the aim of constructing theories that meet certain constraints (2020: 41). There is, as far as success of the discipline is concerned, no need to settle on one theory but "the valuable intellectual work is done when the 'menu of well-worked-out theories is before us'" (44). The constraints in question that theories should meet are captured by what Rosen calls 'acceptability'. Acceptability in metaphysics consists in being "consistent with what we know in other areas" and satisfying certain other constraints that the discipline places on theories such as being "explicit, intelligible to us, explanatorily powerful, relatively complete, and plausible by our lights" (41-42).

I assume that Beebee and Rosen are describing basically the same view here: the aim of metaphysics is to map the space of tenable positions with respect to metaphysical questions, where tenability is understood as being internally coherent, exhibiting various explanatory virtues to a satisfactory degree, and fitting consistently with what we take ourselves to know. I will refer to its conception of metaphysics aim as "the equilibrist aim" in what follows (in line with Beebee's label "equilibrism"). The equilibrist aim seems promising given the objective to accommodate both Unresolved Dispute and Successful Metaphysics. That's because it effectively removes the conflict between the two: to have a plurality of competing accounts of the same phenomenon is just what we are aiming for (indeed, the more the better insofar as we want to map the complete space of constrained possibility), so Unresolved Dispute turns out to be *an important, central part of Successful Metaphysics*.

Nevertheless, the equilibrist proposal needs to be further developed. In particular, equilibrists need to address what I will refer to as the Value Question. The Value Question asks of any proposed epistemic aim for metaphysics as a discipline: what is the epistemic value of achieving or approximating that aim? On the one hand, the Value Question is interested exclusively in the *epistemic* value of metaphysics, and this is independent of whether it has or lacks other

value—as I said, metaphysics may fail to promote the central epistemic aim of the discipline while still producing epistemic goods.

[7] Beebee's thesis is put in terms of philosophy more generally rather than metaphysics specifically, but her motivation for reconceiving for philosophy's aim draws heavily on considerations about methodology in metaphysics.

types of value (e.g. aesthetic, practical, moral).[8] On the other hand, the Value Question is not interested in *any* epistemic value of metaphysics but only in the epistemic value it has in virtue of, or insofar as it, promotes or approximates its aim as a discipline. As already noted above, metaphysical inquiry may have epistemic value—for instance by resulting in some individuals acquiring rational beliefs or becoming better at logical thinking, or by being a type of process that has intrinsic epistemic value[9]—while being in bad epistemic shape, i.e. while consistently failing to promote or approximate its aim. The Value Question is a question about the specifically *epistemic* value of *successful* metaphysics (as defined in terms of its central aim).

As noted, the Value Question can be posed to any view of metaphysics' aim. But with the standard view, it is easier to see what the answer will be: knowledge is a paradigmatic example of something epistemically valuable.[10] With respect to the equilibrist aim, the question is what epistemic good(s) we are securing when we are managing to map the tenable accounts of e.g. causation or property instantiation. This is much less obvious.

I will spend the rest of this paper outlining a proposal according to which successful equilibrist metaphysics is valuable because it creates *resources for non-factive understanding of metaphysics' subject matter*, and such non-factive understanding is epistemically valuable.

4. Metaphysics as Epistemic How-Possibly Explanation

Let me start with what might seem like a detour through the philosophy of science. A scientific practice known in the philosophical literature as *how-possibly explanation* (HPE) has lately attracted quite a bit of attention from philosophers. Scientists in a wide variety of fields engage in something like providing explanations—typically using scientific models—that are not understood as *actual* (i.e. true) explanations of the relevant phenomena, but as *possible* explanations. There is no consensus on how this practice is best analysed (see Verreault-Julien 2019 for a useful overview), but arguably in many (but not all) of these cases the relevant possibilities are supposed to be *epistemic*—that is, they are supposed to be explanations of actual phenomena that might be correct as far as current scientific knowledge is concerned (Bokulich 2014; Brandon 1990: 179; Salmon 1989: 137; Sjölin Wirling and Grüne-Yanoff, forthcoming).

A good example is Alisa Bokulich's (2014) analysis of how scientists approach the phenomenon of tiger bush. Tiger bush is the phenomenon where vegetation in semi-arid areas grow in stripes, separated by barren areas, forming a pattern reminiscent of that in the tiger's fur. Scientists do not know exactly what causes this self-organizing pattern formation. Thus, in their research they construct models—Turing models, kernel models, differential flow models—that all support possible explanations of tiger bush, in the sense that they are all

[8] McSweeney (2023) also considers what she calls the Value Question for metaphysics, but it is less obvious whether she has in mind epistemic value specifically. Likewise for Rosen's (2020) discussion of value.

[9] For a proposal along the latter lines, see Sjölin Wirling 2021.

[10] It is then a separate question whether metaphysical inquiry can deliver knowledge, of course.

compatible with current scientific knowledge, and none of them can be ruled out as *not* the actual explanation.

In short, constructing and charting epistemically possible explanations is considered a legitimate and epistemically valuable research activity in sciences like biology, physics, economics, and so on. We now come to the reason for this detour: I suggest that we should understand equilibrist metaphysics as the constructing and charting of epistemic possibilities too.[11]

This is to an extent already present with Rosen's characterisation of what it takes for a metaphysical model to be acceptable: it must be "consistent with what we know in other areas". The idea that successful metaphysical inquiry teaches us about epistemic possibilities is also floated by Michaela McSweeney (2023). She considers a proposal according to which "we are never really justified in believing any particular metaphysical thesis but […] we still learn things about the world […] for example, that the world might be like *p,* for some *p*", and that one important function of argumentation in metaphysics is to remove obstacles to seeing that a particular view "might be true", i.e. is consistent with other things we take to be true.

I say "a kind" of epistemic possibility because, while being epistemically possible roughly amounts to be compatible with what we know, there are several ways in which epistemic possibility can be defined. The truth-value of a claim of the form "*p* is epistemically possible" depends on how a number of moving parts are fixed: whose knowledge and at what time, what does it take to count as part of the relevant knowledge corpus, and what does it take for *p* to be "compatible" with that corpus (Sjölin Wirling & Grüne-Yanoff, forthcoming). These moving parts can be fixed in different ways to generate different notions of epistemic possibility, and different ones are likely useful in different contexts.

Engineering a concept of epistemic possibility that is useful to metaphysical inquiry is far beyond the scope of this paper. But I will offer three preliminary thoughts on what it might look like, drawing on other metaphysicians' remarks about methodology. First, regarding the relevant corpus—the "what we know in other areas": this will arguably include knowledge of how the world undoubtedly appear to us through experience. Such knowledge has often been said to constrain metaphysical theorising, by providing the "data" that all theories must account for, both in the sense that it *prompts* metaphysical inquiry and that the resulting theories must face its tribunal when their viability is to be determined. Whitehead's description of metaphysical theorising using the metaphor of an airplane which start and lands on "the ground of empirical generalisation" after having taken a "flight in the thin air of imaginative generalisation" in *Process and Reality* (1929: 5) is a nice example of this.[12] Another example is Armstrong's talk of "Moorean facts": [13]

[11] This analogy with HPE-modelling is one way to elucidate the claim that metaphysics is usefully seen as *modelling,* which has been explored in rather different ways by Godfrey-Smith (2006) and Paul (2012). The proposal I offer has most in common with Godfrey-Smith's route, but is different in my explicit appeal to *epistemic* possibility, and in how I elucidate the epistemic goods—the understanding—afforded by the how-possibly "modelling" below.

[12] See Simons 1998 and Maurin 2002, chapter 3, for clarifying summaries of Whitehead's methodological stance.

[13] For a nuanced critical discussion of the Moorean approach, see Rinard 2013.

> the fact of sameness of type is a Moorean fact: one of the many facts which even philosophers should not deny, whatever philosophical account or analysis they give of such facts. Any comprehensive philosophy must try to give some account of Moorean facts. They constitute the compulsory questions in the philosophical examination paper (1980: 442).

In addition to the Moorean facts of common sense, the claims of a viable metaphysical theories plausibly need to be compatible with well-established scientific knowledge.[14] Second, the relevant sense of "not ruled out by" or "compatible with" will need to be fairly liberal, i.e. many different metaphysical accounts will be compatible with what we know. This is suggested by the fact that wildly different theories of the same thing are considered viable, but a nice way to further highlight it is to consider Lewis's poignant observation in response to Armstrong's complaint that Ostrich nominalism fails to account for sameness of type:

> Not every account is an analysis! A system that takes certain Moorean facts as primitive, as unanalysed, cannot be accused of failing to make a place for them. It neither shirks the compulsory question nor answers it by denial. It does give an account (1983a: 352).

In fact, the validity of Lewis' response to Armstrong is strengthened when we view it through the lens of epistemic possibility: saying that *x* is primitive is in no way ruled out by knowledge that *x*. Finally, not only "what we know in other areas" constrain metaphysical inquiry though, but also various principles or explanatory virtues such as coherence, simplicity, intelligibility, explanatory power, parsimony. It is an open question whether we should build these principles into the notion of epistemic possibility—so that arguments that seek to show that a particular theory is e.g. parsimonious, are understood as seeking to establish that the theory is epistemically possible in the relevant sense—or whether they should be seen as constraining the subset of epistemic possibility that we are interested in. Either approach could, in principle, be workable. It is not exactly clear what notion of epistemic possibility is relevant to epistemic HPE in science—there might well be several, suitable for different cases. I am not aware of any systematic inquiry into this issue, but it would not be surprising if this notion too was somehow constrained not only by compatibility with established scientific findings but also with various explanatory virtues generally taken to further scientific understanding.

[14] Partly because of considerations discussed in the next paragraph, this need not amount to what Daly and Liggins (2011) call deferentialism, i.e. "the view that philosophy should uncritically 'rubber stamp' every scientific claim" (334). Philosophers need not (should not!) uncritically accept everything scientists claim to currently know, not least because (as Daly and Liggins point out) different scientific disciplines may claim inconsistent things—it is thus a delicate question what counts as a well-established scientific finding. It is also an interesting question whether the stance taken here requires scientific realism, but it is not my intention that it should. The hope is that the 'scientific knowledge' can amount merely to knowledge of e.g. what the evidence is and suggests. More generally, these issues turn on how the second moving part of the epistemic possibility concept is fixed.

5. Understanding with Epistemic Possibilities

Now, how does viewing metaphysics as the construction and charting of epistemic how-possibly explanations of metaphysically interesting phenomena help with the Value Question? Well, thinking about the value of epistemic HPEs in science might guide us in finding out what the value of the (arguably) similar activity of metaphysics is.

Granted, this may at first seem like a dead end, because it is clear that the perhaps most *obvious* epistemic value that attaches to epistemic HPEs in science is *instrumental* to the acquisition of knowledge of what the actual explanation is. That is, knowing what the epistemically possible explanations are can, in various ways, help guiding research that will lead to knowledge of what the actual explanation is. In the tiger bush example, the idea is that with the accumulation of more empirical evidence, some of the earlier epistemic HPEs will no longer be such, i.e. scientists will rule out this or that mechanism as *not* in fact responsible for producing the phenomenon. The search for the explanation of tiger bush has, in fact, managed to cull some explanations previously considered to be possible, which no longer count as how-possibly explanation (Bokulich 2014: 331-333). Another illustration of this is Massimi's (2019) analysis of HPE modelling in particle physics. In order to fill a gap in the Standard Model, particle physicists have theorised entities referred to as super-symmetrical (SUSY) particles, but they have not been able to empirically confirm that any SUSYs actually exist. To put things very simply, scientists advance research in this area by modelling different ways in which the SUSY particle could be if it existed, given what they know. The array of possibilities is then used to guide empirical testing, where particle accelerators are used in attempts to *rule out* some of these possibilities as non-actual.[15]

But this clearly cannot be what is going with metaphysics, on the equilibrist picture. As was highlighted in the outline of Unresolved Dispute—which is part of what motivates equilibrism—the tools of metaphysical inquiry cannot adjudicate in a truth-conducive way between alternatives, and so there is no next step of metaphysics in which we might use the menu of possible alternatives in deciding on the true answer. Nor is it plausible that some other discipline will be able to take this map of possibilities prepared by metaphysics and go on to (empirically or otherwise) cull some of them on the road to the one true account.

However, it is arguably not always the case that the prospects of using the menu of possible explanations in the search for the *actual* explanation, are particularly good. This may be the case for a number of historical facts, for instance. Is epistemic HPE not epistemically valuable in cases where there is no prospect of being able to use it to find the actual explanation? I suggest that would be an implausible verdict. Theorising about the possible causes of e.g. the fall of Rome or the possible skin colour of dinosaurs is epistemically valuable, even if we have little reason to think that disputes over there matters will be fi-

[15] Not all instance of how-possibly explanation fits this pattern: in some cases the concern is to acquire possibility knowledge. But as I have argued elsewhere (Sjölin Wirling & Grüne-Yanoff, forthcoming), what is characteristic of such practices is that it targets *objective* (and often (known to be) counterfactual) rather than epistemic possibility. Practices of *epistemic* how-possibly explanation tend to behave like the practices described by Bokulich and Massimi.

nally resolved, and even if we were to recognise that we will never know the actual answer.[16] So there must be some value to these practices that is independent of their (perhaps more immediately evident) instrumental epistemic value, which they can have *whether or not* they lead to knowledge of actual explanations.

I will suggest that this value lies in the *understanding* afforded by ranges of epistemic how-possibly explanations. In particular, the idea is that our epistemic position with respect to a phenomenon is aided by access to and grasp of multiple, perhaps partly overlapping but also in central respects partly conflicting, perspectives on that phenomenon. I think that the state underlying this improved epistemic position is a form of understanding. But before I go on to outline more precisely what I take the relevant form of understanding to consist in, I want to consider yet another activity that bears resemblance to epistemic how-possibly explanation in science, and equilibrist metaphysics, which according to Catherine Elgin is epistemically valuable exactly because it increases understanding: academic art interpretation.

Elgin (2017) describes in detail a case where scholars present and defend different interpretations of Cézanne's *Le Compotier*. For instance, is the key to this painting the way Cézanne constitutes mass out of colour, or the way he emphasizes the flatness of the picture plane? (2017: 174-178). This dispute, between highly skilled scholars consists in sophisticated reason-giving: they present arguments in favour of their respective interpretation, compare it favourably to competing accounts, and so on. The debate is constrained by epistemic norms and standards that all parts are under the obligation to heed, it is not the case that *any* interpretation is viable, and scholars are required to lay out their case for their preferred interpretation in a particular way, putative reasons must be accepted as such by all parties (even if the reasons fail to convince the opposition), and so on.

This is just how we would expect a debate over a factual matter to go. But in a paradigmatic factual dispute, the function of arguing, of giving reasons that are supposed to weigh with the other party and/or a neutral audience, is to settle which side of the dispute is giving the true description of the fact in question. Yet, the dispute between interpretations of a painting seems irresolvable in a deep sense. It is not only that we cannot expect it to be resolved, we do not even in principle see what it would take to solve it—the idea of a conclusive reason that would settle the debate makes little sense. This is contrast to factual disputes, even those that are *in practice* irresolvable. For instance, palaeontologists disagree over the skin colour of the dinosaurs, and there is perhaps little reason to think that evidence which will conclusively settle the issue is forthcoming. But palaeontologists nevertheless will have no problem agreeing on what type of evidence *would* in principle settle the issue. The dispute over how Cézanne's painting is to be interpreted is different in that regard. So what is the point of sophisticated reason-giving in the case of art interpretation?

According to Elgin, some inquiries don't have the function of helping us find out and settle on the truth. Some disputes, including those like the Cézanne case, instead have epistemic value in virtue of its increasing our understanding.

[16] Elgin makes a similar point about disputes like that over whether Neanderthals buried their dead (2017: 181).

Here, arguments are like invitations to consider the perspective in question, to see the object through a particular lens. And within the course of inquiry in these fields, as part of this continuous competition between multiple interpretations where publicly available and assessable reasons in support of the various interpretations are put forward, scrutinised, refined, and so on, interconnections between the arguments, reasons and assumptions that make up the interpretation become visible. Their force (or sometimes, relative lack thereof) and relation to the object of inquiry come to be better appreciated by all parts to the dispute.

The idea, I take it, is that to see these various interconnections and relations and perspectives is to understand, to some significant degree, the object that all these perspectives are perspectives on: in this case the artwork, a particular painting. Because note that while different interpretations will highlight different features and display their significance whereas other features will be downplayed, all interpretations are constrained by the object of understanding, i.e., the painting. There is this one thing that all parties have access to, and some common knowledge about it, from which the inquiry proceeds, and which also constrains inquiry. For sure, what we know about the object may allow for very different and to some extent contradictory interpretations, as there might be both many things we don't know and many ways to accommodate that which we do know. But a viable interpretation cannot "float free" of the available facts. We can, although Elgin does not put it this way, extrapolate from this idea to the claim that viable interpretations need to be *epistemically possible* interpretations of the painting.[17]

What I am trying to convey here is that art interpretation is in an important sense similar to epistemic HPE in science, and to metaphysics: in all three cases there is something—some phenomenon that interests the art scholars, or philosophers, or scientists—that is the target of inquiry. This something is an important part of what constrains inquiry: not anything goes. In some cases we know more about the phenomenon, in some cases less, and that determines how many and how different the viable interpretations or accounts or explanations will be. But whatever the size and nature of the set of epistemically privileged propositions is, it constrains what counts as a viable account: it must be epistemically possible in the relevant sense. The art interpretation case shows us what the epistemic value of epistemic how-possibly explanation is—in absence of, or in addition to, its instrumental epistemic value: understanding. And it is not clear why it should matter whether the absence of instrumental epistemic value is *principled* (as it perhaps is in art) or just *in practice*: in science, or philosophy, where the disputes may not be irresolvable in the same deep sense, there can still be epistemic value in the form of increased understanding—in addition to or in place of any instrumental epistemic value. What is distinctive about metaphysics, and perhaps about a lot of other philosophy too, is that more or less *all* cen-

[17] Elgin also stresses that interpretations can be untenable despite not conflicting with what we know about the painting, for instance by being "uninformative or unenlightening", as would be the interpretation according to which *Le Compotier* is simply a picture of a bowl of fruit. Of course, this is true for how-possibly explanation and for metaphysical theories too: there are certain *explanatory virtues* that an account needs to fulfil. Again, this may or may not be built into the notion of epistemic possibility.

tral questions are like this, and so non-factive objective understanding is the *main* epistemic good that metaphysical research brings about.

I anticipate that someone might now object that in the art interpretation case—and perhaps also in most epistemic HPEs in science—there is clearly an *existing object*, to which the interpretations pertain. But several metaphysical disputes concern exactly whether there *does* exists something (e.g. properties, persisting objects), and it would be bad news if views like nominalism or stage theory could in principle not contribute to our understanding in the relevant sense. I think this worry can be mitigated if we take the notion of "objects of inquiry" less literally, i.e. as not necessarily pointing to an *object* which metaphysics inquires into. I take the "objects" of metaphysical inquiry to be more like *phenomena*—observable events—of scientific inquiry, which does not assume anything about the causes of the event. Property nominalism and universal realism then, can both (in principle) contribute to the understanding of *the phenomenon* of e.g. property instantiation. I note, however that the question of just how we are to pick out the *explananda* for metaphysical theories is a non-trivial question that a full-blown version of this kind of equilibrism would need to address.

5.1 Non-Factive Objectual Understanding

I am going to end this paper by outlining in some more detail the kind of understanding that I have suggested metaphysics (and art interpretation, and epistemic how-possibly explanation in science) is well placed to bring about, given the equilibrist picture of metaphysics. I call this state non-factive objectual understanding.

Philosophers do not agree on what understanding is, and I cannot go through all the different accounts here (see Hannon 2021 for a recent overview). What I will present here does not fully line up with any account currently on the market, but it has affinities with several of them. In any case, most people seem to agree that there are several different kinds of understanding (e.g. understanding *that,* understanding *why*, understanding a subject matter, understanding a language…), and I am happy to embrace such pluralism. All I need for the purposes of this paper is that there is an epistemic state with roughly the characteristics I outline below, and that this state is of epistemic value. Hopefully some accurate uses of 'understanding' capture something like that state, but at the end of the day, terminology is not particularly important to me.

First, the type of understanding which equilibrist metaphysics promotes is *objectual*. The term 'objectual understanding' is due to Kvanvig (2003), and denotes understanding of a subject matter, such as when one understands the periodic table, the Comanche dominance of the southern plains of North America between the 17th and 19th century, current foreign policy in Russia, Freud's theory of the unconscious, moral responsibility, or the nature of causation. Clearly, the objects of understanding can be anything from theories (including false ones) to actual phenomena. An epistemic subject S understands subject matter X insofar as S *grasps* a reasonably comprehensive amount of information about X. In particular, most authors will agree that understanding X requires grasping *connections* or *relations* between pieces of information pertaining to X. These can be e.g. explanatory, probabilistic, or logical relations, of coherence, of relative importance.

My claim is that the objects of understanding are not philosophical *theories* (or, in the case of science, some explanatory story; or in art, some interpretation). Instead, the objects of understanding are the phenomena we normally think of metaphysics as investigating, e.g. properties rather than *theories* of properties (in science, they are the phenomena in the world, in art, the artwork). But the theories are *vehicles of understanding* (Greco 2014: 293) in the sense that they are what we grasp, and it is via grasping them that we get understanding of the phenomena that are the objects of understanding. I differ here from McSweeney (2023), who has recently also proposed that metaphysics is primarily aimed at understanding, which is epistemically valuable. But the objects of understanding she has in mind appears to be *metaphysical theories*. That is, good metaphysical inquiry helps metaphysicians understand their own, and competing theories, better. I have no quarrel at all with that claim, but I do not think this is the main source of epistemic value for the equilibrist aim. Objectual understanding *of theories* may be instrumentally epistemically valuable, e.g. in the tiger bush case, good understanding of the various how-possibly *models* might be very important in guiding search for the actual explanation. And similarly, understanding of metaphysical theories may be an important precondition for the objectual understanding I have in mind, but this latter understanding—from which I think the instrumental value of the equilibrist aim flows—does not have theories as its objects.

Second, as most other forms of understanding I take the kind relevant here to psychologically involve *grasping* a set of propositions. Grasping is often said to involve a "*seeing"* of how things fit together, that bring with it a form of *cognitive command or control*, characterised by giving the subject who has the grasp a number of abilities to "do" things with, or "manipulate" the information in question in various ways, such as reason with and within the body of information, apply explanations to novel cases, draw novel inferences, and so on. Importantly however, in my view grasping does *not* involve belief: you can grasp a set of propositions in the relevant sense without believing them to be true. Here the kind of objectual understanding I have in mind differs from Kvanvig's. Why? Because the understanding I have in mind typically involves grasping the information provided by more than one theory of the same subject matter, and these sets of information will typically contradict each other. Successful metaphysics makes available these sets of jointly epistemically possible propositions, sets that are mutually incompatible. The idea is that grasping the information provided by this plurality of theories about, say, causation, is a way of increasing one's understanding of causation. So it won't work to require that a subject who grasps a plurality of mutually inconsistent theories about causation needs to believe the information that she grasps. For others who deny that understanding requires belief, see e.g. Dellsén 2017 and Elgin 2017.

For similar reasons, the objectual understanding that equilibrist philosophy—or more generally, epistemic how-possibly explanation—produces the resources for is *non-factive*. That is, it is not required—in contrast to how e.g. Kvanvig describes objectual understanding—that the "central propositions" in the grasped amount of information be true. That is, the central propositions in the set, about e.g. causation, need not be true for there to be understanding of causation. Giving up on factivity will seem radical to some, and some will suspect that it is *ad hoc*. But other theorists of understanding reject factivity too, including Elgin (2007), Riggs (2009), and Potochnik (2020), for reasons that do

not have anything to do with the epistemic quality of metaphysics. So there is room, and even demand, for understanding-like, epistemically valuable cognitive states that are non-factive.[18] Note also that factivity isn't given up just to save the idea that *metaphysics* is supposed to deliver understanding. I am trying to capture a state which can explain the epistemic value of not just metaphysics on the equilibrist picture, but also other activities that seem epistemically valuable, even in the absence of instrumental value to knowledge of what the actual answer is, including scientific HPE and art interpretation.

In sum then, S has non-factive objectual understanding of, say, causation, when S grasps—can represent and has cognitive control over—a set of epistemically possible comprehensive subsets of propositions about causation (i.e. philosophical accounts or theories), and the relations (both inside a subset and between subsets) between these propositions. S is not required to believe any of these propositions, and it is not required that the central propositions in these grasped subsets are true. The main idea is that our epistemic position with respect to a phenomenon is aided by this kind of access to and grasp of multiple, perhaps partly overlapping but also in central respects partly conflicting, perspectives on that phenomenon. I think it's aptly called understanding, but I'm not much into fighting for the term. *Something* like this idea is present also in the work of others, including some of those cited above, such as Elgin and Potochnik. It is not easy to pin down exactly how and in what sense this grasp of multiple competing perspectives on the same phenomenon, 'informs us' about or improves our epistemic position with respect to the phenomenon in question, since we do not in any straight-forward way learn new truths about what it is like. But maybe one way to put it is that it improves our grasp of that which we *do* know; what that does and does not imply—e.g. by highlighting the different aspects of the phenomenon as we know it; by illuminating and emphasizing how these features sometimes pull in different directions; by exploring all the things that are compatible with what we know and thereby helping us see what we do *not* know, or what we cannot rule out. Whether or not this is aptly called 'understanding', I think a good case can be made that it is epistemically valuable.

All of the above is obviously compatible with it being *more* epistemically valuable to have e.g. knowledge or factive understanding in domains where there are matters of fact to discover. Having knowledge or factive understanding of e.g. causation in the form of grasping the one correct comprehensive theory of causation, and being able to rule the competing ones out, may well be more valuable and desirable from an epistemic point of view than the non-factive understanding which metaphysics can give us. But the running assumption here is that we don't and won't have that—what we are concerned with is accounting for the epistemic value of that which we (when things go well) *can* achieve given equilibrism. It is my view that contributing to resources for non-factive understanding accounts for the lion's share of the epistemic value of metaphysics'

[18] My use of "non-factive" is meant to mark distance from what is normally meant by "factive" understanding. That said, the type of understanding of interest here could in principle be called factive in the sense that it needs to be true that the (sets of) propositions grasped are epistemically possible in the relevant sense.

equilibrist aim and the processes and practices that, when things go well, help approximate it.

5. Conclusion

The primary aim of this paper was to further develop and supplement the "equilibrist" view that the epistemic aim of metaphysics is to find out what viable alternative theories there are with respect to metaphysical questions. In particular, I argued that equilibrism faces what I called the Value Question: the challenge to explain the alleged epistemic value of this equilibrist aim and inquiry that promote it. First, I proposed to understand this aim in terms of *epistemic possibility*, drawing on an analogy with epistemic how-possibly explanation in science. Second, I argued that metaphysics thus conceived—and activities with a similar structure that constructs and charts multiple, perhaps partly overlapping but also in central respects partly conflicting, perspectives on, or explanations of, one and the same phenomenon—provides resources for a non-factive understanding of the objects of (in this case, metaphysical) inquiry. Such non-factive objectual understanding is arguably of epistemic value, not just in metaphysics, but also in other fields where there are irresolvable (whether in principle or practice) disputes, and this includes art interpretation as well as science.[19]

References

Armstrong, D.M. 1980, "Against 'Ostrich' Nominalism: A Reply to Michael Devitt", *Pacific Philosophical Quarterly*, 61, 440-449.

Armstrong, D.M. 2010, *Sketch for a Systematic Metaphysics*, Oxford: Clarendon Press.

Barnett, Z. 2019, "Philosophy Without Belief", *Mind: A Quarterly Review of Philosophy*, 128, 109-138.

Beebee, H. 2018, "Philosophical Scepticism and the Aims of Philosophy" *Proceedings of the Aristotelian Society*, 118, 1, 1-24.

Bokulich, A. 2014, "How the Tiger Bush Got Its Stripes: 'How Possibly' vs. 'How Actually' Model Explanations", *The Monist*, 97, 3, 321-338.

Brandon, R.N. 1990, *Adaptation and Environment*, Princeton: Princeton University Press.

Brennan, J. 2010, "Scepticism About Philosophy" *Ratio*, 23, 1-16.

Cappelen, H. 2017, "Disagreement in Philosophy: An Optimistic Perspective" in D'Oro, G. and Overgaard, S. (eds.), *The Cambridge Companion to Philosophical Methodology*, Cambridge University Press.

Chalmers, D.J. 2015, "Why Isn't There More Progress in Philosophy?" *Philosophy*, 90, 1, 3-31.

Daly, C. and Liggins, D. 2011, "Deferentialism", *Philosophical Studies*, 156, 3, 321-337.

Dellsén, F. 2017, "Understanding without Justification or Belief", *Ratio*, 30, 3, 239-254.

[19] Thanks to Helen Beebee, David Liggins, Frans Svensson, participants in the research seminar in theoretical philosophy at the University of Gothenburg, and an anonymous reviewer of this journal for helpful comments on this paper. My research was generously funded by the Swedish Research Council grant no. 2019-00635.

Dellsén, F., Lawler, I. and Norton, J. (forthcoming), "Thinking About Progress: From Science To Philosphy", *Noûs*.

Elgin, C.Z. 2007, "Understanding and the Facts", *Philosophical Studies*, 132, 1, 33-42.

Elgin, C.Z. 2017, *True Enough*, Cambridge, MA: MIT Press.

Frances, B. 2017, "Extensive Philosophical Agreement and Progress", *Metaphilosophy*, 48, 1-2, 47-57.

Fumerton, R. 2010, "You Can't Trust a Philosopher", in Feldman, R. and Warfield, T.A. (eds.), *Disagreement*, Oxford: Oxford University Press, 91-110.

Godfrey-Smith, P. 2006, "Theories and Models in Metaphysics", *The Harvard Review of Philosophy*, 14, 1, 4-19.

Goldberg, S.C. 2009, "Reliabilism in Philosophy", *Philosophical Studies*, 142, 1, 105-117.

Goldberg, S.C. 2013, "Defending Philosophy in the Face of Systematic Disagreement", in Machuca, D.E. (ed.), *Disagreement and Skepticism*, New York: Routledge, 277-294.

Greco, J. 2014, "Episteme: Knowlede and Understanding", in Timpe, K. and Boyd, K. (eds.), *Virtues and the Vices*, Oxford: Oxford University Press, 285-301.

Hannon, M. 2021, "Recent Work in the Epistemology of Understanding", *American Philosophical Quarterly*, 58, 3, 269-290.

Keller, J.A. 2017, "Philosophical Individualism", in Keller, J.A. (ed.), *Being, Freedom, and Method: Themes From the Philosophy of Peter van Inwagen*, Oxford: Oxford University Press.

Kelly, T. 2016, "Disagreement in Philosophy: Its Epistemic Significance", in Cappelen, H., Gendler, T.S. and Hawthorne, J. (eds.), *The Oxford Handbook of Philosophical Methodology*, Oxford: Oxford University Press, 374-394.

King, N.L. 2012, "Disagreement: What's the Problem? or A Good Peer is Hard to Find", *Philosophy and Phenomenological Research*, 85, 2, 249-272.

Kornblith, H. 2013, "Is Philosophical Knowledge Possible?", in Machuca, D.E. (ed.), *Disagreement and Skepticism*, New York: Routledge, 260-276.

Kriegel, U. 2013, "The Epistemological Challenge of Revisionary Metaphysics", *Philosophers' Imprint*, 13.

Kvanvig, J.L. 2003, *The Value of Knowledge and the Pursuit of Understanding* (Vol. 113), Cambridge: Cambridge University Press.

Lewis, D. 1983a, "New Work for a Theory of Universals", *Australasian Journal of Philosophy*, 61(December), 343-377.

Lewis, D. 1983b, *Philosophical Papers Vol. I*, New York: Oxford University Press.

Lowe, E.J. 1998, *The Possibility of Metaphysics*, Oxford: Oxford University Press.

Massimi, M. 2019, "Two Kinds of Exploratory Models", *Philosophy of Science*, 86, 5, 869-881.

Maurin, A.S. 2002, *If Tropes*, Dordrecht: Kluwer.

McSweeney, M.M. 2023, "Metaphysics as Essentially Imaginative and Aiming at Understanding", *American Philosophical Quarterly*, 60, 1, 83-97.

Nado, J. 2019, "Who Wants to Know?", in Gendler, T.S. and Hawthorne, J. (eds.), *Oxford Studies in Epistemology*, Vol. 6, 114-136.

Paul, L.A. 2012, "Metaphysics as Modeling: the Handmaiden's Tale", *Philosophical Studies*, 160, 1, 1-29.

Potochnik, A. 2020, "Idealization and Many Aims", *Philosophy of Science*, 87, 5, 933-943.

Riggs, W. 2009, "Understanding, Knowledge, and the Meno Requirement", in Haddock, A., Millar, A. and Pritchard, D. (eds.), *Epistemic Value*, New York: Oxford University Press, 331-338.

Rinard, S. 2013, "Why Philosophy Can Overturn Common Sense", in Gendler, T.S. and Hawthorne, J. (eds.), *Oxford Studies in Epistemology*, Vol. 4, Oxford: Oxford University Press, 185-213.

Rosen, G. 2020, "Metaphysics as a Fiction", in Armour-Garb, B. and Kroon, F. (eds.), *Fictionalism in Philosophy*, Oxford: Oxford University Press, 28-47.

Rotondo, A. 2015, "Disagreement and Intellectual Scepticism", *Australasian Journal of Philosophy*, 93, 2, 251-271.

Salmon, W.C. 1989, "Four Decades of Scientific Explanation", in Kitcher, P. and Salmon, W.C. (eds.), *Scientific Explanation*, Minneapolis: University of Minnesota Press.

Simons, P. 1998, "Metaphysical Systematics: A Lesson from Whitehead", *Erkenntnis*, 48, 2-3, 377-393.

Sjölin Wirling, Y. 2021, "Non-Uniformism and the Epistemology of Philosophically Interesting Modal Claims", *Grazer Philosophische Studien*, 98, 629-656.

Sjölin Wirling, Y. and Grüne-Yanoff, T. (forthcoming), "Epistemic and Objective Possibility in Science", *British Journal for the Philosophy of Science*.

Stoljar, D. 2017, *Philosophical Progress*, Oxford: Oxford University Press.

Verreault-Julien, P. 2019, "How Could Models Possibly Provide How-Possibly Explanations?", *Studies in History and Philosophy of Science Part A*, 73, 22-33.

Whitehead, A.N. 1929, *Process and Reality: An Essay in Cosmology*, New York: Macmillan.

The Thesis of Revelation in the Philosophy of Mind: A Guide for the Perplexed

Bruno Cortesi

University School for Advanced Studies IUSS Pavia

Abstract

The thesis of experiential revelation—*Rev* for brevity—in the philosophy of mind claims that to have an experience—i.e., to be acquainted with it—is to know its nature. It is widely agreed that although at least moderate versions of *Rev* might strike one as plausible and perhaps even appealing, at least up to a certain extent, most of them are nonetheless inconsistent with almost any coherent form of physicalism about the mind. Thus far, the issue of the alleged tension between *Rev* and physicalism has mostly been put in the relevant literature in terms of phenomenal concepts—those concepts which refer to phenomenal properties, or qualia, and characterize them in terms of the peculiar quality(ies) they exhibit—and some kind of "special feature" those concepts allegedly possess. I call this version of *Rev C-Rev*. This paper aims to suggest that while it is true that phenomenal concepts reveal the nature of their referent(s)—i.e., it is *a priori*, for a subject possessing the concept and just in virtue of possessing it, what it is for the referent(s) of the concept to be part of reality—this feature of them, in turn, rests on a non-conceptual non-propositional kind of knowledge, namely, *sui generis* introspective knowledge by acquaintance of one's own phenomenally conscious states. I call this version of *Rev A-Rev*. §1 provides some introductory material. In §2 I discuss two arguments that have recently been put forth to undermine the cogency of *C-Rev* against physicalism. §3 elaborates on the historical roots of *C-Rev*. §4 presents some of the major arguments which have been offered for *A-Rev*. A few concluding remarks close the paper.

Keywords: Revelation, Physicalism, Knowledge by acquaintance, Propositional knowledge, Phenomenal consciousness.

1. Introduction

David Chalmers has written:

Argumenta 10, 1 (2024): 107—126
ISSN 2465-2334

DOI 10.14275/2465-2334/20230.cor
First published: 27 March 2023

> We know consciousness far more intimately than we know the rest of the world, but we understand the rest of the world far better than we understand consciousness. Consciousness can be startlingly intense. It is the most vivid of phenomena; nothing is more real to us. But it can be frustratingly diaphanous (Chalmers 1996: 3).

The verb 'to know' appears twice in the passage above. Yet I think one might ask: was it meant to convey the same meaning in both of its instances? Or did Chalmers, instead, intend to use it to refer to two distinct kinds of state?

This essay aims to suggest that our knowledge of the phenomenology of our own phenomenally conscious states—*i.e.*, those states there is something it is like for a subject to be in (Nagel 1974)—is of a fundamentally different kind with respect to our knowledge of what Chalmers refers to as "the rest of the world".

I also take it that it is *because* we know consciousness *so* intimately that it resists a reductive naturalistic explanation as the one that has been—and/or *is being*—offered for an astonishingly vast variety of *explananda* at least since the development of modern science: the mysteriousness of consciousness with respect to a naturalistic viewpoint broadly construed—*i.e.*, what Chalmers (e.g., 1995) labels "the hard problem of consciousness"—is *rooted* in its being more vivid than any other phenomenon to anyone who has ever been conscious.

It follows that even thinking about addressing the hard problem of consciousness without *eo ipso* also addressing the issue of our epistemic relation with phenomenal properties, will be inevitably doomed to fail as an endeavour: the hard problem, as a metaphysical issue, *forces us* to reconsider the way in which we know the phenomenology of our experiences. The reverse is also true: epistemological considerations, in the case of phenomenal consciousness, might have a huge import on the metaphysical investigation of the mind and of reality in general. It is no coincidence that the major arguments that have been offered against materialism about phenomenal consciousness[1] in the last decades—e.g., Chalmers' (e.g., 1996;

[1] There is not, still, unanimous consensus on how physicalism about phenomenal consciousness should be formulated. According to type-identity materialism (Place 1956; Feigl 1958; Smart 1959; Armstrong 1968), types of phenomenal experience—say painful experiences—are identical to specific types of neural activations tacking place in the brain—say c-fibers firing. Notoriously, this version of physicalism suffers from an objection raised by Putnam (1967) and Fodor (1975), among others. The main idea behind such an objection is that (conscious) mental states are *multiply realizable*: other species besides the human one do have (conscious) mental states very similar if not identical to our own (e.g., they do feel pain) despite having significantly different nervous systems, which is clearly incompatible with types of (conscious) mental states being *identical* to *specific types* of neural activations. Despite having often been considered as a fatal objection to type-identity materialism, this is not the only objection which may be raised against it (see, e.g., Kripke 1980). Even leaving type-identity theories aside, however, there are several options a materialist might resort to when trying to specify the kind of metaphysical relation she believes to hold between physical facts, states, processes and/or properties and phenomenal/conscious ones, including—but not limited to—*Realization* (Melnyk 2003, 2006, 2018; Shoemaker 2007, 2014) *Constitution* (Pereboom 2011) and *Grounding* (Dasgupta 2014, Kroedel & Schulz 2016, O'Conaill 2018, Goff 2017). To complicate the matter, even providing a precise characterization of the *relata* in the very first place is far from being an easy task. My own preferred version of physicalism is the

2009) *conceivability-argument* and Jackson's (1982) *knowledge-argument* above all—all revolve around an attempt to draw metaphysical conclusions from epistemological premises. Likewise, as Stalnaker (2008: 26) has noted cogently, most of the major attempts to counter those arguments in one way or another attempt to decouple items of knowledge—facts, for instance—from metaphysical distinctions between possible situations in which those items obtain.

The thesis of experiential revelation—*Rev* for brevity—has come in the philosophical literature on the mental in a variety of slightly differently nuanced formulations. The term 'Revelation' was introduced by Johnston (1992) to refer to Strawson's (1989) claim whereby the nature of colors is fully revealed in color experiences, but already in Russell (1910; 1912: 47) one can find what is arguably a version of the thesis. *Rev* is generally understood as a thesis about the *essence* of phenomenal properties, where phenomenal properties, or qualia, in turn, are typically defined as properties of conscious mental states which type those states by what it is like for a subject to have them (Nagel 1974). *Rev* has sometimes been phrased (e.g., by Trogdon 2016; Nida-Rümelin 2007; Goff, 2011; 2015; 2017) in terms of phenomenal concepts—those concepts which refer to phenomenal properties and characterize[2] them in terms of the peculiar quality(ies) they exhibit (I will elaborate on this version of *Rev* in a moment). Others (e.g. Majeed 2017; Chalmers 2016) have phrased *Rev* in terms of introspection. Liu (2019; 2020; 2021; forth: 3) offers a rather general characterization of *Rev*: Given an experiencing subject S and a phenomenal property Q, "By having an experience-token with phenomenal property Q, S is in a position to know

one Coleman (2008: 93) calls *conventional physicalism*, namely, a view which states that phenomenal properties supervene upon the *non-experiential* physical. Conventional physicalism consists in the combination of two claims: (a) phenomenal properties supervene upon (fundamental) physical ones, that is, every metaphysically possible world that is a minimal physical duplicate of the actual world must also be a duplicate of the actual world with respect to every conscious property, *i.e.*, a *C-duplicate* of the actual world. This is Jackson's (e.g., 1998) version of physicalism. The addition of the word "minimal" is meant to avoid the so-called problem of epiphenomenal ectoplasms, namely, pure phenomenal entities of some kind which do not interact causally with anything else there is in a given possible world. A minimal physical duplicate of the actual world is a world which duplicates all the physical properties of the actual world *without adding anything else*. According to Lewis (1983), (a) suffices for what he calls minimal physicalism; (b) There are no fundamental phenomenal properties, that is, the view known as Russellian monism is false. It is widely acknowledged in the relevant literature that (metaphysical versions of) physicalism must imply at the very least the supervenience of phenomenal properties upon physical ones. As far as I know, only Montero (2013) and Montero & Brown (2018) deny this. The view they put forth, however, is definitely minoritarian among physicalists. As we shall see, Damnjanovic (2012) defends a version of the identity thesis. In what follows, unless otherwise specified I will use the words 'physicalism' and 'materialism' interchangeably to refer to conventional physicalism. This note owes a lot to my colleague and dear friend Giacomo Zanotti: see Zanotti 2020, 2021, 2022.

[2] The notion of a characterization is just aimed at capturing the idea that concepts always do characterize their referent(s) as being in a certain way or present it/them under a peculiar aspect. As Trogdon (2016) notes, this construal of what a characterization is requires a Fregean/two-factors account of reference and meaning according to which the referent and the cognitive significance of a concept are distinct.

that 'Q is X', where the predicate 'X' captures the essence of Q".[3] In other words: to have an experience—*i.e.*, to be acquainted with it—and possibly to attend to it, *is* to know its essence: just by having, say, a headache (and attending to it), one is put in a position to come to know what pain—or better, the painfulness of her experience—*essentially* is.

For the purposes of the present essay, I do think that a rather broad understanding of what the essence of something is will suffice. Since Kripke published his (1980) and Fine his (1994), talk of essence has regained a central poignancy in many debates in metaphysics, and is now deemed as perfectly legitimate (see Tomasetta 2016). Along with Fine (1994; 1995a; 1995b), Hale (2013) Lowe (2012) and Tomasetta (2016)—among several others—I do think the notion of 'essence' is primitive and not further analyzable. I will adopt a non-modal/definitional/Finean (Fine 1995a, 1995b; Dasgupta 2014; Liu, forth.) approach whereby the essence of a certain item *x* is what makes x the thing it is/belongs to *x*'s most core respects. *X*, thus, will be said to have a certain property *p* essentially if *p* belongs to the class of *x*'s most core respects, that is to say, to the class of those properties which make *x* the thing it is.[4]

It is widely agreed that although some moderate versions of *Rev* might strike one as *prima facie* plausible and perhaps even appealing, at least up to a certain extent, most of them are nonetheless inconsistent with almost any coherent form of physicalism about the mind. David Lewis' (1995) *Should a materialist believe in qualia?* is arguably one of the *loci* where the tension between *Rev*—which Lewis refers to as *the identification thesis*—and materialism emerges most clearly. There (1995: 141-42) Lewis writes:

> Unfortunately there is more to the folk-psychological concept of qualia than I have yet said. It concerns the modus operandi of qualia. Folk psychology says, I think, that we identify the qualia of our experiences. We know exactly what they are—and that in an uncommonly demanding and literal sense of 'knowing what' […] If qualia are physical properties of experiences, and experiences in turn are physical events, then it is certain that we seldom, if ever, [know the nature of] the qualia of our experiences. Making discoveries in neurophysiology is not so easy!

[3] Another broad characterization of the main idea behind *Rev* is offered by Stoljar (2009: 115).

[4] As we shall see, one of the main ideas behind *Rev* is that phenomenal properties belong essentially to the states bearing them. I do believe that this might be shown to be the case under a modal account of the distinction between essential and accidental properties—as the one Balcan Marcus (1967), Kripke (1980), Zalta (2006), Correia (2007) and Brogaard and Salerno (2007a; 2007b; 2013) (among others) defend (see also Robertson Ishii and Philip 2020)—as well. Grossly, under a modal account of essentiality, a property p belongs essentially to an item x iff it is necessary that *x* has *p*, and it is necessary that *x has p* iff *x* has *p* in all possible worlds—or at least in all the possible worlds where *x* exists. Suppose now there's someone, say Thomas son of Mary (Damnjanovic 2012), who's feeling a sharp pain. Imagine now a possible world W^1 where instead of being acquainted with a "painful" phenomenal quality, Thomas is acquainted with a "joyful" one. Would you really say that it is *pain* that Thomas is feeling in W^1? Suppose now that Thomas is having a visual experience of, say, a red circle in the actual world, and is instead acquainted with a "bluish squarish" phenomenal character in W^1. Would you really say, again, that it is *the same experience* Thomas is having in the two possible worlds? The remarks Kripke draws in his (1980: 150-52) seem to go in the same direction.

The main idea Lewis wants to convey here is rather straightforward. On the one hand, as Goff (e.g., 2017: 107-108) and Stoljar (2009: 115), among others, have emphasized, we seem to be in a rather peculiar—not to say unique—epistemic situation with respect to the phenomenology of our own conscious states. However, it is obviously not the case that one can learn anything about the complex neuro-physiology of her brain *just by being in pain* (and attending to her painful experience). In light of this *impasse*, one is apparently left with two options:

¬*Rev:* Our relation to the phenomenal properties of our own phenomenal mental states does not, indeed, have any or most of the special features it appears to have. Therefore, nothing truly essential is—nor could be—actually revealed to us by the mere instantiation (and attentive awareness) of those properties. ¬*Rev* is compatible with any form of physicalism/functionalism about phenomenal properties.

Rev: The nature of the phenomenal properties of our own mental states is revealed to us by the mere instantiation (and attentive awareness) of them. If so, then those properties are arguably not identical nor completely reducible to a number of physical/functional properties and/or processes or states.

Thus far, the issue of the alleged tension between *Rev* and physicalism has mostly been put in the relevant literature in terms of phenomenal concepts and some kind of "special feature" those concepts allegedly possess. From now on, I will refer to this version of *Rev* as *C-Rev.* According to *C-Rev*, phenomenal concepts provide a (full) essential characterization of their referent(s) (see Trogdon 2017). A concept *C* is said to provide a *partial* essential characterization of its referent(s) iff there are some properties *p, q, r* (at least one) such that *C*'s referent(s) has/have those properties essentially and *C* characterizes its referent(s) has having those properties. *C* is said to provide a *full* essential characterization of its referent(s) iff for *any* property *p*, if *C*'s referent(s) has/have *p* essentially, then *C* characterizes it/them in terms of *p*.

Versions of *C-Rev* have been defended by Nida-Rümelin (2007) and Goff (2011; 2015; 2017; 2019) among others. Nida-Rümelin (2007) argues that via phenomenal concepts one is allowed to *grasp* the properties they refer to, where to grasp a property is to understand what that property essentially consist in, and to do so without any background knowledge besides the one provided by those concepts themselves. Likewise, Goff (2011) argues that phenomenal concepts are *transparent* where a concept is said to be transparent (Goff 2011: 15) "just in case it reveals the nature of the entity it refers to, in the sense that it is a priori (for someone possessing the concept and in virtue of possessing the concept) what it is for that entity to be part of reality". More specifically, Goff (2011: 194) offers the following taxonomy: *transparent* concepts reveal the nature of their referent(s)—*i.e.*, provide a full essential characterization of their referents in the sense provided above; *translucent* concepts reveal part of, but not all, the nature of their referents—*i.e.*, provide a partial essential characterization of their referents; *mildly opaque* concepts do not reveal any essential property of their referent(s) but reveal some accidental features of them which uniquely identify it/them in the actual world; *radically opaque* concepts reveal neither essential nor accidental properties of their referent(s). Opaque concepts, that is, merely *denote* their referents, but say little or nothing about what it is for them to

be part of reality. The amount of what is revealed by a concept of its referent(s) coincides with what that concept allows to know *a priori* about it/them.

What I wish to suggest is that while it is true that phenomenal concepts allow a subject to grasp the properties they refer to *just by being had by her*, this feature of them, in turn, rests on a more primitive, pre-conceptual non-propositional kind of knowledge, which may be understood in analogy with what Pitt (2011) calls *acquaintance-as-knowledge* or a*cquaintance-knowledge,* not to be conflated with knowledge *by* acquaintance, the latter being, for Pitt (2011), propositional in kind. Pitt's notion of acquaintance-knowledge draws from Levine's (2011) distinction between *implicit* and *explicit self-knowledge of thought.* The latter is, for Levine (2011: 108), "what we have when we explicitly formulate a meta-cognitive thought, such as 'I believe that San Francisco is a beautiful city'"; *implicit self-knowledge of thought*, by contrast (2011: 108-109) "is not the result of any explicit formulation or reflection. Rather, it's the knowledge that seems to come with the very thinking of the thought itself. [...] To implicitly know what one is thinking is just to think with understanding".

On the view I endorse, to "acquaintance-know" what it is like to have an experience—which I consider to be an essential property of the experience itself—would be, to paraphrase Levine, just to experience (with focusing). I will call this version of *Rev A-Rev.*

Here is how the paper is structured. In §2 I discuss two arguments that have recently been put forth to undermine the cogency of *C-Rev* against physicalism, namely, those put forth in Damnjanovic 2012 and Trogdon 2016. §3 elaborates on the historical roots of *C-Rev*. §4 presents some of the major arguments which has been offered for *A-Rev*. Few concluding remarks close the paper.

2. Damjanovic and Trogdon on *C-Rev*

Following on Lewis' discussion of 'the identification thesis', Nic Damnjanovic writes:

> […] Lewis speacquaintance-knowledges of experiences 'identifying' qualia in a demanding way. But it is clear that to 'identify' qualia in this way—to know exactly what qualia are—*is to have propositional knowledge of their nature*, just as, as he explicitly says, knowing exactly what potassium is requires knowing its atomic number (Damnjanovic 2012: 72, emphasis mine).

It honestly does not strike me as obvious, as Damnjanovic seems to be here implying, that *any possible piece* of essence-revealing knowledge we might ever come to have—with the possible exception of knowledge how[5]—*must* be propositional in kind. That *any possible piece* of essence-revealing information about *any possible item* in the universe—or at least about those items whose essence we might ever come to know given our cognitive architecture—can only be conveyed by a (number of) proposition(s)—let alone a (number of) proposition(s) expressing

[5] Even though there are authors—e.g., Stanley and Williamson (2001); Stanley (2011); Brogaard (2011), Williamson (2000)—who believe that even knowledge-how might indeed consist in the knowledge of a number of propositions.

some fundamental physical facts—is not a truism.[6] Yet, surprisingly, it is merely taken for granted by Damnjanovic without being argued for at all.

Suppose now that someone, call him Thomas son of Mary, is tasting peaches for the very first time in his life. Damnjanovic's (2012: 73, emphasis mine) own proposed version of the argument from *Rev* against physicalism has the following form. Note that physicalism is here being treated as equivalent to a version of the identity thesis:

1. If Identity is true and Thomas is in a position to know the full nature of the taste of peaches, then Thomas is in a position to *know that p*.
2. Thomas is in a position to know the full nature of the taste of peaches [Revelation]
3. Thomas is not in a position to *know that p*.

Therefore

4. Identity is false.

I think this version of the argument from *Rev* misconstrues the actual meaning of the thesis in the very first place. Nonetheless, it is interesting to note—as Stoljar (2009: 124) has also done—that it—as well as similar versions of it that have been offered—would function against almost any identity statement, whether the alleged identity is between phenomenal properties and physical properties, or between phenomenal properties and "spiritual" properties, or between phenomenal properties and "aesthetic" ones, and so on.

Whilst I agree with both Stoljar and Damnjanovic that *Rev* as thus understood might imply an "uncompromising version of primitivism about experience according to which [qualia] are primitive items in the world, wholly distinct from everything else" (Stoljar 2009: 124), I disagree with them in that I do not regard this as a reason to dismiss the thesis; rather, I regard it as a rather natural conclusion stemming from it, a conclusion I am indeed willing to accept. As Tomasetta says,

> That physicalism is indeed more a worldview than a well-grounded philosophical thesis is further buttressed by the almost religious fervency with which materialist views are often held (Bonjour 2010: 4). A fervour that is evident, for example, in Dennett's (1989: 37) declaration that "dualism is to be avoided at all costs", a position which is certainly not well suited to a rational inquirer (Tomasetta 2015: 107).[7]

Just as (this version of) the argument from *Rev* would function against any kind of alleged identity between kinds of properties, the cogency of (one version of) the knowledge-argument largely depends on what we substitute for 'p' in the (allegedly propositional) new piece of knowledge Mary would acquire once confronted with a red item for the very first time. There are authors (e.g., Church-

[6] Note, also, that *Prima facie* this *de facto* precludes non-linguistic individuals like newborns and animals from the possibility of knowing anything.

[7] Pitt (2011: 2) says that skepticism about the existence of a distinctive, individuative and proprietary phenomenology of conscious thought is "more often based on prior theoretical commitment, or overreaching confidence in the explanatory resources of contemporary Naturalism […] than on unbiased reflection upon our conscious mental lives, or careful evaluation of the arguments in its favor". I believe his concerns may as well be raised with regard to skepticism against *Rev*.

land 1989; Bigelow and Pargetter 1990; Conee 1994, Balog 2012, just to mention some) who believe that there is no new proposition Mary would nor could learn. Rather, she would just *become acquainted* with a new phenomenal property. I quite agree with this; yet, again, I don't think this account, when properly developed, would undermine the cogency of the knowledge-argument—nor that of the argument from *Rev*—against physicalism. This is so because (a) I endorse a *constitutive* account—as opposed to a *causal* one (more on this taxonomy momentarily)—of the notion of knowledge by acquaintance whereby such kind of knowledge is *essentially constituted* by the relation of acquaintance rather than being merely caused or enabled by it; (b) I take this knowledge to be essence-revealing.

Once we interpret Revelation as claiming that by having an experience with a quale Q one is put in a position to gain complete *knowledge by acquaintance* of Q, according to Damnjanovic (2012: 76, emphasis mine) "The argument from Revelation fails, therefore, because i*t incorrectly supposes that Thomas' complete knowledge of the taste of peaches implies that he knows certain truths about the nature of peaches*". This does not seem to be right. *Rev* claims that by tasting peaches Thomas is put in a position to grasp the essence of the experience of tasting peaches—or what it is like to taste peaches; it does not claim, though—or at least it does not *have* to claim, that Thomas comes to know any new proposition about peaches.

Let us now have a look at the remarks Trogdon draws about *C-Rev*. Trogdon's (2016: 4-5) own proposed version of the argument from *C-Rev* against materialism goes like this:

1. PHENOMENAL RED provides an essential characterization of its referent, phenomenal red.
2. PHENOMENAL RED doesn't provide a physical/functional characterization of phenomenal red.
3. If PHENOMENAL RED provides an essential but not a physical/functional characterization of phenomenal red then this property isn't a physical/functional property.
4. Hence, phenomenal red isn't a physical/functional property.

Where for a concept to provide a physical/functional characterization of its referent(s) is for it to characterize that/those referent(s) as physical/functional in kind. Trogdon believes this version *C-Rev* against materialism fails to achieve its goal.[8] The fact, according to Trogdon, is that while the first premise is plausible if 'essential characterization' is read as *partial* essential characterization, the linking premise only makes sense if 'essential characterization' is read as *full* essential characterization. That is to say: the concept 'PHENOMENAL RED' might characterize the property 'phenomenal red' as having *some* of the properties it has essentially; *prima facie* there is no reason, though, to think that 'PHENOMENAL RED' characterizes 'phenomenal red' as having *all* the properties it has

[8] Let me emphasize, though, that according to Trogdon (2016: e.g. 1) (his reading of) *Rev* indeed poses an *indirect* challenge to physicalism. More specifically, it has the potential to undermine the so-called *phenomenal concepts strategy, i.e.*, one of the main strategies physicalists may invoke to respond to typical dualist objections against their view, including explanatory gap-style objections (Levine 1983) and the conceivability-argument.

essentially. More specifically, phenomenal red might have the property of being a physical/functional property essentially and still 'PHENOMENAL RED' might not characterize it has having such property—while nonetheless characterizing it has having some other property(ies) it has essentially. This seems compatible with materialism.

I have got some worries with this. The major worry I have is that there seems to be something wrong in taking *Rev* to be *primarily* and *only* a feature of concepts—rather than a feature of mental *states* or *events*—in the very first place. I will come to that in a moment. At any rate, as Goff (2011: 197) argues, it is dubious whether taking phenomenal concepts to offer only a *partial* essential characterization—or, which is the same, to be *translucent* rather than *transparent*—really can help the (*a posteriori*)[9] physicalist. In fact, claiming that phenomenal properties are *wholly physical* ones, the physicalist is committed to say that *any component* of properties is wholly physical. A part of something *wholly* physical is wholly physical. Thus, even if phenomenal concepts were to reveal only an essential part of their referents, they should reveal such part to be physical, which they clearly don't.

3. More on *C-Rev*

The roots of *C-Rev* are to be traced back to Kripke's (1980) and Putnam's (1975) seminal work in the Seventies. According to what may be called the received theory of reference and meaning, the intension of a term/concept—namely, the peculiar manner in which the referent of that term/concept is selected—*determines* the referent/the set of referents of that term/concept—its *extension*—by fixing a set of conditions—and, some authors (e.g., Carnap 1947) argue, even a set of *criteria*—for being that referent or for belonging to that set; it follows that while two terms may have the same extension but different intensions—as in the 'creature with a kidney' versus 'creature with a hearth' case—the reverse cannot be the case: for two terms to differ in extension is for them to differ in intension.

Against what the received theory would hold, both Kripke and Putnam urged us not to conflate the way in which the reference of a notion is fixed (in a given possible world when that world is taken as actual)—which pertains to the epistemological/psychological domain and might be said to coincide, with some level of approximation, with what Chalmers (e.g., 1996; 2009) calls the *primary intension* of a concept—with the referent(s) of that notion, let alone its/their essence—which instead pertain to the metaphysical domain and is labelled by Chalmers as the *secondary intension* of the notion.

Severing the epistemological domain from the metaphysical one leads Kripke to conclude that *necessary a posteriori* judgments can indeed be formed

[9] Chalmers (1996) distinguishes between *type-a*—or *a priori*—and *type-b*—or *a posteriori*—materialism. Although type-a views come in a broad range of varieties, they share the claim that the mental is logically supervenient on the physical, *i.e.*, is always possible to *a priori* deduce facts about consciousness from physical facts. Typically, type-a theorists deny both that phenomenal zombies are conceivable and that Mary learns anything new once set free from her black-and-white prison. Maintaining (at the very least) that phenomenal facts *metaphysically* supervene upon physical ones, type-b materialists, in turn, concede that consciousness is not *logically* supervenient on physical facts, *i.e.*, they accept the so-called standard story of the explanatory gap (Levine 1983; Schroer 2010).

and justified (*contra* Kant, 1781 [2016])*:* in fact, the *a priori*/*a posteriori* distinction is epistemological in scope, whereas the notion of necessity is metaphysical.

The reference of a term/concept can be fixed in various ways, namely, via an "original" ostensive gesture/baptism—as is typically the case with personal proper names such as 'Francesco'—or by pointing to a property that is or seems to be shared by all the members of a given sample, or a number of them—as is typically the case with those concepts which refer to natural kinds such as 'HEAT', 'BRONZE' or 'TIGER'. Also, the way in which the reference of a notion is fixed—*i.e.* the primary intension of the notion—does not, most often, depend from empirical factors: Being that which *determines* the way in which the actual world should turn out to be in order for a given concept to have a certain extension, it is not itself *dependent upon* how the actual world turns out to be.[10] A term like 'water' will therefore have the same primary intension both in the actual world and in TWIN-EARTH: in all the possible worlds in which it is not void, in fact, it picks the clear drinkable liquid which fills the oceans, etc.

The secondary intension of a notion, by contrast, *does* depend upon empirical factors: one needs to do research to get to know that water is actually H_2O rather than XYZ. TWIN-EARTH, thus, is not a world where water is XYZ; rather, it is just a world without water, or better, a world in which something that *is not* water merely gets *called* 'water'. In light of this, the judgment 'water is H_2O' is necessary—*i.e.*, true in all possible worlds—but still *a posteriori*, as it is justified empirically.

Crucially, both Kripke (1980*:*150-52) and Chalmers (e.g., 1996: 131) agree that phenomenal notions do constitute a notable exception to the framework I have just tried to outline.

In most cases, in fact, the referent of a term/concept is picked by pointing towards a property which belongs *only contingently* to that which is referred to by it. The primary intension of a concept like 'HEAT', for instance, would be something like 'the phenomenon which causes the sensation S in humans'. A certain amount of empirical research having been done, we now know that heat essentially is molecular motion, whereby we are able to identify 'molecular motion' as the secondary intension of the concept 'HEAT'. Heat is thus identical to molecular motion in any possible world, including a world populated with creature whose somatosensory apparatus does not produce the experience S, or even one with no conscious subject at all.

Consider now a state like pain. The referent of the concept like 'PAIN' is presumably fixed by pointing towards a class of experiences which share the same phenomenology, namely, *painful* experiences. This is not a contingent property of pain, though: to be an experience with a "painful phenomenology" *just is* to be an instance of pain. Phenomenal concepts, thus, have identical primary and secondary intensions—thus being transparent/providing a full essential characterization of their referents. In other words, in the case of phenomenal consciousness the epistemological sphere collapse on the metaphysical and *vice versa*. To conceive a world in which people are acquainted with the feeling of pain, again, *just is* to conceive a world where there is pain.

[10] Also, the primary intension of a notion fixes an *explanandum*. If I were to ask someone "what is water?", I would in effect be asking her to explain to me what the liquid transparent thing, which fills the oceans, etc., is.

Liu (forth.; see also Pitt: 2011: 146) labels the principle whereby there is no distinction, in phenomenal consciousness, between appearance and reality NARD (No Appearance-Reality Distinction). A formulation of NARD can already be found in Nagel (1974: 444-45); other formulations of it are also spelled out in Searle (1997: 456) and Horgan (2012: 406), among others. Most notably, however, NARD has been made famous by the arguments Kripke draws for it in his (1980).

What I wish to suggest is that while it is true that phenomenal concepts are transparent/provide a full essential characterization of their referent(s) in the sense given above,[11] this feature of them rests on a form of non-propositional knowledge—acquaintance-knowledge—of phenomenal properties. Let me unpack this.

4. Introspective Knowledge by Acquaintance: Causal Versus Constitutive Approaches

Russell (1910; 1912) distinguished between two kinds of knowledge one might have: *knowledge of truths* and *knowledge of things*. Knowledge of truths is ordinary propositional knowledge, *i.e.*, the kind of knowledge one has when she *knows that* something is the case*,* e.g., that Joe Biden is the president of the United States. Knowledge of things, instead, is a kind of *objectual* knowledge: what one knows in knowledge of things is an *item,* rather than a (body of) proposition(s). In turn, knowledge of things can be of two kinds: *knowledge by acquaintance* and *knowledge by description*. Knowledge by description is grounded on the subject having at least some propositional knowledge concerning the item she knows. Knowledge by acquaintance, on the other hand, does not depend on the subject forming *any propositional judgment* about the item she knows. It is also described by Russell as a kind of *direct* knowledge: in acquaintance we are immediately and directly presented with specific (mental) particulars.

Accordingly, *introspective* knowledge by acquaintance will be defined as the kind of knowledge we have of what we are directly aware of—or presented with—in introspection. There is not, still, unanimous consensus on what objects of introspection are, namely, on what is that one would allegedly have access to via introspection. In his (1910: 110) Russell claims that the objects of introspection are complexes consisting of objects plus various cognitive and conative relations we entertain towards them. So, in seeing the sun and introspecting her visual act, one would become aware both of the sun itself and of her seeing the sun. In (1912), in turn, Russell explicitly says that what we are aware of in introspection are the sense-data which make up physical objects, at least when we introspect our own perceptual states. Here, unless otherwise specified, along with Giustina (2022)—among others—I will assume that the objects of introspection are one's own conscious states.

Now, there are at least two possible ways to construe the expression 'knowledge by acquaintance'—thus the notion of knowledge by acquaintance itself, namely, a *causal* approach and a *constitutive* one. According to a causal ac-

[11] Whether only phenomenal concepts are transparent is debatable. Goff (e.g., 2011; 2017) argues that geometrical concepts—e.g., the concept 'SQUARE' 'TRIANGLE' etc.—are also of this sort.

count the relation of acquaintance—*i.e.*, a kind of direct and immediate access to specific (mental) particulars—is only epistemically relevant inasmuch as it causes, or enables, or justifies knowledge by acquaintance but is not epistemically relevant *per se* (see Depoe 2018; Hasan and Fumerton 2020; Gertler 2011). Moreover, causal views typically take knowledge by acquaintance to be propositional, therefore not *sui generis* (Giustina 2022). A given piece of knowledge is *sui generis* iff it cannot be reduced to any other kind of knowledge. According to a causal approach to knowledge by acquaintance, thus, the only possible *sui generis* kinds of knowledge available to a subject are propositional knowledge and (possibly) knowledge-how.

Under a constitutive account of knowledge by acquaintance, instead, the expression 'knowledge by acquaintance' is interpreted as 'the knowledge which is *constituted* by acquaintance'. Thus, these views take the relation of acquaintance to *be, in itself,* a *sui generis* kind of knowledge. Constitutive views, although still regarded as heterodox, are now beginning to gain currency, and are held by (among others): Duncan (e.g., 2020; 2021), Giustina (e.g., 2021; 2022), Fiocco (2017), Coleman (2019). This is also the view Russell (1910; 1912) most likely had.

I do believe that a constitutive account of introspective knowledge by acquaintance offers the best explanation for the transparency of phenomenal concepts. There are a number of arguments that may be provided for a *A-Rev*. In what follows I will mention those which strike me as more cogent.

4.1 Ordinary Propositional Knowledge and (Non-Propositional) Knowledge by Acquaintance Have an Analogous Normative Status (Duncan 2020, 2021)

As Duncan (2020: 7 and below) notes, phenomenal experience *simpliciter* seems to display several "hallmarks" which give the impression of a "rational or otherwise normative status parallel to that of justification for beliefs". For instance, it seems that at least some of our perceptions and/or somatosensory states can be rationally adjusted and are under our voluntary control—at least up to a certain extent: we can selectively focus on certain specific aspects or components of the perceptual field we are acquainted with, use learning and habituation to improve our capabilities of discrimination, discard hallucinations or optical illusions as non-veridical, and so on.

Moreover, the more attentively one introspects her own experiences, the larger the amount of details and of (non-propositional) information she will be put in a position to detect and get to know. (see Giustina 2022: 20)
Thus, on this approach 'to justify an experience' would amount to providing reasons for its veridicality (e.g., "I was paying attention").

4.2 An Argument for the Best Explanation (Giustina 2022)

Giustina (2022) argues that taking introspective acquaintance-knowledge to be *sui generis* provides the best explanation for cases where there is—or there seems to be—an epistemic asymmetry between subjects which cannot be exhaustively explained by an appeal to differences in the amount of propositional knowledge those subjects have.

People who are affected by an extremely rare pathological condition called *congenital analgesia* cannot experience physical pain. Suppose now you're trying to get a congenital analgesic to know what pain feels like. Arguably, no matter how hard you try, you won't manage to convey an informative, non-circular and non-trivial (e.g., "pain is painful") characterization of the peculiar qualitative character of pain. Imagine now a possible world—call it NON-PROPOSITIONAL-EARTH—where people, although capable to introspect their phenomenal experiences, for some reason—say due to how their cognitive architecture is structured—are unable to form any propositional judgment about them. Take now a subject A and a subject B on non-propositional earth and suppose that A has felt pain at least once in his life whereas B has not. According to Giustina (2022) there would still be, in NON-PROPOSITIONAL-EARTH, an epistemic asymmetry between A and B that is taking to the one there is between you and the congenital analgesic.[12]

4.3 The Argument for Phenomenal Concept Acquisition (Giustina 2021)

Phenomenal concepts can either be *basic* or *non-basic*. *Basic* phenomenal concepts provide the foundational layer upon which all other phenomenal concepts are formed (Giustina 2021: 7). The class of phenomenal concepts include concepts like 'PHENOMENAL YELLOW', 'OLFACTORY EXPERIENCE', 'THIRST', 'HOT' and so on. *Non-basic* phenomenal concepts, by contrast, are formed by combining basic ones: these are concepts like 'EXCRUCIATING ITCHING', 'BITTERSWEET GUSTATORY EXPERIENCE', 'PHENOMENAL ORANGE' etc. The argument from phenomenal concepts acquisition for the existence of a *sui generis* kind of introspective acquaintance-knowledge of the what-it's-like-ness of phenomenal experiences has the following form (Giustina 2021: 8): Unless one wants to buy a very implausible form of nativism whereby *all* or *the vast majority of* our phenomenal concepts—including 'MELANCHOLY' or 'PHENOMENAL RED'—were innately possessed by us, we must concede that (almost) all basic phenomenal concepts are acquired. Moreover, it is most likely that they are acquired via introspection. If all introspective states were conceptual/propositional in nature, however, it could not be the case that most of our basic phenomenal concepts were acquired via introspection, therefore we must conclude that at least some of our introspective states are not conceptual/propositional in kind.

4.4 The Argument(s) from Immediate Identification of Conscious Mental Particulars (Pitt 2004, 2009, 2011, 2019)

The arguments Pitt draws in his (2004; 2009; 2011; 2019) are mainly aimed at defending the existence of a proprietary, distinctive, and individuative phenom-

[12] I do think this argument to be reminiscent of the knowledge-argument against materialism. Pitt (2011: 148) writes: "When Mary leaves the Black and White Room, she comes to know what it's like to see red when she experiences it. In having the experience of red, she acquaintance-knows what seeing red is like". Note that what Pitt calls 'acquaintance-knowledge' arguably corresponds, with some level of approximation, to Giustina's notion of primitive introspection.

enology of cognitive states. I do believe, however, the remarks he makes to apply to more paradigmatic instances of phenomenal states as well.

Dretske (1969) has drawn a distinction between *simple seeing* and *epistemic seeing*. A subject S *simply sees* an object *O* iff she is able to differentiate it from its immediate environment immediately and non-inferentially, that is, only on the basis of how it looks to her. For Dretske, one does not need to identify[13]—*i.e.*, know—what is that she is seeing in order to be able to differentiate it from its environment in such an immediate way, as this ability does not require the formation of any explicit judgment[14] a given perceptual content—say, an apple—just "strike" one as different from its immediate surroundings—the table, the pen…—it appears so independently of whether one does know that it is an apple that she is seeing or not. Thus, for Dretske, simple seeing does not amount to knowledge. In order for *S* to see that *O* is *F* by being acquainted with it—have knowledge by acquaintance of it—a number of conditions must verify.[15]

Now, it is the opinion of Pitt (e.g., 2004) that the distinction between simple seeing—which is a form of simple acquaintance—and epistemic seeing—knowledge by acquaintance—can be generalized not only to other kinds of sensory experiences but to any kind of conscious state whatsoever, including cognitive states. But, Pitt's (*Ibid.*) argument goes on, this would not be possible unless those states had a proprietary, distinctive and individuative phenomenology, thus we must conclude they have one. Dretske (1969) is clearly in favour of a causal reading of the notion of knowledge by acquaintance in the sense specified above. I have a couple of remarks on this, though.

(1) I do agree with Pitt that the distinction between simple acquaintance and (propositional) knowledge by acquaintance can be generalized to all kinds of conscious mental particulars. If this is the case, though, I really cannot see how one could be able to differentiate the phenomenal properties of her own experiences from each the others without *eo ipso* somehow (non-conceptually, non-propositionally) *identifying* them: to be able to differentiate a phenomenal property—say the redness of an apple—from others she is or has been acquainted with, one must recognize those properties as not identical—e.g., the redness as not identical to the brownness of the table, nor to the painfulness of the headache she has, and so on. I do think this should be regarded, if not as a full-fledged form of knowledge, at the very least as a cognitive achievement *by itself*.

[13] Pitt (2004; 2009; 2011) says that in being attentively aware of her own conscious states one is immediately—*i.e.*, without the intermediary of any explicit judgment or reflection—able to *identify* her own experiences—e.g., to distinguish each of them from the others. This choice of words strikes me as particularly interesting, as Lewis (1995) refers to revelation as 'the *identification* thesis'.

[14] Likewise, as we have seen, for Levine (2011: 108) implicit self-knowledge of thought "is not the result of any explicit formulation or reflection".

[15] S is said to see that *O* is *F* iff: (i) *S* simply sees *O* (*i.e.*, is acquainted with *O*); (ii) *O* is *F*; (iii) the conditions under which *S* simply sees *O* are such that it would not look to *S* as it does unless it were *F*; (iv) *S* believes (iii) to obtain; (v) *S* believes *O* to be *F*. Notice, also, that *O* does not necessarily have to appear as *F* to *S* in order for her/him to see (i.e., have knowledge by Acquaintance) that it is *F*: in fact, a given object *O*—say an apple—might appear, e.g., brown to me but I might know that—say, due to a particular law of refraction of the light in this room—it would not appear brown unless it were red, thereby knowing that it is red *via my being acquainted with his brownness*.

(2) Is there something more obvious than the fact 'being painful' is an essential property of an experience or pain, or that 'being red'—where 'red' here refers to a specific phenomenal quality—is essential for an experience of a red surface to be the experience it is (see Kripke 1980: 150-52)?

5. Concluding Remarks

I do think that taking the awareness we have of our own phenomenal mental states to constitute *per se* a peculiar kind of knowledge and taking this knowledge to be essence-revealing might have severe implications upon a materialist framework broadly construed about phenomenal consciousness and about reality in general.

I do believe that *Rev* threatens what Coleman (2008) calls conventional physicalism, namely, a view which consists in the combination of a positive claim and a negative one: phenomenal properties supervene upon (fundamental) physical ones and there are no fundamental phenomenal properties—i.e., the view known as Russellian Monism is false. Since it is widely acknowledged in the relevant literature that any coherent form of physicalism must at the very least imply the supervenience of phenomenal properties upon physical ones, if conventional physicalism is threatened, *a fortiori* more committed forms of physicalism such as the one that Damnjanovic (2012) defends—*i.e.*, form that spell out the relation between phenomenal and physical properties in terms of *identity* or *grounding*—are also threatened. I have also suggested that phenomenal properties should be considered as essential properties of the state bearing them both under a definitional/non-modal and under a modal account of essentiality.

As Giustina (e.g., 2022) has noted, contemplating the idea that the relation of acquaintance is *in itself* peculiar a kind of knowledge might be a way of gaining new insights on how we understand the notion of knowledge in the very first place.

There is a number of issues left open that might be worth to be addressed in the future, spanning from metaphysical issues (do phenomenal concepts provide only a *partial* essential characterization of their referents—phenomenal properties—or do they provide a *full* essential characterization of those properties? is partial *Rev* compatible with physicalism?) to issues in epistemology (does acquaintance *alone* suffice for knowledge? What is the role of attention in introspective acquaintance-knowledge? Is introspective acquaintance-knowledge infallible? Is it knowledge of types or knowledge of tokens? Is introspective acquaintance-knowledge the only kind of acquaintance-knowledge one might have or are there other possible kinds of acquaintance-knowledge? What about, for instance, *perceptual* acquaintance-knowledge, *intuitional* acquaintance-knowledge and so on? How can one use acquaintance-knowledge to build a specific repertoire of concepts? And in particular, how does one use it to build a repertoire of concepts that are at the very least translucent if not transparent?) and even to issues in aesthetics and the philosophy of art (can one imagine phenomenal experiences she has never been acquainted with? can art elicit acquaintance-knowledge?)

I also do think that envisaging the possibility that our epistemic access to our minds and to reality outstrips the possibilities of our propositional knowledge may bring us to reconsider the role of the humanities and of the liberal arts in the academia and in our cognitive endeavour overall. Lodge (2003)

has argued that literature can offer a type of knowledge that is essential and complementary (not opposite) to scientific one. Paying to the view that experience is knowledge (Duncan, 2020) as well as to *Rev* the attention they merit may help further develop Lodge's ideas: in producing e.g., an emotional condition in those who read them, great novelists and poets would not just be merely entertaining us: they would as a matter of fact be revealing to us nature of our very own conscious states and thus, ultimately, of ourselves.[16]

References

Armstrong, D.M. 1968, *A Materialist Theory of Mind*, New York: Humanities Press.

Balog, K. 2012, "Acquaintance and The Mind-Body Problem", in Hill, C. and Gozzano, S. (eds.), *New Perspectives on Type Identity: The Mental and the Physical*, Cambridge: Cambridge University Press, 16-43.

Bayne, T. and Montague, M. (eds.) 2011, *Cognitive Phenomenology*, Oxford: Oxford University Press.

Bigelow, J. and Pargetter, R. 1990, "Acquaintance with Qualia", in Ludlow, Stoljar, and Nagasawa 2004, 179-98.

Block, N. 1995, "On a Confusion about the Function of Consciousness", *Brain and Behavioural Sciences*, 18, 227-47.

Brogaard, B. 2011, "'Knowledge-How: A Unified Account'" in Bengson, J. and Moffett, M. (eds.), *Knowing How: Essays on Knowledge, Mind, and Action*, Oxford: Oxford University Press, 136-60.

Brogaard, B. and Salerno, J. 2007a, "Why Counterpossibles Are Non-Trivial?", *The Reasoner*, 1, 1, 5-6.

Brogaard, B. and Salerno, J. 2007b, "Knowability, Possibility and Paradox", in Hendricks, J. and Pritchard, D. (eds.), *New Waves in Epistemology*, Basingstoke: Palgrave Macmillan, 270-99.

Brogaard, B. and Salerno, J. 2013, "Remarks on Counterpossibles", *Synthese*, 190, 639-60.

Byrne, A. 2017, "The Epistemic Significance of Experience", *Philosophical Studies*, 174, 947-67.

[16] I wish to sincerely thank the anonymous reviewers of this paper for having devoted some of their time and of their attention to my work. Versions of this paper have been presented at the IUSS *Nidi* 2021 cycle of seminars, at the *6th Italian Conference in Analytic Metaphysics and Ontology* hosted by the University of L'Aquila, at the *2022 Conference of the Australasian Association of Philosophy*, at the *19th Conference of the Francophone Society for Analytic Philosophy* hosted by the university of Neuchâtel, and the 2022 *SoPhia Salzburg Conference for Young Analytic Philosophy* hosted by the University of Salzburg: let me thank all those who have attended my talks for having listened with interest and attention and for having provided me with their feedback, some of which have really helped me deepen my thoughts. This work could not have been written without the help of my Supervisors Prof. Alfredo Tomasetta and Prof. Michele Di Francesco. It has also benefited a lot from several conversations I had with my colleagues and dear friends Simone Nota, Jacopo Pallagrosi, Arianna Beghetto, Giacomo Zanotti, and Marco Facchin. Some of them even read previous drafts of the paper and provided me with their comments. To all of them I am deeply grateful.

Carnap, R. 1947, *Meaning and Necessity: A Study in Semantics and Modal Logic*, Chicago: University of Chicago Press.

Chalmers, D.J. 1996, *The Conscious Mind: In Search of a Theory of Conscious Experience*, New York: Oxford University Press.

Chalmers, D.J. 2009, "The Two-Dimensional Argument Against Materialism", in McLaughlin, B.P., Beckermann, A. and Walter, S. (eds.), *The Oxford Handbook of Philosophy of Mind*, Oxford: Oxford University Press, 313-36.

Chalmers, D.J. 2017, "The Combination Problem for Panpsychism" in Godehard, B. and Jaskolla, L. (eds.), *Panpsychism: Contemporary perspectives*, New York: Oxford University Press, 179-214.

Churchland, P. 1989, "Knowing Qualia: A Reply to Jackson", in Ludlow, Stoljar, and Nagasawa 2004, 163-78.

Coleman, S. 2009, "Mind under Matter", in Skrbina, D. (ed.), *Mind that Abides*, Amsterdam: John Benjamins Publishing, 83-108.

Coleman, S. 2019, "Natural Acquaintance", in Raleigh, J.K. (ed.), *Acquaintance: New Essays*, Oxford: Oxford University Press, 49-74.

Conee, E. 1994, "Phenomenal Knowledge", in Ludlow, Stoljar, and Nagasawa 2004, 197-216.

Correia, F. 2007, "(Finean) Essence and (Priorean) Modality", *Dialectica*, 61, 63-84.

Damnjanovic, N. 2012, "Revelation and Physicalism", *Dialectica*, 66, 69-91.

Dasgupta, S. 2014, "The Possibility of Physicalism", *The Journal of Philosophy*, 111, 9/10, 557-92.

DePoe, J.M. 2018, "Knowledge by Acquaintance and Knowledge by Description", *Internet Encyclopedia of Philosophy*, http://www.iep.utm.edu/knowacq.

Dretske, F. 1969, *Seeing and Knowing*, Chicago: The University of Chicago Press.

Duncan, M. 2020, "Knowledge of Things", *Synthese*, 197, 3559-92.

Duncan, M. 2021, "Experience is Knowledge", in Kriegel, U. (ed.), *Oxford Studies in Philosophy of Mind*, 1, Oxford: Oxford University Press, 106-29.

Feigl, H. 1958. "The 'Mental' and The 'Physical'", *Minnesota Studies in the Philosophy of Science*, Minneapolis: University of Minnesota Press, 370-497.

Fine, K. 1995a, "Ontological Dependence", *Proceedings of the Aristotelian Society*, 95, 1, 269-90.

Fine, K. 1995b, "Senses of Essence", in Sinnott-Armstrong, W., Raffman, D., and Asher, N. (eds.) 1995, *Modality, Morality and Belief: Essays in Honor of Ruth Barcan Marcus*, Cambridge: Cambridge University Press, 53-73.

Fodor, J. 1975, *The Language of Thought*, New York: Cromwell.

Gertler, B. 2011, *Self-Knowledge*, New York: Routledge.

Giustina, A. 2021, "Introspection Without Judgment", *Erkenntnis*, 86, 407-27.

Giustina, A. 2022, "Introspective Knowledge by Acquaintance", *Synthese*, 200, 2, 1-23.

Goff, P. 2011, "A Posteriori Physicalists Get Our Phenomenal Concepts Wrong", *Australasian Journal of Philosophy*, 89, 2, 191-209.

Goff, P. 2015, "Real Acquaintance and Physicalism", in Coates, P. and Coleman, S. (eds.) 2015, *Phenomenal Qualities: Sense, Perception, and Consciousness*, Oxford: Oxford University Press, 121-43.

Goff, P. 2017, *Consciousness and Fundamental Reality*, Oxford: Oxford University Press.

Hasan, A. and Fumerton, R. 2020, "Knowledge by Acquaintance vs. Description", in *The Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/archives/spr2020/entries/knowledge-acquainde scrip/.

Horgan, T. 2012, "Introspection and Phenomenal Consciousness: Running the Gamut from Infallibility to Impotence", in Smithies, D. and Stoljar, D. (eds.), *Introspection and Consciousness*, Oxford: Oxford University Press, 405-22.

Jackson, F. 1982, "Epiphenomenal Qualia", *The Philosophical Quarterly*, 32, 127, 127-36.

Jackson, F. 2008, *From Metaphysics to Ethics*, Oxford: Clarendon Press.

Johnston, M. 1992, "How to Speak of the Colors", *Philosophical Studies*, 68, 3, 221-63.

Kant, I. 2016, *Critique of Pure Reason*, edited by Guyer, P. and Wood, A.W., Cambridge: Cambridge University Press.

Kripke, S. 1980, *Naming and Necessity*, Oxford: Basil Blackwell.

Kroedel, T. and Schulz, M. 2016, "Grounding Mental Causation", *Synthese*, 193, 6, 1909-23.

Levine, J. 1983, "Materialism and Qualia: The Explanatory Gap", *Pacific Philosophical Quarterly*, 64: 354-61.

Levine, J. 2011, "What is Cognitive Phenomenology, and Do We Have it?", in Bayne and Montague 2011, Chapter 5.

Lewis, D. 1983, "New Work for a Theory of Universals", *Australasian Journal of Philosophy*, 61, 4, 343-77.

Lewis, D. 1995, "Should a Materialist Believe in Qualia?", *Australasian Journal of Philosophy*, 73, 1, 140-44.

Liu, M. 2019, "Phenomenal Experience and the Thesis of Revelation" in Shottenkirk, D., Curado, M., and Gouveia, S.S. (eds.), *Perception, Cognition and Aesthetics*, New York: Routledge, 227-51.

Liu, M. 2020, "Explaining the Intuition of Revelation", *Journal of Consciousness Studies*, 27, 5-6, 99-107.

Liu, M. 2021, "Revelation and The Intuition of Dualism", Synthese, 199, 3-4, 11491-11515.

Liu, M. forthcoming, "Revelation and the Appearance/Reality Distinction", in Kriegel, U. (ed.), *Oxford Studies in Philosophy of Mind*, 4.

Lodge, D. 2003, *Consciousness and The Novel*, London: Vintage.

Ludlow, P., Stoljar, D., and Nagasawa, Y. 2004 (eds.), *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*, Cambridge, MA: The MIT Press.

Majeed, R. 2017, "Ramseyan Humility: The Response from Revelation and panpsychism", *Canadian Journal of Philosophy*, 47, 1, 75-96.

Marcus, R.B. 1967, "Essentialism in Modal Logic", *Noûs*, 1, 91-96.

Melnyk, A. 2003, *A Physicalist Manifesto: Thoroughly Modern Materialism*, New York: Cambridge University Press.

Melnyk, A. 2006, "Realization and the Formulation of Physicalism", *Philosophical Studies*, 131, 1, 127-55.

Melnyk, A. 2018, "In Defense of a Realization Formulation of Physicalism", *Topoi*, 37, 3, 483-93.

McGinn, C. 2008, "Consciousness as Knowingness", *The Monist*, 91, 237-49.

Montero, B.G. 2013, "Must Physicalism imply the Supervenience of The Mental on The Physical?", *The Journal of Philosophy*, 110, 2, 93-110.

Montero, B.G. and Brown, C. 2018, "Making Room for a This-Worldly Physicalism", *Topoi,* 37, 3, 523-32.

Nagel, T. 1974, "What Is It like to Be a Bat?", *The Philosophical Review*, 83, 4, 435-50.

Nida-Rümelin, M. 2007, "Grasping Phenomenal Properties", in Alter, T. and Walter, S. (eds.) 2007, *Phenomenal Concepts and Phenomenal Knowledge: New essays on Consciousness and Physicalism*, Oxford: Oxford University Press, 274-307.

O'Conaill, D. 2018, "Grounding, Physicalism and Necessity", *Inquiry*, 61, 7, 713-30.

Pereboom, D. 2011, *Consciousness and The Prospects of Physicalism*, New York: Oxford University Press.

Pitt, D. 2004, "The Phenomenology of Cognition or 'What is it Like to Think That p?'", *Philosophy and Phenomenological Research*, 69, 1-36.

Pitt, D. 2009, "Intentional Psychologism", *Philosophical Studies*, 146, 117-38.

Pitt, D. 2011, "Introspection, Phenomenality and the Availability of Intentional Content", in Bayne and Montague 2011, chapter 7.

Pitt, D. 2019, "Acquaintance and Phenomenal Concepts", in Coleman, S. (ed.), *The Knowledge Argument*, Cambridge: Cambridge University Press, 87-101.

Place, U.T. 1956, "Is Consciousness a Brain Process", *British journal of psychology*, 47, 1, 44-50.

Putnam, H. 1967, "Psychological Predicates", in Capitan, W.H. and Merrill, D.D. (eds.), *Art, Mind and Religion*, Pittsburgh: University of Pittsburgh Press, 37-48.

Putnam, H. 1975, "The Meaning of 'Meaning'", in *Mind, Language and Reality*, *Philosophical Papers*, Volume 2, Cambridge: Cambridge University Press, 215-71.

Robertson Ishii, T. and Philip, A. 2020, "Essential vs. Accidental Properties", *The Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/archives/win2020/entries/essential-accidental/

Russell, B. 1910, "Knowledge by Acquaintance and Knowledge by Description", *Proceedings of the Aristotelian Society*, 11, 1, 108-28.

Russell, B. 1912, *The Problems of Philosophy*, New York: H. Holt and Company.

Schroer, R 2010, "What's the Beef? Phenomenal Concepts as Both Demonstrative and Substantial", *Australasian Journal of Philosophy*, 88, 505-22.

Searle, J. 1997, "Reductionism and The Irreducibility of Consciousness", in Flanagan, O.J., Block, N., and Guzeldere, G. (eds.), *The Nature of Consciousness*, Cambridge, MA: MIT Press, 451-60.

Shoemaker, S. 2007, *Physical Realization*, New York: Oxford University Press.

Smart, J.J. 1959, "Sensations and Brain Processes", *The Philosophical Review*, 68, 2, 141-56.

Stalnaker, R. 2008, *Our Knowledge of The Internal World*, Oxford: Oxford University Press.

Stanley, J. 2011, "Knowing (How)", Noûs, 45, 207-38.

Stanley, J. and Williamson, T. 2001, "Knowing How", *Journal of Philosophy*, 98, 411-44.

Stoljar, D. 2009, "The Argument from Revelation" in Braddon-Mitchell, D. and Nola, R. (eds.), *Conceptual Analysis and Philosophical Naturalism*, Cambridge, MA: The MIT Press.

Strawson, G. 1989, "Red and 'Red'", Synthese, 78, 2, 193-232.

Tomasetta, A. 2015, "Physicalist Naturalism in The Philosophy of Mind (far less Warranted Than Usually Thought)", *Discipline Filosofiche*, XXV: 89-110.

Tomasetta, A. 2016. "Knowledge by Experience: Or Why Physicalism Should not be Our Default Position in Consciousness Studies", *Rivista internazionale di filosofia e psicologia*, 7, 1, 37-47.

Trogdon, K. 2016, "Revelation and Physicalism", *Synthese*, 194, 7, 2345-66.

Williamson, T. 2000, *Knowledge and Its Limits*, Oxford: Oxford University Press.

Zalta, E. 2006, "Essence and Modality", *Mind*, 115, 659-93.

Zanotti, G. 2020, "Sul Perché l'Argomento Naturalista non Può Fornire Ragioni a Sostegno del Fisicalismo", *Sistemi intelligenti*, 32, 3, 575-96.

Zanotti, G. 2021, "Physicalism and the Burden of Parsimony", *Synthese*, 199, 3, 11109-32.

Zanotti, G. 2022, "Consciousness, Neuroscience, and Physicalism: Pessimism About Optimistic Induction", *Acta Analytica*, 2022, DOI: https://doi.org/10.1007/s12136-022-00512-5

The Feasibility Approach to Imagination as a Guide to Metaphysical Modality

Daniel Dohrn

University of Milan

Abstract

I present a novel approach to modal imagination as a means of knowing metaphysical possibilities. Hume calls the link between imagining and possibility an 'established maxim'. I ask: what makes it seem so natural to use imagination as a guide to modality? (1.) I draw some lessons on my motivational question from the current debate. (2.) I develop my answer: we use imagination to creatively simulate solutions to feasibility issues. (2.1.) To corroborate my answer, I consider everyday feasibility issues. (2.2.) I then extend the account to more remote feasibility issues. (2.3.) I point out a special connection between imagination and creativity (3.) I show how the feasibility approach bears on issues of metaphysical possibility. (3.1.) I outline how imagination allows to retrieve and test modal constraints. (3.2.) I support my argument by examples from the philosophical debate. (3.3.) I answer my original motivational question. (4.) I address objections.

Keywords: Imagination, Conceivability, Possibility, Modality.

There is a long-standing philosophical tradition of using imagination as a guide to modal knowledge ('modal imagination'). As Hume put it:

> Tis an establish'd maxim in metaphysics, *That whatever the mind clearly conceives includes the idea of possible existence*, or in other words, *that nothing we imagine is absolutely impossible* (Hume 1739-40: 1.2.2.8, 32).

Yet *what makes it seem so natural to use imagination as a guide to modality*? I shall develop one answer to this *motivational question*. I concentrate on metaphysical possibility, setting aside necessity for reasons of space.[1]

[1] An anonymous reviewer has reminded me that the notion of metaphysical possibility needs clarification. Metaphysical possibilities have been characterized as 'absolute' in the sense of being the most inclusive objective (as contrasted to epistemic, deontic) possibili-

Argumenta 10, 1 (2024): 127—144
ISSN 2465-2334

DOI 10.14275/2465-2334/20240.doh
First published: 30 November 2024

1. The Motivational Gap – Lessons from the Debate on Conceivability

Disregarding the historical connections to Hume etc., I list some aspects of modal imagination as discussed in recent literature. All these aspects are contentious, but I select those which I take to be most amenable to a non-sceptical answer to the motivational question:

(1) Modal imagination, imagination properly used to figure out possibility, is a subcase, distinguished from other uses of imagining, e.g. imagining epistemic alternatives (Yablo 1993). When I henceforth talk of imagination without further qualification, I have in mind modal imagination.

(2) Often it is emphasized that imagination recruits 'structural representations' (like diagrams, maps) as contrasted to 'conceptual' ones (Ichikawa and Jarvis 2012: 151). The epistemic contribution of imagination is sometimes even restricted to that of qualitative or quasi-perceptual content (see section 4). Still many authors take a more *holistic* approach. Imagination may recruit any mental resources in *simulating* some reality (Williamson 2007: 143), even canonical world descriptions (Chalmers 2002).

(3) Modal imagination is often described as *objectual* as contrasted to propositional. This does not mean that objectual imagination cannot proceed via describing its object. The object of imagining *p* may be a complete world verifying *p* (Yablo 1993, Chalmers 2002).[2]

(4) As far as the object of imagining *p* goes beyond *p*, imagination tends to come with elaborating a *p*-scenario in some detail (Yablo 1993, Chalmers 2002). There are doubts that we can elaborate far-fetched scenarios in *sufficient* detail, though (Van Inwagen 1998).

The picture drawn so far does not yet answer my motivational question. Imagining *p* is not obviously sufficient for *p* being possible.

Looking for a way to close the gap, I shall consider two exemplary ways of answering the motivational question as discussed in the literature.[3] The *first view*, advocated by Stephen Yablo, is that imagining raises an *appearance of possibility*:[4]

ties (Hale 1996). However, it has been argued that there are more inclusive objective possibilities such as the diverse systems of logical possibilities (e.g. Clarke-Doane 2021, Priest 2021). To deal with this problem, I suggest to understand 'absolute' in the sense of lifting any *contextual* constraints on circumstantial objective possibilities (as exemplified by the skunk and the mountaineering example to come), leaving only *general* metaphysical constraints. Logical possibilities do not result from such a process of lifting contextual constraints on circumstantial possibilities. It is to be seen whether the laws of nature form general metaphysical constraints, or whether their generality is more limited.

[2] I use '*p*' as a variable for propositions. However, I allow myself locutions like 'the possibility of *p*', 'a *p*-scenario' by which I mean the possibility *that p*, a scenario such *that p* and so on. I do not think that such loose talk is detrimental to my argument.

[3] See Evnine's distinction between two claims: imaginability entails possibility or it merely 'gestures in its direction' (Evnine 2008: 666). Yet I do not take my two alternatives to be exhaustive.

[4] Yablo uses 'imagining' to spell out 'conceiving'.

> Just as someone who perceives that *p* enjoys the appearance that *p* is true, whoever finds *p* conceivable enjoys something worth describing as the appearance that it is possible. In slogan form: conceiving involves the appearance of possibility (Yablo 1993: 5).

Later Yablo says:

> Just as to perceive that *p* is to be in a state that (i) is veridical only if *p*, and that (ii) moves you to believe that *p*, to find *p* conceivable is to be in a state which (i) is veridical only if possibly *p*, and (ii) moves you to believe that *p* is possible (Yablo 1993: 12).

Yablo's talk of an appearance of possibility seems a promising way of addressing the motivational issue. Perceptual seemings are a natural start for cognizing the world. The same may go for some presentational phenomenology coming with certain imaginings. However, several gaps remain to be filled. Firstly, even if imagination issues in an appearance of possibility, what motivates us to use imagination in the first place? How do we anticipate that it may come with such an appearance? In the case of perception, elementary seemings might be expected to be simply given. They spontaneously arise from external stimuli. But the same does not obviously go for using imagination.

Moreover, the appearance does not simply arise when we *somehow* represent *p*. We have to imagine a world verifying *p*. It is not a matter of course that we react to a possibility issue by imagining *a world*, and that we have an idea of how to do that. Yablo (1993: 37) suggests that we leave most of the world unspecified by treating it *as determinate*.[5] However, we may not simply treat any detail as determinate on pain of trivialization. The motivational issue rearises: how do we come to adopt a practice of imagining a world, treating irrelevant details as indeterminate and relevant details as determinate? One answer is that we use imagination to test *p* for *coherence* in a suitable sense. This brings me to the second view.

The *second view* is presumably most widespread, and it comes in several variants. It is a somewhat daring enterprise to lump these variants together, but I reckon it worth the attempt. The unifying idea is to use imagination for a *coherence* test.

One variant of this view is that there is a *rational* or *a priori* connection: ideal conceivability as given by a complete and coherent canonical world description entails possibility (Chalmers 2002). Being aware of this connection, we take our exercises of imagination as a test for ideal conceivability.

Another variant is that the connection is *conceptual* (Sidelle 1989, Ichikawa and Jarvis 2012). Conceptual knowledge provides access to a space of conceptual possibilities. We use imagination to check *p* for coherence with the constraints imposed by conceptual knowledge and empirical knowledge.

A third variant of the view uses the equivalency with counterfactuals: $\Diamond p \equiv \neg(p \Box\!\!\rightarrow \bot)$ ($\bot$ being a logical falsehood, Williamson 2007: 163). Reasoning in accordance with the equivalence is part of our competence of everyday counter-

[5] More precisely, Yablo distinguishes between ignoring the rest of the world as irrelevant and treating the fully determinate way in which *p* is realized as determinate. I use the *as determinate* clause so as to cover both alternatives.

factual reasoning. We imaginatively develop a counterfactual supposition. If we do not encounter a contradiction after sufficient development, we judge that *p* is possible.

All these approaches motivate the use of imagination only if we already appreciate certain connections between the possibility and the coherence of a scenario, be they rational, conceptual, or built into the logics of counterfactuals. I harbour the suspicion that this reverses the order in which we first come to know certain possibilities: we first have an immediate tendency to use imagination in a constrained way to figure out possibility; *then* we may come to appreciate the connection between the possibility of a scenario and its coherence.

In the next section, I shall propose an answer to the motivational question. The answer takes inspiration from both views, imagination seen as a coherence test and imagination as coming with an appearance of possibility. I shall build on several features mentioned so far to guide my discussion:

(1) Imagination may recruit any mental resources in simulating some reality.
(2) Imagination is object-directed.
(3) Imagining *p* comes with coherently fleshing a larger scenario that verifies *p*.
(4) When imagining informs modal belief, it does so by raising an appearance of possibility.

2. Imagination and Feasibility

2.1 Addressing Everyday Feasibility Issues

I shall answer the basic motivational question by pointing to the use of imagination in figuring out *practical solutions to feasibility issues*. The close connection between possibility and the feasibility of a course of action has been noticed before:

> Plausibly, the idea of possibility has a primitive association with action: the world at large determines how things *are*; we determine what to *do*, and in these episodes we take ourselves to choose from possibilities. From there, a sense of possibility projects backward and sideways. We see other events, including past events, as embedded in a cloud of ways- things- might-have-been… Action gives us the idea of possibility, and also an accompanying idea of dependence: *if I do this, things will go like that*. The forward models used in planning can also be applied to testing (*if I do this, I expect things to look like that—unless I am wrong*). The sense of possibility thus gains an epistemic role (Godfrey-Smith 2020: 166).

Godfrey-Smith points out that a capacity of exploring different possible courses of action may be evolutionarily hardwired and even be found in animals:

> …as rats make a spatial decision, they activate a collection of neural paths that sweep ahead of the animal's representation of its current position, running "first down one path and then the other," apparently representing future possibilities... (Godfrey-Smith 2020: 166).

As an example of how partly sensory imagination may be used to address everyday feasibility issues, I consider Neil van Leeuwen's (fictive?) report of how he encountered a skunk while on a run:

SKUNK:

I visualized the skunk spraying, imagined myself running across the street to a distance beyond where I imagined the skunk spray going, and then ran across along the route I had imagined (Van Leeuwen 2011: 69-70).

I shall assume that van Leeuwen's runner could indeed have used imagination to figure out a near-optimal route around the skunk. I shall work within the broad paradigm of imagination as a capacity of *simulating* aspects of reality, perhaps partly by re-creating mental processes like perception 'off-line' (see Currie and Ravenscroft 2002: 11; Williamson 2007). The runner simulates sensorimotor experience as of a not-yet actual reality in which he runs along a certain route. The imagined route tracks the contextually restricted possibility of pursuing one's course without entering the spraying range. It seems plausible that, at some point, the runner might have enjoyed an appearance of possibility, at least if he had pondered the question of feasibility: a distinctive appearance as of the route as feasible.

I shall try to remain as neutral as possible about the minutes of this appearance, but I follow Yablo in suggesting that it moves the runner to believe the route to be possible, and that it has the veridicality condition that the route indeed is possible. It is a matter of further debate whether the appearance may take the form of perceiving an affordance (Gibson 1966) or some sort of potentiality (X-ability, viability, Nanay 2011), and whether there is some implicit reasoning involved (Fodor and Pylyshyn 1981). I also hope to stay clear from commitments with regard to the debate on the format of imagery (see Pylyshyn 2002).

To prepare my transition to more theoretical possibilities, I shall stipulate that the runner first had a purely theoretical knowledge about the danger of getting sprayed and the circular spraying range of about 6m. Theoretical knowledge had to be translated into a structurally represented tangential curve. Bringing to bear his theoretical knowledge on the case, the runner faced the problem of how to adapt the goal of running straight to the unexpected obstacle. He used imagination to find a feasible way of overcoming the obstacle which optimally reconciled the goal with newly encountered constraints. The solution was easy but not trivial. It took a minimal innovative effort to figure it out.

To bring out the innovation, I add two comparisons. First, I contrast the imaginative effort to the formidable alternative of *calculating* the route in the abstract. Calculating would involve a substantial step, which is so much facilitated as to become barely noticeable by imaginatively manipulating the perceived situation. Second, I compare SKUNK to a related case:

MOUNTAIN:

A skilled climber is faced with the explicit issue of whether the north flank of a mountain can be ascended by free climbing. Looking at the mountain, she imaginatively traces several routes but finds them blocked. She makes an innovative effort to figure out a new route, being well-aware of her limited range of movement. At some point in her imaginative tracing of the route, she suddenly enjoys a positive appearance as of the route being feasible.

I suggest that MOUNTAIN is another typical and unproblematic example of an imagination-based appearance of feasibility. An effort of imagination is the most natural reaction to the feasibility problem. It is intimately linked to one's awareness of the obstacles on the route and the innovative effort to overcome them.

I discern a pattern which guides fleshing out a scenario. The climber begins with a dim awareness of the difficulties to be expected in attacking the flank. She has a general idea of the difficulty of overcoming gravity by climbing a near-vertical wall and the solution of exploiting friction with its uneven surface. But the best way to get into view the more *determinate obstacles* is to consider particular candidate routes. Imaginatively tracing one particular route will give the mountaineer a more concrete idea of the pertinent constraints imposed by the precise physical condition of the wall (angle, material…) and ways of meeting them (cracks, edges … to get a hold on). Generalizing: often modal constraints will not simply be manifest; our awareness of more determinate versions of these constraints depends on our going through exemplary ways for a possibility to be realized.

The cases described show imagination in its life function; why it is useful to have this capacity, when it is properly used, and how the modal use of imagination naturally arises: not yet as a response to abstract modal issues, but as an effort at solving a practical feasibility problem. In representing the solution as feasible, imagination takes on board all relevant information about the actual state of things but goes beyond that actual state in simulating some real situation that is not (yet) actual.

I shall list some characteristic features of using imagination for addressing a feasibility issue:

(1) We start from a concrete actual situation.
(2) A feasibility issue arises: we are more or less dimly alerted by some difficulties in achieving a goal.
(3) We set out to imaginatively simulate some particular solution: some not-yet-actual way to change the situation such that the goal is attained.
(4) The solution does not straightforwardly follow from our current informational state. It takes some innovation.
(5) Many details of the solution will be left open, though we may tend to fill the scenario with features of the actual world.
(6) Imagination works holistically: the simulation may recruit any informational resources and any mental capacity we have, in particular sensorimotor representation, but also propositional information.
(7) Different pieces of (partly tacit) information in different formats are activated, interact, and are transformed by concocting the imagined scenario.
(8) Our awareness of the more specific obstacles to be overcome viz. constraints to be met gradually emerges in the course of imaginatively developing the exemplary solution.
(9) Our imaginative effort is reliably constrained by our awareness of the obstacles: a phenomenology of feasibility ('appearance') arises only upon imagining a suitable solution.

2.2 More Remote Feasibility Issues

The feasibility issues considered arise from our perceptual acquaintance with concrete actual situations. Philosophical possibilities often completely detach from such situations. Still I suggest that the use of modal imagination preserves core features of the normal application of imagination to situational feasibility issues. The key function of imagination remains the simulation of some varia-

tion of reality by a creative albeit restrained departure. Certain additional tendencies distinguish the use of imagination from general theoretical inquiry, although both may go together and the distinction only be one of degrees:

(1) Imagination is *case*-directed. In responding to a possibility issue, it tends towards *simulating a concrete scenario* that confirms the possibility at issue. Yet the scenario will typically be left partly indeterminate. It can be multiply realized and thus is only treated *as if* it were an individual.
(2) Imagination is *holistic*. Due to its case-directedness, it tends towards sensorimotor representation, but it recruits any informational resources and any mental capacities that bear on a possibility issue; in particular, it is highly sensitive to information about the restrictions which delimit a solution to the issue at stake.
(3) The creative development aims at exploring ways of meeting the pertinent constraints and thus testing whether they preclude *p* from being possible. Our understanding of both these constraints and ways of satisfying them grows the more determinate the case imagined becomes.

To get these tendencies into view, I shall consider a new example. The use of imagination must not be confined to manipulating the perceived situation, and the general structure outlined should be transferred to theoretical/propositional content. I shall introduce a use of imagination meeting these conditions by anecdotal evidence. I do not aim at historical accuracy. Instead, I follow Amy Kind (2016: 154) in assuming that the case described is typical for the way imagination can be used:

TESLA: *Nicola Tesla's invention of the alternating current motor*.

Tesla's proficiency in using imagination was noted by his biographers:

> Before I put a sketch on paper, the whole idea is worked out mentally. In my mind, I change the construction, make improvements, and even operate the device. Without ever having drawn a sketch, I can give the measurement of all parts to workmen, and when completed these parts will fit, just as certainly as though I had made accurate drawings (O'Neill 1944: 257).

Tesla reportedly used his imaginative powers in a dispute with his teacher Poeschl in Graz whether a motor without a commutator was (technically) possible:

> In his mind he constructed one machine after another, and as he visioned them before him he could trace out with his finger the various circuits through armature and field coils, and follow the course of the rapidly changing currents (O'Neill 1944: 50).

The climax of the anecdote is that, taking a walk with a friend in Budapest, Tesla envisioned the working alternating current motor with a rotating magnetic field replacing the commutator, exclaiming:

> I have solved the problem. Can't you see it right here in front of me, running almost silently? It is the rotating magnetic field that does it (O'Neill 1944: 57).

Judging from his avowal, Tesla was under the impression of having solved the feasibility problem. The seeming he enjoyed was intimately connected with a visualization of the motor ('see… running silently').

There is also a deflationary reading of the case: Tesla's modal knowledge of the motor was entirely justified by an applied physical theory. Nevertheless I think that the following alternative has some plausibility: at some point, Tesla's justificatory basis for his feasibility claim was holistic. The state of the art in physics and engineering did not yet settle the dispute with Poeschl. Tesla's base comprised a partly explicit physical theory, but *as applied* to an imagined object. Tesla imaginatively simulated a concrete working exemplar of the motor; that objectual imagination first gave him the veridical appearance of possibility that rationalized his belief that the motor was feasible.

The imagination that rationalized Tesla's modal belief was the result of a series of efforts at creative problem-solving: 'In his mind he constructed one machine after another.' It took Tesla several trials to come up with a motor that satisfied the technical constraints. The trials formed a series of innovative steps. They were not simply pre-determined by the pertinent constraints. At each step, Tesla attained a better understanding of the technical constraints and ways to meet them. Eventually, Tesla 'saw' the last of these trial pieces running in accordance with the laws of electromagnetism. He enjoyed a positive appearance of possibility, coming with a case confirming this possibility.

The example illustrates the transition towards a more detached use of modal imagination. While still addressing an issue of practical feasibility, Tesla's visualization completely detached from his actual perceptual environment (the road in Budapest). It has been criticized that my feasibility approach is too centred on *imagining actions*. However, Tesla did not imagine how to build the motor. He imagined the motor itself working in a certain way. It took a further step to draw consequences for how to build the motor.

Before pursuing the continuity to issues of metaphysical modality, I shall add another motivational consideration, which further supports my focus on *creativity* as a main feature of the role of imagination in addressing feasibility problems.

2.3 Imagination and Creativity

I have emphasized that the use of imagination for solving practical problems is most pronounced when it takes some ingenuity to come up with a solution. Thus, I draw a close connection between imagination and creativity. I illustrate this association by results on *pretense*.

There are substantial differences between the exercise of imagination in many pretense games and in modal reasoning. Pretense may aim at verisimilitude, but it usually does not aim at settling possibility issues. Still I suggest that there are commonalities in the general function of imagination. One of them lies in creatively projecting an as-if particular situation. To illustrate the role of creativity in games of pretense, Nichols and Stich report an experiment in which participants were supposed to play waiters in a restaurant:

> WAITER:
> … in one of our fancy restaurant pretenses, the waiter pretended to decapitate one of the diners! A theory of pretense needs to be able to accommodate these

> kinds of elaborations as well as the more sober inferential elaborations (Nichols and Stich 2000: 119).

This shockingly unexpected albeit not illicit move in a standard pretense game testifies to the creative function of imagination. The use of imagination for solving feasibility issues explains this striking feature. It is part and parcel to the use of imagination in addressing non-trivial feasibility issues to generate and test innovative solutions. In contrast to normal feasibility issues, the pretense game is only minimally constrained by the premise of playing waiter. It invites eccentric ways of filling the role. The aspect of creativity prevails.

One may doubt that creativity is *part* of imagination. Imagination, one may say, only serves to spell out an independent pretense premise (or a supposition). Creativity lies only in coming up with the premise. Such doubts neglect that the continuous exercise of creativity is not simply a prerequisite but part of imaginative development. The idea of decapitating the guest may not have been premeditated but arisen spontaneously from enacting the pretense premise that one is a waiter.

To see creativity at work in the use of imagination to figure out metaphysical possibilities, I consider an example of Frank Jackson's. Jackson discusses how to assess

CAT: There could be a cat which is not an animal.

Jackson here is interested in questions of aprioricity, but his remarks are relevant to my discussion:

> Our failure to decide in advance how we would jump in fantastical, remote cases gives philosophers with their notorious ability to think up fantastical, remote cases, plenty of scope to come up with a case for which it is undecided whether, as it just might be, 'cat' and 'animal' apply, and so is a case where we can be induced, without going against anything determinate in the meaning of the terms, to apply, say, 'cat' and not apply, say, 'animal'. Thus, the case becomes one where cats are not animals (Jackson 1998: 54).

I use the quote to illustrate my main point: general metaphysical considerations and general conceptual analysis may be relevant. But such resources provide no alternative to imagining 'fantastic, remote *cases*' like perfect mechanical facsimiles of cats in order to test the metaphysical constraints on being a cat.

3. Feasibility and Metaphysical Possibility

3.1 The Feasibility Approach to Modal Constraints

I shall now generalize my feasibility approach to metaphysical possibility. A first requirement is *detachment*. In TESLA, I have illustrated how an exercise of imagination can detach from actual perception and interact with theoretical background knowledge. Nevertheless Tesla was still faced with a practical feasibility problem. In contrast, interesting metaphysical possibilities do not reduce to feasibility *for us*. Still there are relevant parallels. We have a tendency to tackle questions whether *p* is possible as *how possible?*-questions. Just as we imagine a solution to a feasibility issue, we more generally try to imagine *how it could be that p*.

One may frame such *how could it be?*-questions in a way that comes closer to feasibility issues. One main use of imagination is to put oneself into the shoes of other subjects.[6] For instance, the runner may imagine how far the skunk could spray. In metaphysical considerations, one may even detach from *any* normal subject. Philosophers sometimes raise issues of metaphysical necessity by asking what a *god* could have made real (e.g. Chalmers 2002: 146; Fine 2005: 259). In a similar vein, we may ask how an immensely powerful subject, call it nature, God, or a metaphysical engineer, could make it the case that *p* while abiding by metaphysical constraints.

Another key requirement for generalizing my everyday examples is to generalize the interplay between appreciating pertinent *constraints* on feasibility and envisaging *creative* solutions for how to meet them. We cannot simply presume these constraints to be manifest. We need empirical knowledge of the corresponding facts, and we need an awareness of their modal resilience. On pain of circularity, this awareness must not amount to outright modal knowledge, though (see Roca-Royes 2011).

I shall draw on Williamson's suggestion that the pertinent constraints are *implicit* in our imaginative exercise. Consider:

GOLD: Gold could have an atomic number different from 79.

> '…we need not judge that it is metaphysically necessary that gold is the element with atomic number 79 *before* invoking the proposition that gold is the element with atomic number 79 in the development of a counterfactual supposition. Rather, *projecting* constitutive matters such as atomic numbers into counterfactual suppositions *is part of our general way of assessing counterfactuals*. The judgment of metaphysical necessity originates as the output of a procedure of that kind; *it is not an independently generated input* (Williamson 2007: 170, m.e.).

To get a better idea of Williamson's suggestion, consider his account of the folk physics backing our everyday counterfactual assessments:

> …the folk physics needed to derive the consequents of counterfactuals such as [If the bush had not been there, the rock would have ended in the lake] from their antecedents may be stored in the form of some analogue mechanism, perhaps embodied in a connectionist network, which the subject cannot articulate in propositional form... the supposed premises may not be stored in a form that permits the normal range of inferential interactions with other beliefs, even at an unconscious level (Williamson 2007: 145).

Judging from this picture, our awareness of modal constraints is largely inexplicit and needs suitable cues to be activated. The constraints often need interpretation, precisification, and weighing, but such tasks of qualification cannot always be performed in the abstract. Often they can only be tackled by exploring suitable ways of embedding *p* (the possibility at issue) into an overall situation. The ways considered should help us with our limited minds to get a hold on the pertinent constraints.

To see how the difficulty of retrieving the pertinent constraints is addressed within the feasibility approach, consider again MOUNTAIN. The climber's at-

[6] I do not take stance on the theory vs. imagination debate on mindreading.

tention to the minutes of the route provides the right cues for her to become aware of the obstacles to be overcome. She starts with a general idea of the different kinds of obstacles arising in climbing a mountain. A more specific take on the pertinent obstacles will partly depend on specifying candidate routes. The result is a process of weighing. The mountaineer will adjust her route such as to overcome certain obstacles; more specific obstacles will emerge; and so on.

In a parallel vein, addressing a *how possible?* issue sharpens our sense for the metaphysical requirements of making something possible. One starts from a general take on the metaphysical restrictions that bear on *p*. But often this take will not be specific enough to be directly applied to the question whether *p* is reconcilable with the pertinent constraints. Sometimes it can be developed further by general metaphysical considerations. But if imagination is useful in addressing a possibility issue, this is because of the epistemic interplay between getting a grip on more determinate metaphysical constraints and coming up with a concrete solution of how they may be reconciled with *p* being true. The conceivability test takes the form of creatively rehearsing ways for *p* to be fitted into the metaphysical structure of the world.

3.2 Examples

I shall present some examples illustrating the creative use of imagination in metaphysics. My first example is a standard conceivability argument for possibility, Bohn on the possibility of junky worlds (everything is a proper part of something else). The example shows the maieutic aspect of the feasibility approach, making a solution palatable to our limited capacities:

> Now consider the following scenario. Everything in this world is spatially extended and just one half of something else that is also spatially extended. That is, for any thing in this world, there is something else of which it is a spatial proper part. Or consider this scenario. Our universe is a miniature replica universe housed in a particle of a bigger replica universe, which is again a miniature replica universe housed in a particle of an even bigger replica universe, and so on ad infinitum. Conceiving of these scenarios amounts to conceiving of worlds in which *everything* is a proper part. Let's call such worlds, *junky worlds*. Official definition: world *w* is *junky*$=_{df}$ anything in *w* is a proper part.
>
> Having thus conceived of junky worlds, we seem provided with some prima facie reasons to think such worlds are possible (Bohn 2009: 28).

Bohn does not simply ponder the possibility of a world in which there is no universal object, he uses imagination creatively to conjure up *two* recipes for how such a world could be made true, in one case putting halves together infinitely, in the other case a Chinese box- or matryoshka doll-like encapsulation of universes. These recipes are crafted such as to make the abstract mereological requirements of junk more accessible to us by an easy algorithmic structure: wholes are assembled from parts which have obvious and non-gerrymandered mereological features themselves.[7] To be sure, we do not imagine assembling junk-worlds ourselves, but the repetitive procedure displays some analogy to action recipes. It seems that we could go on and on in the same way in reproduc-

[7] See also Giberman's (2015) imagination of a 'junky spruice'.

ing the structure of the junk-world imagined. This intuition supports our developing an appearance of possibility.

My second example aims at illustrating the relevance of creative solutions in testing metaphysical constraints. Take Williamson's

GOLD: Gold could have an atomic number different from 79.

Metaphysicians in the tradition of Kripke tend to deny GOLD. Yet we should not naïvely assume that atomic number wears its modal status on its sleeve. On pain of circularity, an epistemological account of how we assess GOLD should not start from outright modal knowledge that atomic number is metaphysically necessary. It must start from the role of atomic numbers in our scientific world view: the atomic number of gold plays a key role in explaining the overall chemical behaviour of gold (see Tahko 2015: 813). We take into account the full extent of molecular chemistry. Still textbook chemistry is unlikely to straightforwardly answer the modal question.

One salient way of approaching GOLD is by general considerations which embed the chemistry of gold into a metaphysical framework, which may be assessed by its explanatory virtues, as in recent neo-aristotelean proposals of an empirically informed *essentialism* (Mallozzi 2021 has an overview of the literature). However, if I am right about modal imagination, our appreciation of the metaphysical status of atomic numbers may depend on enriching the general metaphysical framework by considering particular ways for gold to have a different atomic number.

The use of imagination for tackling GOLD can be framed analogously to Tesla's problem of a motor without a commutator. We ask a *how possible?* question: how could nature or god make it true that gold has an atomic number different from 79? We try ways for gold to have a different atomic number, starting from our initial grip on the theoretical bond between gold and atomic number. One salient option is to vary further aspects of our world to see whether they might *compensate* for the differences in theoretical roles of different atomic numbers. Perhaps a stuff with a different atomic number could come sufficiently close to gold to *be* gold if the chemical laws for the constitutive particles like protons, neutrons, electrons, positrons, are slightly twisted in this or that direction.

In performing the task, we may simulate exemplary manipulations. For instance, we may start with considering changing the atomic number of gold to 80. We realize that this yields mercury, which is not gold. We consider ways of solving this problem like a change in the laws for protons and electrons such that 80 protons and electrons exert the same gravitational and electromagnetic forces as 79, go through the corresponding changes for other elements, and so on. The more determinate the scenario becomes, the more specific our awareness of the modal status of atomic number will become. If atomic number is necessary, the changes will prove too substantial to preserve gold as part of our system of elements. But we may not find out unless we try. In any case we get a more precise idea of the essential status of atomic numbers as related to the overall theoretical roles of the particles involved.

My third example is the necessity of origin (the standard example used in Roca-Royes 2011 against conceivability-based modal epistemologies):

ORIGIN:

Aristotle could not have originated from a different zygote than he actually came from.

General metaphysical considerations bear on ORIGIN (see Rohrbaugh and DeRosset 2004), but again they might have to be supplemented by test scenarios, trying to figure out ways for Aristotle to originate from a different zygote. The purported constraint that binds Aristotle to the zygote he actually came from might permit qualification: perhaps Aristotle could have emerged from something that came close enough to the actual zygote to play the metaphysical role of the latter. To check, we might consider one of Jackson's 'fantastic, remote cases', e.g. a scenario in which Aristotle developed not from the actual zygote but from some perfect molecule-per-molecule replica implanted by some advanced extra-terrestrial scientists into the body of his mother at the very moment of his conception.

In sum, the creative use of imagination in thought experimenting seems an often helpful and sometimes even indispensable device for clarifying the modal status of metaphysical constraints.

3.3 Filling the Motivational Gap

I shall now elaborate how the feasibility approach fills the motivational gap. It seems that, in an individual's development, modal issues first arise in issues of feasibility: how can she attain or miss her goals (see Papafragou 1998)? We are immediately disposed to solve feasibility issues like SKUNK and MOUNTAIN by imaginatively simulating a solution. A reliable simulation must recruit any relevant mental resources, propositional thinking, imagery, explicit and tacit knowledge activated by suitable cues.

There is a natural tendency to extend this established practice to more detached possibility issues like TESLA. Responding to a debate of feasibility, Tesla imagined a motor without a commutator without having in mind one particular course of action. There is a continuity even to more detached issues of metaphysical possibility. They do not concern what anyone can do but what could be the case. Imagination works holistically; it may even be confined to propositional content. Still the use of imagination is special compared to principled metaphysical arguments. The original use of imagination in devising a particular solution to a feasibility issue is preserved in the *case*-directedness of modal imagination. The focus is on creatively crafting a concrete recipe for meeting the pertinent constraints on a *p*-situation. The recipe is instrumental in getting a grip on the determinate constraints and their modal status.

The proposal takes on board both the view of imagination as a coherence test and the view of imagination as raising an appearance of possibility. As for the former, just as it is crucial for solving a feasibility issue to come up with a sufficiently concrete solution which meets the relevant restrictions, it is crucial for modally imagining *p* that we can come up with a scenario that (i) verifies p, (ii) brings out the pertinent modal constraints and (iii) reconciles them with *p*. As for the latter, if imagination functions properly, an appearance of possibility arises precisely if the scenario meets these conditions, just as it plausibly arises from a use of imagination for tackling more everyday feasibility issues.

4. Objections and Replies

CIRCULARITY OBJECTION: we need modally qualified knowledge (e.g. knowledge of essences) to constrain imagination (Roca-Royes 2011). A more

recent internalist challenge is that we should give reasons why imagination is suitably constrained (Vaidya and Wallner 2021).

REPLY: I have already used the general circularity worry to outline how imagination is used to manifest *implicit* constraints in the first place, drawing on any available knowledge of the actual world. My resulting feasibility account also lends itself to a reflective justification of why the modal use of imagination is suitably constrained.

ENABLING OBJECTION: The work of imagination is confined to meeting enabling conditions or to a context of discovery. The real justificatory work is done by general arguments.

REPLY: Principled arguments may settle many issues of modality, but I have used my examples TESLA, GOLD, and ORIGIN to argue that they often have to be supplemented by using imagination. Imagination plays a genuine justificatory role in devising concrete solutions for some *p* to be made true.

EXCEPTIONALISM OBJECTION: Modal imagination cannot be integrated into a naturalistic picture which explains epistemic capacities by their life role (see Morato 2019).

REPLY: It is part and parcel to my feasibility approach to bring out a continuity between the use of imagination in tackling everyday issues of feasibility and an eligible way of addressing more remote modal issues. The feasibility approach perfectly fits into a naturalistic epistemology.

UNIQUENESS OBJECTION: Imagination is not our only pathway to modal knowledge, and it does not cover all cases of such knowledge, e.g. the necessity of mathematics.

REPLY: My argument shows how imagination may play a key role in addressing modal issues, but it does not support stronger claims to uniqueness. I shall remain neutral about the format of an integrative modal epistemology. One model for such an epistemology is given by the Kripkean tradition, in particular Chalmers's (2002) notion of ideal conceivability in terms of surveying the space of possible worlds by canonical descriptions. Theoretical considerations and more limited exercises of imagination may play a role in preparing canonical descriptions. Another model would be that the results of using imagination become part of a general metaphysical theory, which does not have to conform to canonical world descriptions but may integrate them.

IMAGISTIC OBJECTION: There is a strong tendency to delimit the epistemic contribution of imagination by its qualitative content, driving a wedge between my pre-philosophical and my philosophical examples (Tidman 1994, Byrne 2007, Fiocco 2007, Kung 2010, Kind 2016, Berto and Schoonen 2018, Jago 2021). In SKUNK and TESLA, qualitative content plays a key role. It is not a matter of course that the same goes for philosophical examples. I outline three motivations for the imagistic view.

The first motivation is the *definitional* issue: how are we to define imagination if not by imagery?

The second motivation lies in confining the genuine *epistemic contribution* of imagination. One obvious answer is that it consists in providing imagery or qualitative content.

The third motivation concerns the *limits and freedom* of imagination. On the one hand, as far as its qualitative content goes, imagination seems very limited. We cannot sensorily imagine things like a ten-dimensional space. Most authors grant that imagination may take on board propositional content, though.[8] Once admitted, propositional content greatly expands the range of imagination. We might *assign* almost any content.: 'I imagine myself receiving the Fields medal for proving Goldbach's conjecture. … I imagine (and I suggest that you have imagined too) that I *really have proved it*. I can also engage in a similar imaginative project: I can imagine disproving Goldbach's conjecture.'(see Kung 2016: 96). In this vein, Priest (2017) claims that we can imagine anything we can grasp. Thus, the propositional content of imagination does not seem properly restrained to provide modal knowledge on its own.

REPLY: I harbour broadly Moorean misgivings about the imagistic objection: an 'established maxim' of using imagination in philosophy is challenged on the basis of a highly debatable hypothesis about how imagination works (see Lam 2018, 2167). When in doubt, we should sacrifice the latter rather than the former, especially given the salient alternative of a holistic view of imagination (see Williamson 2007). But the challenge becomes to tell why that approach yields a notion of *imagination*.

I shall use my feasibility approach to rebut the three motivations of the imagistic objection. The first and the second line of motivation can be tackled together. My feasibility account along broadly simulationist lines provides material for defining imagination and identifying its core epistemic functions. One core function of imagination is to approach issues of feasibility by simulating limited variations of the current situation. The function transmits to more detached issues of possibility. Other uses of imagination like pretense can be connected to this core function (see section 2.3.). The core function supports a holistic view of imagination. The latter may recruit any mental resources required to simulate solutions for feasibility issues.

Coming to the third line of motivation, as illustrated by SKUNK, MOUNTAINEER, and TESLA, imagination recruits any capacities, representational resources, and information available to the mind. It combines them in a more complicated way than presupposed in the objection. A feasibility issue *streamlines* the use of imagination beyond concocting imagery and a free propositional gloss. Streamlining goes beyond explicitly and voluntarily observed constraints. It is largely triggered by thoroughly addressing an issue how p is possible. Determinate versions of implicitly known constraints are not explicitly imposed. They emerge in imaginatively developing a solution for how to make p true. The constraints apply to our entire representation of the scenario. They delimit both qualitative content and assigned content. This explains why the epistemic role of imagination goes far beyond the contribution of qualitative content. Yet again, there are other uses of imagination than the modal one, which come with different requirements and restrictions.

[8] A middle position would be to admit rich quasi-perceptual content (see Byrne 2007).

FREEDOM OBJECTION: As contrasted to perception, imagination is free. We can manipulate its content at will. How can such a manipulation yield independent evidence (see Balcerak-Jackson 2018)?

REPLY: My account shares the deeper motivation of the freedom objection but forges an intimate connection between the epistemic function of imagination and its freedom. We exert the freedom of imagination in creatively coming up with innovative solutions to feasibility issues, but this freedom is also limited by the constraints thereby activated. Imagination in my account resembles a tool. Within limits, we can use a tool in many ways, among them dysfunctional ones. But we can also use it in line with its proper functioning. There are (relatively) free uses of imagination as in WAITER. But if we intentionally use imagination to seriously address a feasibility issue, it is constrained by this purpose. Then it can provide knowledge.

OBJECTION OF FAR-FETCHEDNESS: Does imagination provide a firm grip on remote, fantastic cases like perfect replicas of zygotes and mechanic cats? Relatedly: we cannot simply rely on actuality to fill in the neuralgic details of far-fetched worlds; do we have a suitable grip of them (van Inwagen 1998)?

REPLY: Again the continuity to our normal use of imagination in addressing issues of feasibility provides an answer. Our competence of imagining differentiated action plans as in MOUNTAIN calibrates our imaginative powers. It also comes with *implicit monitoring* when a scenario is sufficiently developed to permit a confident assessment, comparable to our automatic monitoring of perception as to whether it is differentiated enough to support perceptual judgements (see Williamson 2007: 153-155; Gregory 2020). A skilled mountaineer would not be confident about some particular route being feasible if her plan were not suitably developed. The skilled engineer Tesla would not have been satisfied with his vision of the motor if the latter had not been suitably detailed and accurate. In a similar vein, a diligent modal reasoner may be occasionally misled, but she would not generally base her modal verdicts on underdeveloped imagined scenarios, which leave open how to satisfy the pertinent constraints.

APOSTERIORITY OBJECTION: The classical objection to imagination-based accounts is that we can imagine a posteriori impossibilities like water not being H_2O.

REPLY: Imagination within the broad confines of a simulation account can be used in many ways, among them to track epistemic possibilities from viewpoints that differ from ours, e.g. viewpoints from which it is open whether water is H_2O.[9] But modal imagination as modelled on feasibility issues is sensitive to any relevant information, including empirical knowledge. We pay due respect to such information, and we are at a loss how to imagine a suitable way for *p* to be made true in sufficient detail when we lack crucial information, as in Yablo's example of Goldbach's Conjecture (Yablo 1993: 10).

[9] See Yablo 1993: section VIII; Chalmers's (2002) *primary* conceivability.

5. Summary

I have raised and answered a basic motivational issue about the modal use of imagination: what motivates us in using imagination in the first place? My answer is: there is a natural inclination to use imagination in simulating solutions to everyday feasibility issues. There is a continuity between this natural use of imagination and the use of imagination in tackling philosophical possibility issues.

References

Balcerak-Jackson, M. 2018, "Justification by Imagination", in Macpherson, F. and Dorsch, F. (eds.), *Perceptual Imagination and Perceptual Memory*, Oxford: OUP, 209-226.

Berto, F. and Schoonen, T. 2018, "Conceivability and Possibility: Some Dilemmas for Humeans", *Synthese*, 195, 2695–2715.

Bohn, E.D. 2009, "An Argument Against the Necessity of Unrestricted Composition", *Analysis*, 69, 27–31.

Byrne, A. 2007, "Possibility and Imagination", *Philosophical Perspectives*, 21, 125-144.

Chalmers, D. 2002, "Does Conceivability Entail Possibility?", in Gendler, T.S. and Hawthorne, J. (eds.), *Conceivability and Possibility*, Oxford: Clarendon Press, 71–125.

Clarke-Doane, J. 2021, "Metaphysical and Absolute Possibility", *Synthese*, 198, S1861–S1872.

Currie G. and Ravenscroft, I. 2002, *Recreative Minds*, Oxford: OUP.

Evnine, S. 2008, "Modal Epistemology: Our Knowledge of Necessity and Possibility", *Philosophy Compass*, 3, 664–684.

Fine, K. 2005, *Modality and Tense. Philosophical Papers*, Oxford: OUP.

Fiocco, M.O. 2007, "Conceivability, Imagination and Modal Knowledge", *Philosophy and Phenomenological Research*, 74, 364-380.

Fodor J.A. and Pylyshyn Z.W. 1981, "How Direct is Visual Perception? Some Reflections on Gibson's 'Ecological Approach'", *Cognition*, 9, 139–196.

Gibermann, D. 2015, "Junky Non-Worlds", *Erkenntnis*, 80, 437-443.

Gibson, J.J. 1966, *The Senses Considered as Perceptual Systems*, Boston: Houghton-Mifflin.

Godfrey-Smith, P. 2020, "Models, Fictions, and Conditionals", in Godfrey-Smith, P. and Levy, A. (eds.), *The scientific imagination*, Oxford: Oxford University Press, 154-177.

Gregory, D. 2020, "Imagery and Possibility", *Noûs*, 54, 755-773.

Hale, B. 1996, "Absolute Necessities", *Philosophical Perspectives*, 10, 93–117.

Hume, D. 1739–40, *A Treatise of Human Nature*, Selby-Bigge (ed.), Oxford: Clarendon Press, 1896.

Ichikawa, J. and Jarvis, B. 2012, "Rational Imagination and Modal Knowledge", *Noûs*, 46, 127-158.

Jackson, F. 1998, *From Metaphysics to Ethics*, Oxford: OUP.

Jago, M. 2021, "Knowing How Things Might Have Been", Synthese, 198, S1981–S1999.

Kind, A. 2016, "Imagining Under Constraints", in Kind, A. and Kung, P. (eds.), *Knowledge Through Imagination*, Oxford: OUP, 145-159.

Kung, P. 2010, "Imagining as a Guide to Possibility", *Philosophy and Phenomenological Research*, 81, 620-663.

Kung, P. 2016, "You Really Do Imagine It: Against Error Theories of Imagination", *Noûs*, 50, 90-120.

Lam, D. 2018, "Is Imagination too Liberal for Modal Epistemology", *Synthese*, 195, 2155-2174.

Mallozzi, A. 2021, "Putting Modal Metaphysics First", *Synthese*, 198, 1937-1956.

Morato, V. 2019, "Conceivability, Counterfactual Thinking and Philosophical Exceptionality of Modal Knowledge", *Topoi*, 38, 821-833.

Nanay, B. 2011, "Do We See Apples as Edible?", *Pacific Philosophical Quarterly*, 92, 305–322.

O'Neill, J. 1944, *Prodigal Genius. The Miracle Life of Nicola Tesla*, New York: Ives Washburn (reprint 2009).

Nichols, S. and Stich, S. 2000, "A Cognitive Theory of Pretense", *Cognition*, 74, 115-147.

Papafragou, A. 1998, "The Acquisition of Modality: Implications for Theories of Semantic Representation", *Mind and* Language, 13, 370–99.

Priest, G. 2017, "Thinking the Impossible", *Argumenta*, 2, 181-194.

Priest, G. 2021, "Metaphysical Necessity: A Skeptical Perspective", *Synthese*, 198, S1873–S1885.

Pylyshyn, Z.N. 2002, "Mental Imagery: In Search of a Theory", *Behavioral and Brain Sciences*, 25, 157–182.

Roca-Royes, S. 2011, "Conceivability and De Re Modal Knowledge", *Noûs*, 45, 22-49.

Rohrbaugh, G. and DeRosset, L. 2004, "A New Route to the Necessity of Origin", *Mind*, 113, 705-725.

Sidelle, A. 1989, *Necessity, Essence, and Individuation*, Ithaca: Cornell University Press.

Siegel, S. 2014, "Affordances and the Contents of Perception", in Brogaard, B. (ed.), *Does Perception have Content?*, Oxford: OUP, 51-74.

Tahko, T. 2015, "Natural Kind Essentialism Revisited", *Mind*, 124, 795-822.

Tidman, P. 1994, "Conceivability as a Test for Possibility", *American Philosophical* Quarterly, 31, 297–309.

Vaidya, A. and Wallner, M. 2021, "The Epistemology of Modality and the Problem of Modal Epistemic Friction", *Synthese*, 198, S1909-S1935.

van Inwagen, P. 1998, "Modal Epistemology", *Philosophical Studies*, 92, 67-84.

van Leeuwen, N. 2011, "Imagination Is Where the Action Is", *Journal of Philosophy*, 108, 55-77.

Williamson, T. 2007, *The Philosophy of Philosophy*, Oxford: Blackwell.

Yablo, S. 1993, "Is Conceivability a Guide to Possibility?", *Philosophy and Phenomenological Research*, 53, 1-42.

The Pragmatics of Metaphysics Explanation: An Epistemology of Grounding

James Lee

State University of New York at Oswego

Abstract

Explanation can be distinguished between linguistic practices and metaphysical relations. At least with respect to metaphysical explanation, some are skeptical that any knowledge gained via explanation qua linguistic practices confers knowledge of explanation qua metaphysical relation. I argue that this skepticism is unfounded. Engaging in the linguistic practice of explanation gives us no reason to skeptical in beliefs about corresponding metaphysical relations like causation or grounding. Moreover, those very linguistic practices can provide resources to justify beliefs in those relations. So, exploring those practices can move us forward in developing an epistemology of grounding and metaphysical explanation.

Keywords: Grounding, Metaphysical explanation, Essence, Contrastivity.

1. Introduction

There is a voluminous and growing literature on grounding. However, the literature on the epistemology of grounding is relatively sparse. There are numerous contributions to the literature addressing questions about the nature of grounding. Is grounding irreflexive? Asymmetric? Transitive? Is grounding well-founded? Is grounding properly expressed as a relation or as an operator? Does grounding relate facts or entities of any ontological category? There aren't as many contributions addressing questions about how it is that we come to know claims about grounding. This paper is a contribution to the latter project. I will argue that our explanatory practices can confer justification for beliefs about grounding claims, i.e., claims about metaphysical explanation.

Some are skeptical about this link. Some argue that the nature of explanatory practices, is such that they cannot confer justification for beliefs about worldly relations that purportedly underwrite or serve as the truthmaking basis for these practices. According to this line of thought, our explanatory practices are subjective in some pernicious sense. Worldly relations are objective. Beliefs

Argumenta 10, 1 (2024): 145—159
ISSN 2465-2334

DOI 10.14275/2465-2334/20240.lee
First published: dd November 2024

justified via subjective means cannot justify beliefs about objective claims. Therefore, explanatory practices cannot justify beliefs about worldly relations like grounding. I will argue that this argument is unsound. Not only is it false that explanatory practices fail to justify beliefs about worldly relations, but it is also the case that those practices themselves can provide resources to justify beliefs about worldly relations like grounding.

In section 2, I will provide a brief survey of metaphysical explanation. In section 3 I will exposit the skeptical argument against the epistemic connection between grounding and explanatory practices. Sections 4 and 5 will show that the argument is unsound. Section 6 will show how our explanatory practices can confer justification for beliefs about grounding or metaphysical explanation.[1]

2. Metaphysical Explanation

Explanations are generally what people give in response to why-questions. It is typical for an explanation to follow an indicator term like "because". Explanations come in different varieties. A common form of explanation is a causal explanation. We often give causal explanations in response to questions about why something happened. Someone asks a doctor, "Why did I get sick?", and the doctor might give an explanation that identifies particular causes, like a bacterial infection. Another common form of explanation is an appeal to reasons. We appeal to such reasons in response to questions about an individual's actions. Someone might ask why I went to the kitchen and in response, I might say that I wanted to get something to eat and that I believed that there were leftovers in the refrigerator.

Alongside these kinds of explanations, there is another possible category of explanation. These explanations are what metaphysicians call "metaphysical explanations". Rather than asking why something happened, or why someone acted in some way, we might ask why something is the way that it is. Such questions might be answered by appealing to causes. When we ask why diamonds are hard, we might answer this by providing the causal antecedents that led to the formation of diamonds. However, there is a different way of answering this question. We might explain why diamonds are hard by identifying their underlying composition and structure. Diamonds are hard because they are composed of carbon atoms arranged in a crystalline pattern. Note that this way of explaining doesn't tell us why or how something came to be. Rather this explanation tells us why something is the way that it is by appealing to some other, usually more fundamental, fact about that very same thing. Philosophers often use the phrase "in virtue of" to signal such an explanation. Here are some common examples of this kind of explanation.

(1) The statue has a particular weight in virtue of the weight of its constituting matter.

[1] It is worth asking whose explanatory practices confer justification. Would it be the explanatory practices of ordinary individuals? I hold that the explanatory that confer justification are those conducted by metaphysicians in the ontology room. Since why-questions and answers to why-questions require interpretation, identifying the explanatory practices of metaphysicians as justification-conferring will at least provide some clarity that may be missing in ordinary language practices. Thanks to an anonymous reviewer for raising this point.

(2) Moral claims are true in virtue of natural facts.
(3) Mental states occur in virtue of the occurrence of corresponding brain states.

These sorts of explanations, i.e. explanations that involve some kind of non-causal determinative connection between explanans and explanandum, are what falls into the category of metaphysical explanation.

I will assume without argument that for at least some kinds of explanations correspond with some kinds of metaphysical relations.[2] Examples of such explanations are causal explanations. Hume and Humeans aside, it is commonly supposed among contemporary analytic metaphysicians that the truth of causal claims corresponds to some mind-independent feature of the world. To say truthfully that *x* causes *y* is to say that some particular worldly relation, perhaps counterfactual dependence or minimal sufficiency, holds between *x* and *y*. In like manner many, but not all, metaphysicians hold that the truth of metaphysically explanatory claims corresponds to some mind-independent feature of the world.[3]

The metaphysical relation that is most commonly associated with metaphysical explanation is the widely discussed notion of grounding. There are few if any, uncontroversial claims to be made about the nature of grounding.[4] That said, grounding is generally thought to have the formal features of irreflexivity, asymmetry, and transitivity. Grounding is also generally thought to be necessitating. If *x* grounds *y*, then necessarily if *x*, then *y*. For this paper, one primary question to address is what the relationship is between grounding and metaphysical explanation.

What we observe from the proceeding is that the term "explanation" is ambiguous. It can refer to the sorts of things that individuals communicate to each other. Alternatively, it can refer to mind-independent relations out in the world. We can say that Jones gave an explanation to Smith. We can also say that a bacterial infection explains why someone is sick, apart from anyone stating that there is a bacterial infection. Henceforth I will call the former *explanatory practices*. I will call the latter *worldly relations*. "Metaphysical explanation" can either refer to an explanatory practice or a worldly relation, such as grounding. There is nothing in this paper that hangs on whether the term "metaphysical explanation" should be reserved exclusively for the explanatory practice or the worldly relation.

3. The Skeptical Problem of Metaphysical Explanation

At least concerning metaphysical explanation, there have been concerns expressed about the epistemic relation between our explanatory practices and corresponding metaphysical relations. Suppose that Wong asks a why-question and Garcia answers Wong's question. Suppose that Wong forms a justified belief as a result of Garcia's answer. Does this justified belief also confer justification for believing that some corresponding worldly relation holds? Some philosophers have argued that at least with respect to metaphysical explanation, we should adopt a skeptical stance about that question. In her (2016) Naomi Thompson states:

[2] See Audi 2015 and Roski 2021 for defenses of the realist thesis.
[3] See Miller and Norton 2017 for a dissenting opinion.
[4] See Bliss and Trogdon 2021 for a survey.

> If metaphysical explanation is like ordinary explanation but in a metaphysical context, then (assuming we can meet the challenge of specifying what this context is) the problem is that metaphysical explanation, like ordinary explanation, will have pragmatic features. What makes for successful metaphysical explanation will depend (to an extent) on features of agents… But that straightforwardly contradicts [the thesis that] grounding relations are supposed to be entirely objective and mind-independent (397-398).

Anna-Sofia Maurin, in her (2018) argues as follows:

> More precisely, if grounding is a mind-independently obtaining worldly relation, adopting separatism amounts to saying of explanation that it is not a mind-independently obtaining and worldly relation. Rather, explanation is mind-involving, pragmatic, and/or 'epistemic' (whatever we take those locutions to mean more precisely). But then, as part of what it is to be an explanation is to be this mind-dependent and epistemic thing, why think that explanation having the properties it does, justifies our thinking that those are properties had by worldly and mind-independent grounding? No good reason comes to mind (1578-1579).

The above passages suggest this line of reasoning, which I will call the *main skeptical argument*:

(1) Worldly relations are mind-independent.
(2) Explanatory practices are mind-dependent.
(3) If explanatory practices are mind-dependent and worldly relations are mind-independent, then justified beliefs brought about by explanatory practices do not confer justified beliefs in corresponding worldly relations.
(4) Therefore, justified beliefs brought about by explanatory practices do not confer justified beliefs in corresponding worldly relations.

I will assume that premise (1) is true by definition. I will discuss premise (2) in the next section. Why think that premise (3) is true? Suppose that it is the case that explanatory practices are mind-dependent. Explanatory practices include both the formation of the why-question and the answering of the why-question. Thus, to say that explanatory practices are mind-dependent is to say that either what counts as a why-question is mind-dependent or that what counts as an acceptable answer to a why-question is mind-dependent. Suppose that what counts as a why-question and an acceptable answer to the why-question are mind-dependent. Finally, suppose that *x* is justified in believing an explanation *e* just in case *e* is an acceptable answer to a why-question. What seems to follow from this is whether one is justified in believing *e* will be subject at least partially to mind-dependent factors. Such factors might include aesthetic preferences, practical considerations, or even wishful thinking.

Given what I said above about justification being a function of mind-dependent factors, the argument for premise (3) goes as follows. Mind-dependent factors like aesthetic or practical preference do not reliably track the truth of claims about worldly relations. If the factors by which one forms a belief that *p* are unreliable with respect to claims about q, then one is not justified in believing *q* on the basis of *p*. In other words, such unreliability undercuts one's justification for believing *q*.[5] For instance, suppose some epistemic subject S looks outside and

[5] For more on undercutting and rebutting defeaters, see Pollock 1986.

sees that it is raining. S forms the belief that it is raining outside. S is justified in believing that it is raining outside on the basis of S's perception that it is raining. However, S has some reason to think that her perception is unreliable. Perhaps S took a drug earlier that produces hallucinogenic effects. Given that S has reasons to think that her perception unreliably tracks entities external to my mind, her justification for believing that it is raining outside has been defeated.

Since the factors that bring about one's belief in *e* via explanatory practice unreliably track the truths regarding corresponding worldly relations, while one may be justified in believing that *e* is an acceptable answer to a why-question, one is not justified in beliefs about some corresponding worldly relations on the basis of *e*. This reasoning is then applied to metaphysical explanation. The means by which we judge an answer to a metaphysical why-question to be satisfactory unreliably tracks corresponding metaphysical relations. As such, justified beliefs that arise from metaphysical explanatory practices do not confer justification for beliefs in corresponding metaphysical relations.

It's worth noting that beliefs about worldly relations fall into at least two categories. Such beliefs can be about the nature of such relations. Secondly, such beliefs can be about whether such a relation holds between certain relata. A strong form of skepticism would hold that justified beliefs formed via explanatory practices do not confer justification for either kind of belief about worldly relations. A weaker form of skepticism would allow for the possibility of justification for one of the two kinds of beliefs about worldly relations on the basis of justification via explanatory practice. This essay aims to show that justified beliefs via explanatory practice can confer justification for both kinds of beliefs about worldly relations.

4. The Pragmatics of Explanation: Why-Questions as Mind-Dependent

Premise (2) of the main skeptical argument says that explanatory practices are mind-dependent. Why think that this is true? One can derive support for this claim by appealing to work done on the pragmatics of explanation in the philosophy of science. A particularly influential account is given by Bas van Fraassen in his classic (1980). According to van Fraassen, explanations are answers to why-questions. Why-questions themselves are sensitive to context along three dimensions.

First, why-questions have a topic. Why questions have the form "Why *p*?" where *p* is some proposition. *p* is the topic of the question. The topic of a why-question is sensitive to context in all the usual ways. Suppose someone asks, "Why did the students get sick?" The topic of this question is the proposition, <The students got sick>. Answering this question in any satisfactory way will require that we specify contextual parameters like time, location, the specific individuals designated by "the students", etc. Moreover, the topic of a why-question is considered a *presupposition* (see Bromberger 1966). The topic of the why question must be assumed to be true in some sense in order for the question itself to be felicitous. Asking a question like "Why is Los Angeles the capital of the United States?" would be considered infelicitous.

Second, why-questions have a contrast class. A contrast class is a set of alternatives that specifies the appropriate answer to a why-question. By specifying a contrast class, an answer to a why question must not only explain why the topic

of a question is true but also explain why the members of the contrast class are false. Consider the following example:

Why did Suzy hit Jimmy with a pie?

The topic of this question, i.e. that Suzy hit Jimmy with a pie, can be associated with the following three contrast classes:

Why did Suzy, rather than (Jane, Angela, Eloise, etc.) hit Jimmy with a pie?
Why did Suzy hit Jimmy rather than (Bob, Steven, Marcus, etc.) with a pie?
Why did Suzy hit Jimmy with a pie rather than a (cake, doughnut, sundae, etc.)?

Consequently, there can be at least three different kinds of appropriate answers to this why question, depending on which contrast class we specify. Which contrast class we specify will be sensitive to context.

Third, there are considerations with respect to explanatory relevance when attempting to answer a why-question. Even after specifying the topic and contrast class of a why question, such a question can still admit of multiple correct answers. As an example, van Fraassen asks, "Why does blood circulate through the body?" A relevant answer for someone wanting to know what makes the blood circulate would be "because the heart pumps the blood through the arteries". A relevant answer for someone wanting to know the function of blood circulation would be "to bring oxygen to every part of the body tissue". According to van Fraassen, a proposition that is an explanatorily relevant answer to a why-question will bear a relevance relation to the ordered pair $<P_K, X>$, where P_K is the topic and X is the contrast class. There are a number of different ways in which some answer to a why-question can bear a relevance relation to $<P_K, X>$. Which relation is the right one will be sensitive to context. What relevance relations there are for any given $<P_K, X>$ will be important for what follows.

How does van Fraassen's account provide support for premise 2 of the main skeptical argument? Premise (2) states that explanatory practices are mind-dependent. In this case, the explanatory practices are the asking of why-questions, which includes contextual specification along the three parameters discussed above. Such practices are mind-dependent primarily because they are *interest-relative*. The topic of a why-question, i.e. what the question is about, is determined by the interests of the asker. It seems obvious enough that when someone asks a why-question, what they ask about will be determined by what they are interested in learning. If a person isn't interested in *p*, then we wouldn't expect them to ask why-*p*. The contrast class of a why-question is also determined by interest in several ways. First, the topic of a why-question can admit of more than one contrast class, as is the case with the pie sentences above. Which contrast class to focus on will depend on what the asker wants to know. Second, what to include in a contrast class can be a function of interest-relativity. Suppose that Suzy hit Jimmy with a pie. Suppose further that it is counterfactually true that four other individuals could have hit Jimmy with a pie. Rather than including all four other individuals in the contrast class, the asker might only be interested in contrasting with just one of the individuals. Thus, as a result of the asker's interest, the size of the contrast class may vary. Finally, explanatory relevance is also a function of interest-relativity. There are a number of different ways in which an explanation can enter into a relevance relation with $<P_K, X>$, and many of these ways are a function of interest. For instance, there can be a number of correct answers to a why-question that differ with respect to complexity. Which of these answers is

relevantly related to the why-question will depend on the interests of the individual asking the question. Someone with a layperson's understanding of epidemiology will not be interested in a highly technical answer to the question of why diseases spread.

The upshot of the above is unsurprising. Why-questions are mind-dependent in that what they are about and whether they are event asked at all is up to us. If there is no sapient life in the universe, then there are no why-questions being asked. In this sense, explanatory practices are ontologically dependent on minds in that it is essentially an activity conducted by individuals with an interest in seeking certain kinds of knowledge. However, the reader may have noticed a discrepancy between the argument given in section 3 and what was presented here. In this section, a defense of premise 2 was given by showing that the *asking* of why-questions is at least partly a function of interest, and thus mind-dependent. In the previous section, a defense of premise 3 was given by showing that if the *answering* why-questions was a function of mind-dependent factors, then no justification is conferred for beliefs about worldly relations. The main argument thus stands guilty of committing equivocation. Explanatory practices can be mind-dependent in that the asking of why-questions is a function of mind-dependent factors. Explanatory practices can be mind-dependent in that the answering of why-questions is a function of mind-dependent. So, in order for the main skeptical argument to be sound, it must be shown either that the mind-dependence of asking why-questions entails the mind-dependence of answering why-questions, or that the answering of why-questions is mind-dependent for independent reasons. In the next section, I will argue that neither is the case. The mind-dependence of answering why-questions does not necessarily entail the mind-dependence of answering why-questions. Moreover, it is not the case that the answering of why-questions is necessarily mind-dependent. In arguing for both I will thus show that the main skeptical argument is unsound.

5. The Pragmatics of Explanation: Why-Questions as Mind-Independent

Suppose it is the case that the asking of why-questions is a function of mind-dependent factors such as interest-relativity. Does it follow from this that the answering of why-questions is also a function of mind-dependent factors? The answer is no. The fact that we ask why-questions about what we are interested in does not imply that what we consider to be an acceptable answer is determined by what we consider to be interesting, or by any other mind-dependent factor. In fact, we have some reason to think that the asking of why-questions itself is guided by mind-independent factors beyond interest relativity, such as factors related to identifying truth.

We first begin with the topic of the question. Specifying the topic of a why question involves specifying the context. There are elements to context specification that are objective. For instance, specifying the referent of an indexical term like "I" or "here" is plausibly objective in nature. If David Lewis utters "I am a philosopher", then the referent of "I" in this context is David Lewis. Who "I" refers to is not assessment-sensitive. In other words, reference to such indexicals does not change relative to who happens to be the listener. Once the context establishes that "I" refers to David Lewis, the sentence "I am a philosopher" is true regardless of who happens to be assessing the sentence. Topic

specification seems to generally involve this kind of reference fixing—going from character to content, using David Kaplan's terminology. For instance, consider the question, "Why did the robbery occur?" Fixing the context involves specifying parameters like world, location, and time. If we specify the context such that it results in a true proposition, then we've established the topic for the question. For instance, if we identify the parameters as 2:45 pm on August 4 2021 at The Bank of Princeton in Princeton, New Jersey in the actual world, and if it is indeed true that a robbery occurred at that time, location, and world, then it is the cause that we've specified the topic for the why question. This process is not sensitive to interest-relativity, and so we have reason to believe that this aspect of specifying the why question does not entail that answering a why-question is subject to mind-dependent interest-relativity.

The next aspect is the contrast class. Specifying a contrast class involves engaging in counterfactual reasoning. Selecting members of a contrast class involves identifying relevant alternatives. There are limits to which alternatives are plausibly members of a contrast class, and those limits are for the most part not sensitive to interest relativity. Consider the following example, "Why did LeBron pass the ball to Anthony?" Suppose the topic has it that this question is about a particular action that occurred during an NBA game. Suppose further that we are to form a contrast class for values of x in the following: "Why did LeBron pass the ball to Anthony rather than *x*?" We reason counterfactually in order to determine the appropriate members of this class. Doing so involves substituting names for *y* in the following and evaluating whether the resulting proposition is true: "LeBron could have passed the ball to *y* rather than Anthony". If the sentence is true, then we have a suitable candidate member of the contrast class. Given the features of the context, and given the usual factors that go into determining the closest possible worlds, this sort of counterfactual reasoning places constraints on admissible members of the contrast class. Other teammates on the basketball court at the time of the action would be admissible members of the contrast class. Someone living halfway across the world would not be an admissible member of the contrast class. This goes some way in showing that membership in a contrast class isn't a matter of interest-relativity. We generally don't add things to a contrast class on the basis of pragmatic or practical reasons. So, insofar as specifying a contrast class plays a role in justifying explanation, interest-relativity is not part of the justification process. Thus far any interest-relativity found in formulating the why-question doesn't entail interest-relativity in answering the question.

Of the three why-specifying components, it may seem that relevance relations are the most conducive to interest-relativity. A why-question with a specified topic and contrast class can still have multiple correct answers. Would this not be a case of interest-relativity that would defeat justification for believing in some corresponding objective relation? This needn't be the case. To say that a why-question can have multiple correct doesn't necessarily some anti-realism about the answers. Rather, it can be the case that there are multiple objective relations at work when it comes to answers to a particular why-question. Consider the following question, "Why are diamonds hard?" There are at least two correct answers to this question, and they correspond to different worldly relations. One response is to identify the conditions under which diamonds are formed. Another response identifies the underlying matter and structure of a diamond. Which answer we want is a function of our interests, but the answer still corresponds to

some objective feature of reality. Furthermore, given that explanatory relations are transitive, there can be multiple correct answers involving the same relation. One answer to why question can identify the immediate cause of the question topic. Another answer can identify a cause that is further upstream. The same can be said for grounding relations. There can be answers that identify the immediate grounds or answers that identify the ultimate grounds for the topic of the why question. Again, while it may be the case that which part of the causal or grounding chain we focus on is a matter of interest, this does not imply that whether or not the answer is correct is a matter of interest.

Interest isn't the only thing that factors into relevance. Another might be our ability to understand. Suppose someone asks why people get cancer. What answer is relevant for this individual will depend on their level of understanding with respect to biology and human physiology. A highly technical answer will not be relevant for someone with no background in either. Does this way of measuring relevance imply that answering why-questions is a function of mind-dependent factors? Not necessarily. When we provide different answers to a why question for different levels of cognitive ability, we are not thereby changing the subject and talking about different things. It is plausible to think that different answers are still talking about the same worldly relation under different descriptions. It is often the case that an answer to a why question given to an individual with little to no background will appeal to figurative language. Even at this level we are often still talking about some worldly entity, under the plausible assumption that the figurative language can be translated into a correct literal answer. Consequently, we have some good reasons to think that relevance does not imply that providing an adequate answer to why-questions is subject to mind-dependent factors.

The foregoing considerations hopefully suffice in showing that any mind-dependence in formulating why-questions does not entail any mind-dependence in answering why-questions. We move on to the question of whether the process of answering why-questions itself is a function of mind-dependent factors. If that is the case, then there would be good reason to be skeptical that such answers correspond to objective relations. However, this needn't be the case. It is certainly true that people can deem an answer to a why question as good because it suits their interest, but it's far less certain that every good answer to a why question is based even partly on interest. In many cases what makes an explanation good is that it identifies an objective mind-independent relation.

Methods for identifying such a relation can be derived from the pragmatics of why-questions themselves. For instance, consider Peter Lipton's discussion on contrastive inference (Lipton 1991). Lipton demonstrates that the very act of producing a contrast class provides a way to infer causal relations. Contrastive inference is a variant of Mill's methods of agreement and difference. Mill's methods are one way in which one infers that there is a causal relation. According to the method of difference, we infer that *C* is the cause of *E* when we observe that in a variety of cases where *C* is absent, *E* is also absent. According to the method of agreement, we infer that *C* is the cause of *E* when we observe that *C* is present, *E* is also present throughout multiple cases where the only relevant commonality is both *C* and *E*. Contrastive inference works backward. We observe *E* in scenario 1, but not in other relevant scenarios. Those scenarios form a contrast class. We then apply the method of difference and look for what scenario 1 has that the others lack. Recall that forming a contrast class involves counterfactual reasoning.

Contrastive inference and the methods of difference and agreement are also applications of counterfactual reasoning. Counterfactual reasoning does not necessarily involve any kind of subjective interest. In these cases, what makes an explanation good is not that it serves our interests. Rather what makes an explanation good is that it successfully locates a causal relation.

Such methods can do at least two things for us. First, methods like contrastive inference can justify our belief that some worldly relations hold. For instance, we use contrast classes to justify our belief that some event *x* causes some other event *y*. Second examining such methods can justify our beliefs about the nature of worldly relations. For instance, we can examine how we form contrast classes in order to identify certain features of the worldly relation we take to be doing the explanatory work. When we use contrast classes to identify causes, we note that we do not include potential causes that are outside of the effect's light cone. When we ask why John rather than Suzy hit Joe with a pie, we don't include someone who lives halfway around the world in the contrast class. Moreover, we don't include events in our contrast class that occur after the effect. When we ask why the assassination of Archduke Ferdinand led to World War I, we don't include the assassination of John F. Kennedy in our contrast class.[6] So, not only is it the case that methods such as contrastivity confer justified beliefs about whether corresponding worldly relations hold but investigating how we employ these methods can confer justified beliefs about the nature of corresponding worldly relations.

The upshot of the foregoing is that while interest plays a role in specifying a why question, it does not threaten justification for believing that explanation corresponds to some worldly relation. What kind of why-question we want answered is surely at least a partial function of interest. However, what answer we think is correct is not necessarily a matter of interest. Oftentimes it is not a matter of interest at all. In fact, it is often in our best interest that the answer to a why-question identifies the correct worldly relation. For instance, we recognize that there are some cases where identifying an answer to a why-question as correct on the basis of interest are instances of motivated reasoning. It seems plausible to hold that motivated reasoning is in tension with epistemic rationality. To infer that we are not justified in believing in some corresponding worldly on the basis of a good explanation confuses our interests in asking the question with the methods we use in answering that question. Furthermore, we have reasons to think that our explanatory practices, i.e. how we ask and answer why-questions, can provide us with various justified beliefs about corresponding worldly relations.

It's also worth noting that van Fraassen himself doesn't take the answering of why-questions to be necessarily interest-relative. With respect to answering why-questions, van Fraassen has the following to say.

> How good is the answer *Because A*? There are at least three ways in which this answer is evaluated. The first concerns the evaluation of *A* itself, as acceptable or as likely to be true. The second concerns the extent to which *A favours* the topic *B*

[6] The possibility of causal loops complicates matters when it comes to the temporal ordering of causal relations. However, the general point still stands. Our practices in forming contrast classes can still confer prima facie justified beliefs about the nature of the causal relation.

> as against the other members of the contrast-class… The third concerns the comparison of *Because A* with other possible answers to the same question; and this has three aspects. The first is whether *A* is more probable (in view of *K*); the second whether it favours the topic to a greater extent; and the third, whether it is made wholly or partially irrelevant by other answers that could be given (1980: 146).

What is worth pointing out here is that for van Fraassen, a good answer to a why-question is one that directs us to the truth. Given that van Fraassen is developing a theory of scientific explanation, this should come as no surprise.

Peter Achinstein provides the following analysis of the illocutionary act of giving explanations:

> *S* explains *q* by uttering *u* iff *S* utters *u* with the intention that their utterance render *q* understandable by producing the knowledge, of the proposition expressed by *u*, that it is a correct answer to *q* (Achinstein 1983: 18).

The important part of this analysis is that the explanation must be a *correct* answer to a question like "Why q?" Correctness is not understood in terms of aesthetic preferences or practical interests. Correctness is understood in terms of truth (Achinstein 1983: 42).

Finally David Lewis notes that providing an explanation, in particular a causal explanation, amounts to providing a causal history. Given that we are finite minds, no human has the ability to provide a complete causal history as an answer to a why-question. Consequently, we use pragmatic tools like the ones van Fraassen developed in order to provide a relevant partial history that answers a particular why-question.

> Why-questions, of course, are among the questions that inevitably get partial answers. When partial answers are the order of the day, questioners have their ways of indicating how much information they want, or what sort… One way to indicate what sort of explanatory information is wanted is through the use of contrastive why-questions (Lewis 1986: 229).

The above references should provide additional reasons in favor of rejecting the claim that answering why-questions is a mind-dependent affair. As I mentioned above, this should come as no surprise, given that much of the developmental work on explanations occur in the philosophy of science. A theory of scientific explanation that entails that all answers to why questions are mind-dependent would have disastrous consequences for those who take science to provide us with objective knowledge about the world.

Before proceeding to the next section, I want to make clear that it is certainly true that *some* answers to why-questions are entirely a function of mind-dependent factors. Surely, we look for answers to some why questions that satisfy our aesthetic preferences or our practical interests. Sometimes we look for answers on the basis of wishful thinking or to confirm our own biases. What is important to note here is that while some cases of answering why-questions are mind-dependently determined, some are not. We see from the literature on the methodology of science that we recognize a large class of cases where the process by which a why-question is answered is not governed by mind-dependent factors. So, we see that at least with respect to cases involving causal or scientific explanation, it is not the case that mind-dependent factors in why-question

formulation imply mind-dependence in answering why-questions. Nor should we think that there is anything perniciously mind-dependent about the process of answering why-questions in themselves. Thus, we have reasons to reject the main skeptical argument.

6. The Epistemology of Metaphysical Explanation

The last step is to take what was just said about answering why-questions involving causes and to apply the same methods to metaphysical explanation. In doing so, we can develop an epistemology of metaphysical explanation. What we observed above is that there are tools that we employ in formulating and answering why questions that can also confer justification for believing that some worldly relation. One such tool that I will focus on here is contrastivity. We identify contrast classes in order to specify the why-question we are asking. As Lipton pointed out, we can employ such contrast classes in isolating causes. As I mentioned above, this use of contrast class can both justify beliefs about whether causal relations hold and beliefs about the nature of causal relations. Using contrast classes can confer similar justification for beliefs about metaphysical relations such as grounding. I will discuss one type of case where explanatory practices like contrastivity can justify beliefs.

A common type of question in metaphysics is the "What is *F*?" question. What is time? What are properties? What is possibility? Such questions can be plausibly interpreted as inquiries into essences. To ask, "What is *F*?" is to ask about *F*'s essence. Philosophers have argued that there is a close relationship between grounding and essence.[7] The fact that *x* is *F* is grounded in the fact that *x* is *G*, and *G* constitutes at least a part of *x*'s essence. For example, the fact that Saul is in pain is grounded in the fact that Saul's brain is undergoing a kind of c-fiber activation, and this c-fiber activation constitutes the essence of being in pain. This appeal to essence is explanatory. The reason why Saul is in pain is that Saul's brain is undergoing c-fiber activation.

We can employ contrast classes when answering such questions. When we ask, "Why is *x* an *F*?", we can form at least two contrast classes. We can ask, "Why is *x*, rather than *y*, *F*?" Alternatively, we can ask, "Why is *x* an *F* rather than a *G*?" Forming such contrast classes allows us to engage in contrastive inferences. Such inferences allow us to locate some essence or partial essence in virtue of which *x* is *F*. Furthermore, contrastive inferences enable us to form justified beliefs about both the nature of essences and grounding relations. We can illustrate the use of contrast classes in answering why-questions about natural kinds.

Suppose we ask the question, "Why are whales mammals?" In this question, we are asking what is it that grounds the fact that a whale is a mammal. In other words, we are asking for the essence of mammal-hood, i.e. what it is to be a mammal, such that a whale counts as a mammal. In order to answer this question, and thus identify what it is that grounds the fact that a whale is a mammal, we can create a contrast class to locate the mammal essence. We can why whales rather than squids count as mammals. Creating such a contrast class both specifies the question and enables us to perform contrastive inferences. As mentioned

[7] For the connection between grounding and essence, see Fine 2012, Rosen 2010, and Kment 2018.

previously, making contrastive inferences involves using a variation of Mill's methods of difference and agreement. We note that whales are mammals, but squids are not mammals. This contrast class leads us to search for some F that whales possess and that squids lack that may serve as the full or partial essence of mammal-hood. Just as adding more to the contrast helped Semmelweis to be more precise in identifying the cause of childbed fever, we can add more to a contrast class so as to identify an essence with more precision. Contrasting whales with squids will provide us with some information. Whales have backbones, squids don't. This is insufficient for identifying the essence of mammals. Adding something like sharks to the contrast class helps us get closer to identifying the essence of mammals. Moreover, we can employ different contrast classes to further triangulate the sought-after essence. In addition to asking why whales, rather than sharks or squids, are mammals. We can ask why whales are mammals, rather than reptiles or amphibians. Engaging in this kind of contrastivity with respect to question-asking and answering is a method that can confer at least prima facie justified belief in the claim that *x* being *F** explains why *x* is *F*. This in turn can justify our belief that *x* being *F** serves as the grounds for *x* being *F*.

In addition to justifying beliefs about whether some metaphysical relation holds, explanatory practices can justify beliefs about the nature of metaphysically explanatory relations. For instance, there is currently a debate in the grounding literature about whether there is a unified metaphysically explanatory relation. Call this unified relation "big-G" grounding. Some argue that there is no theoretical utility in positing a big-G grounding relation. Rather, a plurality of "small-g" grounding relations is sufficient for a metaphysician's theoretical purposes.[8] Such relations might include constitution, composition, determinate/determinable, etc.

Examining our explanatory practices may help to move this discussion forward. Recall that a satisfactory answer to a why-question must bear a relevance relation to the topic and contrast class of the question. As I argued above, whether an answer bears a relevance relation to the why-question is not solely a matter of subjective factors. Answers that correspond to appropriate worldly relations like causation also bear a relevance relation to why-questions. We can reframe the debate about the theoretical unity of grounding in terms of relevance relations. Is it the case that there are answers that correspond to a big-G grounding relation that bear relevance relations to why-questions? Or, is it the case that there are no such answers, and that an answer that would bear such relevance relations corresponds to one of a plurality of small-g grounding relations.

Framing the debate about the theoretical unity of grounding around relevance relations allows us to focus on our explanatory practices to see if there is evidence for thinking that there are big-G grounding relation answers that are relevantly related to certain why-questions. For example, Ted Sider observes that there are cases involving general theses about positions like naturalism and physicalism that are best expressed using big-G grounding (see Sider 2020). In other words, there are why-questions about global metaphysical views like naturalism or physicalism that are most relevantly answered by appeals to big-G grounding. Sider's observation is a hypothesis that we can investigate by examining our explanatory practices with respect to the relevant class of why-

[8] See Wilson 2014 and 2021 for the influential criticism of big G grounding. See Schaffer 2016 and Berker 2018 for responses.

questions. Why are true scientific claims true? Why do we have any phenomenal experiences at all? Why is there something rather than nothing? In answering these questions, is it the case that an answer corresponding to big-G grounding is relevantly related, or is it the case that instead some answer corresponding to a small-g grounding relation is relevantly related? If we find that for each question, there is a small-g relation that is relevantly related, then that would be a reason to reject the theoretical unity of grounding. In sum, explanatory practices like contrast classes and relevance relations can be the means by which we arrive at justified beliefs regarding metaphysical relations like grounding.

7. Conclusion

What I've shown in this paper is that realist analytic metaphysicians need not fear epistemic explanations or explanatory practices in general. Rather than being solely governed by subjective or otherwise mind-dependent factors, such practices can offer us a rich vein of insight into how it is that we justify our beliefs about worldly explanatory relations. By incorporating seminal work by philosophers of science on explanation, the paper gestures towards an opportunity for the discussions on grounding to be enriched by the literature on scientific explanation. Further engagement with the literature on scientific explanation will surely advance the discussion on both the metaphysics and epistemology of grounding.

References

Achinstein, P. 1983, *The Nature of Explanation*, Oxford: Oxford University Press.

Audi, P. 2015, "Explanation and Explication" in Daly, C. (ed.), *The Palgrave Handbook of Philosophical Methods*, London: Palgrave-Macmillan, 208-230.

Bennett, K. 2017, *Making Things Up*, Oxford: Oxford University Press.

Berker, S. 2018, "The Unity of Grounding", *Mind*, 127, 507, 729-777.

Bromberger, S. 1966, "Why-Questions" in Colodny, R. (ed.), *Mind and Cosmos: Essays in Contemporary Science and Philosophy*, Pittsburgh: University of Pittsburgh Press, 86-111.

Bliss, R. and Trogdon., K. 2021, "Metaphysical Grounding", in Zalta, E. (ed.), *The Stanford Encyclopedia of Philosophy*, Winter 2021 Edition.

Fine, K. 2012, "Guide to Ground" in Correia, F. and Schneider, B. (eds.), *Metaphysical Grounding: Understanding the Structure of Reality*, Cambridge: Cambridge University Press, 37-80.

Kment, B. 2018, "Essence and Modal Knowledge", *Synthese*, 198, Suppl 8, 1957-1979.

Lewis, D. 1986, "Causal Explanation", in *Philosophical Papers (Vol II)*, Oxford: Oxford University Press, 214-240.

Lipton, P. 1991, *Inference to the Best Explanation*, London and New York: Routledge.

Maurin, A. 2019, "Grounding and Metaphysical Explanation: It's Complicated", *Philosophical Studies*, 176, 6, 1573-1594.

Miller, K. and Norton, J. 2017, "Grounding: It's (Probably) All in The Head", *Philosophical Studies*, 174, 12, 3059-3081.

Pollock, J. 1986, *Contemporary Theories of Knowledge*, Savage: Rowman and Littlefield.

Raven, M. 2015, "Ground", *Philosophy Compass*, 10, 5, 322-333.

Rosen, G. 2010, "Metaphysical Dependence: Grounding and Reduction", in Hale, B. and Hoffman, A. (eds.), *Modality: Metaphysics, Logic, and Epistemology*, Oxford: Oxford University Press, 109-135.

Roski, S. 2021, "In Defence of Explanatory Realism", *Synthese*, 199, 5-6, 14121-14141.

Schaffer, J. 2016, "Ground Rules: Lessons from Wilson", in Aizawa, K. and Gillet, C. (eds.), *Scientific Composition and Metaphysical Ground*, London: Palgrave-Macmillan, 143-169.

Sider, T. 2020, *The Tools of* Metaphysics *and the Metaphysics of Science*, Oxford: Oxford University Press.

Thompson, N. 2016, "Grounding and Metaphysical Explanation", *Proceedings of the Aristotelian Society*, 116, 3, 395-402.

Van Fraassen, B. 1980, *The Scientific Image*, Oxford: Oxford University Press.

Wilson, J. 2014, "No Work for a Theory of Grounding", *Inquiry: An Interdisciplinary Journal of Philosophy*, 57, 5-6, 535-579.

Wilson, J. 2021, *Metaphysical Emergence,* Oxford: Oxford University Press.

What Everett Couldn't Know

Tom Schoonen

University of Amsterdam

Abstract

In an impressive feat of combining modal metaphysics with fundamental quantum mechanics, Wilson (2020) presents a new genuine realist metaphysics of modality: Quantum Modal Realism. One of the main motivations for Wilson's project is to do better than existent realist metaphysics of modality with regards to epistemic challenge: we should be able to explain our knowledge of modality. In this paper, I will argue that there is a significant worry for the epistemology of Wilson's modal metaphysics, one that parallels Rosen's objection to Lewis genuine modal realism. That is, quantum modal realism fails to explain why our ordinary methods for gaining modal knowledge are reliable. I argue that this means that with regards to the epistemic challenge, Wilson's modal metaphysics is, at best, as well off as Lewis', but potentially worse.

Keywords: Quantum modal realism, Epistemic challenge, Epistemology of modality, Naturalised modal metaphysics, Modal realism.

1. Introduction

Modal metaphysics concerns the nature of modality. More generally, a metaphysical theory should meet two requirements. First of all, the metaphysics should allow for a more than nominal role of science in constraining metaphysics. That is, in Bryant's (2020: 1869) words, the metaphysics should not be free range. Call this the **Cooped Up** desideratum. Secondly, for any field of inquiry, the metaphysics of that field should be compatible with a relevant epistemology, so that it can comply with the integration requirement (Peacocke 1999: 1; Roca-Royes 2021: 158; Sjölin Wirling 2021: 5658). Call this the **Integration** desideratum.[1]

[1] My focus will be on **Integration**, so the motivations for **Cooped Up** are not of special importance to us (that is, if it turns out to **Cooped Up** is unmotivated, I need not assume that Wilson's theory satisfies it as I do below, yet the epistemological worries that I raise are unaffected by this). There are, however, some motivations one can give for something like **Cooped Up**. The main worry is that without it, one's metaphysics has to rely on dubious (philosophical) intuitions that lack any epistemological warrant (see, for example,

Argumenta 10, 1 (2024): 161—172
ISSN 2465-2334

DOI 10.14275/2465-2334/20240.sch
First published: 30 November 2024

Rosen (1990: §6) argues that Lewis' (1986) theory of modality fails to satisfy **Integration**. Rosen starts with the assumption that any theory of modality should be able to explain that "our usual methods for forming modal beliefs are generally a good guide to the modal truth". For failing to do so, would "lead rather quickly to modal scepticism, the view that we have no modal knowledge; a claim which, like most strong sceptical theses, is very hard to believe" (339). The central tenet of Rosen's objection is that it is "profoundly puzzling" why our imaginative capacities (which ordinary agents seem to use reliably to find out modal truths) would

> truly describe a domain of objects [i.e., Lewis' possible worlds] with which human beings had absolutely no contact when those principles were being shaped, presumably by a perfectly natural evolutionary process? After all, there might have been creatures whose imaginative principles were quite out of step with the distribution of worlds in modal space. How is it that we are so lucky as to have been given the *right* imaginative dispositions? (1990: 340, original emphasis).

Wilson (2020) agrees and suggests that failure to be compatible with a plausible epistemology is one of the main challenges faced by traditional realist theories of modality, one which they have failed to overcome (6).[2] In an impressive feat of combining modal metaphysics with foundational quantum mechanics, Wilson aims to do better and to provide a naturalised modal metaphysics that is supposed to improve on classic realist theories of modality in relation to **Integration**.

Wilson's modal metaphysics, based on the Everettian, or many-worlds, interpretation of quantum mechanics, is dubbed *Quantum Modal Realism* (QMR). Through a number of elaborate arguments about the nature of the Everettian interpretation, objective chance in such an interpretation, and the overall utility of his theory, Wilson tries to establish the thesis that "[t]o be a metaphysically possible world is to be an Everett world" (22). In this paper, I will assume that Wilson's theory satisfies **Cooped Up** and I will not question his interpretation of Everettian worlds as diverging rather than overlapping.[3] Instead, I want to focus on Wilson's comments on the advantage QMR has when it comes to the epistemology of modality. I will suggest that Wilson's metaphysics potentially does worse than Lewis' when it comes to Rosen's formulation of **Integration**.

I will first briefly set out Wilson's modal metaphysics and the corresponding epistemology that he suggests (Section 2). After this, I will argue that there is a significant worry for Wilson's modal metaphysics that parallels Rosen's objection to Lewis (Section 3). Finally, I will consider a possible response on behalf of Wilson, in Section 4 and argue that it fails. I conclude that, with regards to **Integration**, Wilson's modal metaphysics is, at best, as well off as Lewis', but potentially worse.

Ladyman et al. 2007; Bryant 2020; Wilson 2020 and references in Bryant 2020). Sharpening these arguments, Bryant argues that properly identifying theories that fail to satisfy **Cooped Up** shows them to "not produce justified theories of reality, *since the constraints on its content are not sufficiently robust and their satisfaction secures insufficient epistemic warrant*" (2020: 1868, emphasis added). Thanks to an anonymous reviewer for urging me to point to some motivations for **Cooped Up**.

[2] All page numbers related to Wilson's work are to Wilson 2020 unless otherwise indicated.

[3] See Divers 2022 for a terminological note on the use of 'diverging'.

2. Wilson's Theory of Modality

In this section, I will very briefly set out Wilson's modal metaphysics and the corresponding epistemology.

2.1 Quantum Modal Realism

In quantum mechanics, on the 'standard' interpretation, there are taken to be two (fundamental) rules that describe the way that very small objects (e.g., electrons), and systems composed of them, behave. The Schrödinger equation, which describes the behaviour of unobserved systems, and the Born Rule, which describes the behaviour of systems when observed (e.g., through measuring them). When a quantum state evolves into a superposition, the standard interpretation has it that there is a fundamental indeterminacy to the state. Yet, when we measure something, we never experience such indeterminacy (remember Schrödinger's cat). The problem is that when these two rules are applied hinges on the very vague, and ultimately unclear, notion of "measurement" (this is one way of setting up the measurement problem).

One way out of this problem, which has come to be known as the Everettian or many-worlds interpretation of quantum mechanics, is to hold that there is only *one* fundamental rule, namely the Schrödinger equation, and to 'replace' the indeterminacies of superpositions by a multiplicity of universes. So, whenever the orthodox suggested that one quantum state is in a superposition, and thus ultimately includes some fundamental indeterminacy, Everettians suggest that the quantum state *splits* into two states, each perfectly determinate. As Wilson puts it:

> The quantum dynamics generically evolves quantum states into *superpositions*; where the orthodox interpretation took superposed quantum states to represent single systems with unfamiliar indeterminate properties, Everett proposed taking superposed states to represent multiple systems each with familiar determinate properties. In other words, the central idea of EQM is to replace indeterminacy with multiplicity (77, original emphasis).

This means that whenever a superposition occurs, the complete quantum state splits into two complete universes. One interpretation of this splitting, favoured by Wilson, suggests that these split quantum states are complete, non-overlapping worlds.[4]

With the Everettian multiverse in hand, Wilson suggests that we have all we need to provide a modal metaphysics: *Quantum Modal Realism*. The core tenet, for our purposes, is the claim that "[t]o be a metaphysically possible world is to be an Everett world" (22).[5] This tenet, which Wilson calls *Alignment*, entails two further principles:

> *Individualism*: If X is an Everett world, then X is a metaphysically possible world.
> *Generality*: If X is a metaphysically possible world, then X is an Everett world" (24).

[4] The Everettian interpretation of quantum mechanics is one of the most prominent interpretations of quantum mechanics among those working on the foundations of it (cf. Saunders et al. 2010; Wallace 2012; Carroll 2019; and Wilson 2020: Ch. 2).

[5] Some other core tenets of the theory concern the diverging interpretation of Everett worlds, the indexicality of actuality, propositions as sets of worlds, and the interpretation of objective chance (Wilson 2020: 22).

Individualism concerns a particular way of interpreting Everettian worlds, which is defended by Wilson in Ch. 3, and will not concern us much. For our purposes, *Generality*, is of interest. Given our interest in QMR's ability to satisfy the **Integration** desideratum, it will be worth to quote Wilson's motivation for *Generality* at length:

> Why accept Generality? I will argue for this principle by appeal to the theoretical unity and simplicity of the systematic metaphysics that it makes possible. Without Generality, Everettians must distinguish two fundamental and fundamentally different kinds of possibility; Generality provides theoretical uniformity. Generality also enables a wholly reductive theory of objective modality, and *a straightforward account of modal epistemology which renders it continuous with ordinary scientific inquiry* (26, emphasis added).

Wilson explicitly notes that existent (genuine) realist theories of modality face, what he calls, the *epistemic challenge* (6).[6] As Wilson points out, "[o]ther Lewisian possible worlds bear no constitutive, causal, or other explanatory relations to the observable goings-on within our own world. If Lewisian modal realism is correct, then how we ascended to our current state of modal knowledge is an intractable mystery, even if our current modal beliefs were (inexplicably) formed de facto reliably" (11).[7]

2.2 Science as a Guide to Knowledge

Wilson's modal metaphysics is deeply rooted in quantum mechanics (and a particular interpretation of it). The resulting theory is a realist theory about modality, very much akin to Lewis' (1986) *Genuine Modal Realism* (GMR), with the exception that QMR is supposed to be able to overcome the epistemic challenge. Wilson points out that a realist account of modality has to "help us to make sense of how we know which worlds are possible (*the epistemic challenge*)" (6, original emphasis).[8] That is, Wilson stresses the importance of the **Integration** desideratum mentioned above. I will now discuss the epistemology that Wilson proposes to explain our knowledge of modality.

The first thing to note is that Wilson acknowledges that providing a realist account of modality means that modality is "discovered, not invented" (61). The epistemology in question should accommodate the appropriate humility that results from this. That is, since it is not up to us which modal claims are true or not, we should not presume to have perfectly accurate or complete modal knowledge (see also Lewis 1986: 114). For our purposes, we can simply accept this and focus on the more interesting question: for the kind of modal statements that we *do* know, how do we know them?

[6] Throughout this paper, I will use Wilson's terminology of 'the epistemic challenge' and my terminology of 'the integration desideratum' interchangeably.

[7] As we will see below, in Section 4.2, Lewis does have an epistemology of modality, one that is not dependent on the lack of a causal relation between other possible worlds and the actual world. (Lewis thought that a causal connection is only needed for knowledge of contingencies.)

[8] Phrased like this, the answer for Lewis is obvious: all worlds are possible. Rather, the issue for Lewis is *which* possibilities these worlds represent (see Divers 2002: 274 for a related discussion).

Given the metaphysics presented by Wilson, and the aim to adhere to **Integration**, there is a seemingly straightforward proposal for the epistemology of QMR: let science tell us what is possible.

> In quantum modal realism, modal epistemology is entirely subsumed into general scientific epistemology. When we discover—experimentally or theoretically—that some outcome of some process has a non-zero objective chance, then we can immediately infer that there is a genuine possibility corresponding to it (63).

So, if there is a system that is "sufficiently decohered" (ibid.) and the Schrödinger equation tells us that there is a (non-zero) chance that φ, then it is also possible that φ. This means that we need to turn to theoretical and experimental physics to tell us which states of affairs have a non-zero chance, which then provides us with knowledge that those states of affairs are possible. For example, in a situation similar to that of set-up relevant for Schrödinger's cat thought experiment, physics tells us that there is a non-zero chance that the cat is alive and that there is a non-zero chance that the cat is dead. That is, there is an Everettian world where the cat is alive and one where it is dead. So, science tells us (correctly according to QMR) that it is possible that Schrödinger's cat is alive and that it is possible that the cat is dead. Call this *Quantum Theory-based Epistemology of Modality* (QTEM).

3. Everett Crosses the Street (or, the Return of Rosen)

In this section, I will present an epistemological objection against Wilson's modal metaphysics (a parallel of Rosen's objection to Lewis). The objection has it that Wilson cannot explain the reliability of the methods that ordinary agents use to gain modal knowledge and that theories that can't do so would "lead rather quickly to modal scepticism", a price we shouldn't pay for any theory of modality (Rosen, 1990: 339). This is particularly pressing for Wilson, as addressing the epistemological challenge is one of the main motivations for his theory.

Consider Hugh as he is getting ready to cross a busy street, while deciding which of the diverging streets to take to mail a postcard to Alastair. There are a number of modal judgements that Hugh needs to make, which all rely on the quotidian modal knowledge that he has: can I cross before that car hits me? If I go left, will I arrive at the mailbox before it gets emptied today? *et cetera*. Arguably, Hugh will rely on his imagination (imagining how quickly he can cross the street and how quickly the car approaches) or similarity and analogical reasoning (last week he took the left street and it took him 10 minutes to get to the mailbox) to do so. Both of these methods have been proposed to explain our (philosophically interesting) modal knowledge (see, respectively, Byrne 2005; Kung 2010; Balcerak Jackson 2018; Gregory 2020 and Hawke 2011; Roca-Royes 2017; Dohrn 2019, Schoonen n.a.). It seems that Hugh, like most of us, has swaths of such modal knowledge.

The problem is explaining why "our usual methods for forming modal beliefs are generally a good guide to the modal truth" (Rosen 1990: 339). Call this the *folk challenge*.[9] I claim that Wilson's QTEM fails to meet this challenge.

[9] See Rosen 1990: 337-339, Williamson 2007: 162, and Sauchelli 2010: 347-348, for similar remarks. Combined with the claim that ordinary agents have swaths of modal knowledge, this is an instance of what Alexander & Weinberg (2014) call the *general reliability thesis*:

Note that, with regards to QTEM, in gaining the kind of knowledge exemplified by Hugh, one does *not* put "on a labcoat or fire up a statistics program" (Nolan 2017: 9). That is, it doesn't seem to be the case that ordinary agents, in acquiring their ordinary modal knowledge (which, occasionally needs to be acquired within seconds, Williamson 2016: 116), rely on theories, let alone the findings of experimental and theoretical physics.[10] For example, Fischer (2016: 240, original emphasis), who defends an epistemology of modality similar to QTEM, notes that "[i]t *isn't* plausible that I—with my embarrassingly poor understanding of physics—am in any position to assess what is and isn't possible for neutrinos [or quantum states]. It takes more than a passing familiarity with the relevant theories to make such assessments". Especially since the modal knowledge that we have is crucial for our going about the world (see, e.g., Byrne 2005; Nichols 2006; Williamson 2007), which means that sometimes, modal judgements have to be made in a split second (consider Williamson's (2016) example of jumping a river while being chased by a wild animal). Even if in principle we could do such quantum calculations (and it is not obvious that we can, see footnote 13), this cannot be the method by which ordinary agents gain the modal knowledge relevant for navigating their surroundings.

The above suggests that ordinary agents usually don't perform the required quantum calculations in order to determine what is possible. What about some methods that have been appealed to in order to explain ordinary agents' knowledge of modality: imagination, similarity reasoning, perception, *et cetera*? Might they be able to explain ordinary agents' knowledge of possibilities given Wilson's QMR?[11] It seems that none of these methods are straightforwardly interpreted as being related to modal space as it is described by the Schrödinger equation. That is, even if these methods might explain some of the modal knowledge that ordinary agents have, QMR cannot *explain* why this is so. That is, QMR cannot meet the folk challenge, as there is in general no reason to think that any of the methods that we rely on in knowledge acquisition can provide us with (experiential) evidence of non-zero probabilities in quantum states.

So, QTEM does not seem to be the method through which ordinary agents acquire modal knowledge and QMR cannot explain the reliability of the methods we *do* seem to use in our ordinary modal judgements. This suggests that QMR might not do so well with regards to the epistemic challenge as Wilson suggests. We can phrase the worry more directly in terms of the integration desideratum. Possibilities, on QMR, depend on whether or not some particles are in a superposition and thus split the universe. Assuming that this in fact gives us plenitude and that the world does indeed split for each of the possibilities that

though fallible, ordinary agents' epistemic judgements are generally reliable when concerning mundane cases. The folk challenge can be thought of as a specific instance of the epistemic challenge: the epistemic challenge concerns 'our' knowledge of modality, where this 'our' is interpreted as 'ordinary agents' in the folk challenge (rather than 'philosophers').

[10] This is *not* to say that the cognitive capacities that experimental physicists rely on when doing their quantum calculations are significantly different from (regimented) cognitive capacities used in everyday life. It is just that we don't seem to use the scientific method in order to acquire everyday (modal) knowledge. Thanks to Giacomo Giannini for pushing me to make this clearer.

[11] Note that even if this works, this is already a significant move away from Wilson's preferred epistemology of modality, QTEM.

we think there are,[12] the epistemological challenge is to explain the reliability of the methods of ordinary agents in tracking this (cf. Schechter 2010).[13] However, there is absolutely no reason to assume that there are any methods that ordinary agents use in knowledge acquisition that track superposition or quantum split universes. That is, there seems to be no explanation linking modal judgements of ordinary agents to the metaphysical possibilities that there are on QMR.[14]

4. Wilson, Lewis, and Ordinary Agents

So, Wilson's epistemology leaves the ordinary agent on the street high and dry when it comes to their modal knowledge. I will now consider a possible response on behalf of Wilson: Wilson is simply not concerned with the modal knowledge of ordinary agents. I will first argue that it is not strange to assume that he *should* care about the modal knowledge of ordinary agents given his commitments to naturalism. Secondly, I will argue that, regardless of the previous argument, Lewis *can* explain the modal knowledge of ordinary agents, so if Wilson can't or isn't concerned with it, then his theory is not an improvement over existent genuine realist theories of modality with regards to the epistemic challenge.

4.1 Ordinary Agents' Modal Knowledge

Of course, Wilson might retort that his epistemology is not intended to explain the modal knowledge of ordinary agents and that he simply is not interested in explaining that. I will argue that Wilson, as a naturalist, should care about explaining the modal knowledge of ordinary agents, or, at the very least, that it is not farfetched to think that he should.

Wilson puts a lot of emphasis on his naturalistic methodology with regards to his modal metaphysics (esp., sec. 0.4). This kind of naturalism is, what is sometimes called, *ontological* or *metaphysical* naturalism: what there is in the world is that what science tells us there is (cf. Nolan 2017; Papineau 2021). Nolan (2017: 12-13) suggests that accepting metaphysical naturalism (as Wilson does) *motivates* accepting naturalism with regards to the *epistemology* of modality. He characterises methodological naturalism, in the sense relevant for the epistemology (of modality), as follows:

> *[M]ethodological naturalism*, is the approach that requires that philosophical *methods* be those of the natural and social sciences, or at least that those methods be

[12] See Wilson 2020: Sec. 1.8 for a defense.

[13] There is a stronger worry in the vicinity of this one for Wilson. For it is not at all obvious that we (i.e., theoretical and experimental physicists) can in fact translate quantum mechanical phenomena into macro phenomena and *vice versa*. That is, it is unclear how knowing how to solve the Schrödinger equation in a particular instance can tell us anything about whether the car will turn right or left. I will leave this worry aside for the purposes of this paper. Thanks to Giacomo Giannini for bringing this worry to my attention.

[14] The ignoring of quantum possibilities precisely *because* it seems obvious that ordinary agents are not concerned with them can be found in a broad spectrum of philosophical debates. For example, see Carey 2009 on core cognition; Lewis 2016 on evaluating counterfactuals; Aimar 2019 on evaluating disposition ascriptions; and and Schoonen & Jones (n.a.) and Boardman & Schoonen (n.a.) on imagination.

> of the same general kind and be generally harmonious with the methods of the sciences, particularly the natural sciences (Nolan 2017: 8, original emphases).

On one reading of this definition, Wilson's suggested epistemology is straightforwardly naturalistic: it simply *is* science that tells us what is possible. Call this **Narrow Naturalism** (as I will focus exclusively on methodological naturalism, I will drop the 'methodological'): science and the scientific method provide us with (modal) knowledge.

However, note that this is significantly different from the kind of naturalism we usually find in epistemology (e.g., Quine 1969; Goldman 1986; Kornblith 2002). This kind naturalism has it that epistemologists *turn to* science to see what cognitive capacities or methods they can suggest agents rely on when acquiring knowledge.[15] For example, the naturalistic epistemologies of, e.g., Goldman (1986) and Kornblith (2002) have it that the methods that an epistemology postulates should be beholden to and in line with our best scientific theories. Call the latter kind of naturalism, which turns to the sciences to determine which of our methods are epistemically useful and reliable, **Broad Naturalism**.

Broad naturalism is the kind of naturalism relevant to the folk challenge: we should turn to the sciences to determine which of the methods *used by ordinary agents* reliably results in modal knowledge (and, potentially, explain why this is so). Phrased in this way, this is closely related to Sauchelli's (2010: 347) *feasibility challenge*: "if empirical studies about the means by which our minds process modal judgements are available, then it seems interesting and methodologically correct to take into account such research" (ibid.: 348). Given that the antecedent of the challenge is true (with regards to, e.g., imagination, see Lane et al. 2016; Harris 2021), we better take into account how ordinary agents acquire their modal knowledge. As Sauchelli himself points out, this is supposed to be understood as "a simpler point" than "having a naturalistic stance" (2010: 348). Yet, as we saw, this is something that Wilson's theory fails to do.

From an *epistemological* point of view, it seems to me that **Broad Naturalism** is the most interesting interpretation of methodological naturalism (see also Nolan 2017: 9). It would thus be very much in line with Wilson's naturalistic commitments that he adopts it. If he does, however, he is committed to explain the folk challenge, which, as things stand, his theory seems to be unable to do. Of course, Wilson might put his foot down and stick to **Narrow Naturalism** on the epistemological side, in which case the folk challenge loses its bite. In the next subsection, I will evaluate what this would mean for Wilson's overall project.

4.2 Lewis Crosses the Street

Having to retreat to **Narrow Naturalism** and not addressing the folk challenge is, in light of Rosen's (1990) comments and Sauchelli's (2010) feasibility challenge, in and of itself, a significant strike against Wilson's proposed metaphysics of modality. Worse, I will argue that such a retreat would make Wilson's modal metaphysics *worse off* than Lewis' GMR when it comes to the epistemic challenge. This is particularly worrisome for Wilson as doing better than existent re-

[15] The former is, perhaps, more aptly called a *scientific* epistemology of modality. Thanks to Samuel Boardman for discussion here and for suggesting the label.

alist theories of modal metaphysics is one of the main motivations for Wilson's account (6).

In order to assess whether or not QMR is worse off than Lewis' GMR, it will be useful to quickly rehearse what Lewis says about (our) modal knowledge. For Lewis, what is crucial for which possibilities there are (or are represented) is the *principle of recombination*. This principle is something that the Lewisian needs to defend. However, once defended, we can explain how we get knowledge of modality. In particular Lewis (1986: Ch. 2.4) suggests that the proper method of gaining knowledge of modality is a theoretical understanding of the principle of recombination and what follows from it. "[H]ow do we come by the modal opinions that we in fact hold? […] I think our everyday modal opinions are, in large measure, consequences of a principle of recombination" (Lewis 1986: 113). This is of course very similar to Wilson's suggestion, as for Wilson the Schrödinger equation does the work that the principle of recombination does for Lewis (Wilson 2020: 28, 65-67, 145).

The epistemological work is deferred to theoretical metaphysicians, rather than (quantum) physicists, on Lewis' picture. So, perhaps Lewis has an equally hard time explaining the modal knowledge of Hugh (and ordinary agents in general)? If so, then the problems for QMR don't undermine Wilson's claim that his theory is better at addressing the epistemic challenge than existing realist theories of modality.

However, Lewis does explicitly explain how ordinary agents might gain modal knowledge by relying on imagination, which humans *do* rely on in order to make ordinary modal judgements (cf. Lane et al. 2016; Harris 2021). Given the principle of plenitude, according to Lewis, we can explain why ordinary agents rely on imagination when they are making their modal judgements.

> We get enough of a link between imagination and possibility, but not too much, if we regard imaginative experiments as a way of reasoning informally from the principle of recombination. To imagine a unicorn and infer its possibility is to reason that a unicorn is possible because a horse and a horn, which are possible because actual, might be juxtaposed in the imagined way (Lewis 1986: 90).

That is, the principle that governs the space of possibilities is tracked by the imagination in order to explain some of the knowledge that ordinary agents have of modality (even if, ultimately, philosophers need to study the principle of recombination to get knowledge of extraordinary modal claims, Lewis 1986: 113).[16] One way of understanding what Lewis is doing here, is as explaining why the methods that ordinary agents use in making modal judgements are reliable heuristics. This, in turn, can be seen as giving a proper, broad naturalistic, account of the epistemological side of the **Integration** desideratum.

Granted that humans do rely on imagination to make ordinary modal judgements, Lewis has a story to tell why it is that imagination is reliable when it comes to modal judgements and thus he can account for the folk challenge. For Wilson, however, this is not so clear. As argued above, if we assume, with

[16] Lewis' point can be strengthened by pointing out that imagination, on certain interpretations, does seem to be structured such that it is very likely to mirror the principle of recombination. This is particularly clear on Hume's (1777/1997) picture of imagination. See also Kung's (2017) discussion thereof.

Wilson, that the Schrödinger equation generates modal space, then it is no longer obvious that we can explain the reliability of the methods that ordinary agents rely on in making modal judgements, imagination in particular (again, see Schoonen & Jones (n.a.) and Boardman & Schoonen (n.a.) on imagination and quantum possibilities).

5. Conclusion

Wilson sets out to provide a (metaphysically) naturalistic account of modal metaphysics. This metaphysics is closely related to Lewis' Genuine Modal Realism, but instead of relying on the principle of recombination and concrete spatiotemporally isolated worlds, Wilson suggests that worlds are 'branched' Everettian universes as specified by the Schrödinger equation. The main upside of Quantum Modal Realism over Genuine Modal Realism, according to Wilson, is that it can deal with the epistemic challenge: a modal metaphysics "must help us to make sense of how we know which [possibilities there] are" (6).

Wilson suggests that we know which possibilities there are by relying on theoretical and experimental physics; that is, science has to tell us what is possible and what not. This is, though not in letter, in spirit similar to Lewis' suggestion, who suggests that it is theoretical metaphysicians who have to tell us what is possible and not based on the principle of recombination. However, there is another challenge for theories of modality—i.e., the folk challenge—that requires theories to explain the knowledge that non-expert adults have of possibilities. Interestingly, Lewis seems to be able to address the folk challenge, whereas it is not obvious that Wilson can.

One final retreat for Wilson might be to piggy-back on Lewis' explanation. The rough idea would be that *if* the Schrödinger equation and the principle of recombination create an extensionally identical modal space, then the fact that imagination tracks the principle of recombination would also explain imagination's reliability in modal judgements on the QMR picture. Note, however, that this is a pretty big *if* and it is not obvious that Wilson himself thinks that QMR and GMR are extensionally equivalent (in the sense that they generate the exact same set of possibilities). Also note that even if we grant this assumption, the conclusion is still only that QMR is *as good as* Lewis' GMR when it comes to dealing with the folk challenge and Wilson has not shown us that QMR is in a better position to deal with the epistemic challenge than, e.g., GMR.

I take it that the arguments above show the importance of being able to address the folk challenge for any theory of modality (see also Rosen 1990; Sauchelli 2010; Schechter 2010). So, even though QMR might be considered better at explaining *philosophers'* modal knowledge, it fares no better when it comes to dealing with the folk challenge. In fact, it potentially fares worse in that regard. As it stands, QMR cannot be said to explain the modal knowledge that ordinary agents have. This is particularly worrisome as it questions the foundational motivation of QRM: providing a better solution to the epistemic challenge than existent realist theories of modal metaphysics.[17]

[17] This paper was written during a fellowship at *Human Abilities*, a Centre for Advanced Studies in the Humanities (*Kollegforschungsgruppe*) funded by the *Deutsche Forschungsgemeinschaft* (DFG), where I was hosted by Barbara Vetter and Dominic Perler. Thanks to both for hosting me and allowing me to be part of a greatly stimulating research envi-

References

Aimar, S. 2019, "Disposition Ascriptions", *Philosophical Studies*, 176, 1667–1692.

Alexander, J. and Weinberg, J.M. 2014, "The 'Unreliability' of Epistemic Intuitions", In Machery, E. and O'Neill, E. (eds.), *Current Controversies in Experimental Philosophy*, New York, NY.: Routledge, 128–145.

Balcerak Jackson, M. 2018, "Justification by Imagination", In Macpherson, F. and Dorsch, F. (eds.), *Perceptual Imagination and Perceptual Memory*, Oxford: Oxford University Press, 209–226.

Boardman, S. and Schoonen, T. (n.a.), "Core Imagination", Unpublished manuscript.

Bryant, A. 2020, "Keep the Chicken Cooped: The Epistemic Inadequacy of Free Range Metaphysics", *Synthese*, 197, 1867–1887.

Byrne, R.M.J. 2005, *The Rational Imagination*, London: MIT Press.

Carey, S. 2009, *The Origin of Concepts*, Oxford: Oxford University Press.

Carroll, S. 2019, *Something Deeply Hidden*, Dutton.

Divers, J. 2002, *Possible Worlds*, London: Routledge.

Divers, J. 2022, "Book Review: The Nature of Contingency", *Mind*, 131, 524, 1357-1364.

Dohrn, D. 2019, "Modal Epistemology Made Concrete", *Philosophical Studies*, 176, 2455–2475.

Fischer, B. 2016, "A Theory-Based Epistemology of Modality", *Canadian Journal of Philosophy*, 46, 2, 228–247.

Goldman, A. 1986, *Epistemology and Cognition*, Cambridge, MA.: Harvard University Press.

Gregory, D. 2020, "Imagery and Possibility", *Noûs*, 54, 4, 755–773.

Harris, P.L. 2021, "Early Constraints on the Imagination: The Realism of Young Children", *Child Development*, 92, 2, 466–483.

Hawke, P. 2011, "Van Inwagen's Modal Skepticism", *Philosophical Studies*, 153, 3, 351–364.

Hume, D. 1777/1997, *An Enquiry Concerning Human Understanding*, Indianapolis: Hackett Publishing Company.

Kornblith, H. 2002, *Knowledge and Its Place in Nature*, New York: Oxford University Press.

Kung, P. 2010, "Imagining as a Guide to Possibility", *Philosophy and Phenomenological Research*, 81, 3, 620–663.

Kung, P. 2017, "Personal Identity Without Too Much Science Fiction", in Fischer, B. and Leon, F. (eds.), *Modal Epistemology After Rationalism*, Vol. 378, Cham: Springer, 133–154.

Ladyman, J., Ross, D., with Spurrett, D. and Collier, J. 2007, *Every Thing Must Go*, Oxford: Oxford University Press.

ronment. Additionally, many thanks to Samuel Boardman and Giacomo Giannini for feedback on an earlier version of this paper.

Lane, J.D., Ronfard, S., Francioli, S. and Harris, P.L. 2016, "Children's Imagination and Belief: Prone to Flights of Fancy or Grounded in Reality?", *Cognition*, 152, 127–140.

Lewis, D.K. 1986, *On the Plurality of Worlds*, Oxford: Blackwell.

Lewis, K.S. 2016, "Elusive Counterfactuals", *Noûs*, 50, 2, 286–313.

Nichols, S. 2006, "Imaginative Blocks and Impossibility: An Essay in Modal Psychology", in Nichols, S. (ed.), *The Architecture of the Imagination*, Oxford: Oxford University Press, 237–255.

Nolan, D. 2017, "Naturalised Modal Epistemology", in Fischer, B. and Leon, F. (eds.), *Modal Epistemology After Rationalism*, vol. 378, Cham: Springer, 7–27.

Papineau, D. 2021, "Naturalism", in Zalta, E.N. (ed.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University, Summer 2021 ed.

Peacocke, C. 1999, *Being Known*, Oxford: Oxford University Press.

Quine, W.V.O. 1969, "Epistemology Naturalized", in *Ontological Relativity and Other Essays*, New York: Columbia University Press, 69–90.

Roca-Royes, S. 2017, "Similarity and Possibility: An Epistemology of de re Possibility for Concrete Entities", in Fischer, B. and Leon, F. (eds.), *Modal Epistemology After Rationalism*, Vol. 378, Cham: Springer, 221–245.

Roca-Royes, S. 2021, "The Integration Challenge", in Bueno, O. and Shalkowski, S. (eds.), *The Routledge Handbook of Modality*, New York: Routledge, 157–166.

Rosen, G. 1990, "Modal Fictionalism", *Mind*, 99, 395, 327–354.

Sauchelli, A. 2010, "Concrete Possible Worlds and Counterfactual Conditionals: Lewis Versus Williamson on Modal Knowledge", *Synthese*, 176, 345–359.

Saunders, S., Barrett, J., Kent, A. and Wallace, D. (eds.) 2010, *Many Worlds? Everett, Quantum Theory, and Reality*, Oxford: Oxford University Press.

Schechter, J. 2010, "The Reliability Challenge and the Epistemology of Logic", *Philosophical Perspectives*, 24, 437–464.

Schoonen, T. (n.a.), "Kinds, Categories, and Possibility", Unpublished manuscript.

Schoonen, T. and Jones, M. (n.a.), "Embodied Imagination and Knowledge of Possibilities", Unpublished manuscript.

Sjölin Wirling, Y. 2021, "An Integrative Design? How Liberalised Modal Empiricism Fails the Integration Challenge", *Synthese*, 198, 5655–5673.

Wallace, D. 2012, *The Emergent Multiverse: Quantum Theory According to the Everett Interpretation*, Oxford: Oxford University Press.

Williamson, T. 2007, *The Philosophy of Philosophy*, Oxford: Blackwell.

Williamson, T. 2016, "Knowing by Imagining", in Kind, A. and Kung, P. (eds.), *Knowledge Through Imagination*, Oxford: Oxford University Press, 113–123.

Wilson, A. 2020, *The Nature of Contingency*, Oxford: Oxford University Press.

Essence and Knowledge

Daniele Sgaravatti

University of Bologna

Abstract

In this paper I will attempt to show that there are some essential connections between essence and knowledge, and to clarify their nature. I start by showing how the standard Finean counterexamples to a purely modal conception of essence suggest that, among necessary properties, those that are counted as essential have a strong epistemic value. I will then propose a "modal-epistemic" account of essence that takes the essential properties of an object to be precisely the sub-set of its necessary properties that constitute a significant source of knowledge about it. I will then argue that this view is supported by an inference to the best explanation that starts from some uncontroversial, although sometimes neglected, epistemic roles essences should play.

Keywords: Essence, Definition, Explanation, Kit Fine.

1. Introduction

In this paper I will defend the thesis that the essence of something just is a set of cognitively significant properties with a certain modal profile. More precisely, an essential property of x is a necessary property of x which constitutes a significant source of knowledge about x. And the essence of x is a set of essential properties, ideally sufficient for individuating x, which is as far as possible simple and informative. Because the picture I want to draw is very ample, I will often have to paint with a very broad brush. But the connections I wish to highlight only emerge at this very general level, and I believe this is the reason why they are too easily missed.

The plan of the paper is as follows. In the first section, I will introduce the relevant notion of essence, and I will discuss some widely accepted arguments that show that the notion of essence cannot be reduced to purely modal notions. I will argue that the same arguments already show that essences have a high degree of interest from the epistemic point of view, and I will sketch a view on which this high degree of epistemic interest is part of the definition of essence. In the second section I will look more specifically at various epistemic roles that essences play. The notion of essence is assumed by most theorists to have close connections which notions such as explanation, individuation, and definition (and sometimes induction), and these

Argumenta 10, 1 (2024): 173—187
ISSN 2465-2334

DOI 10.14275/2465-2334/20240.sga
First published: 30 November 2024

are epistemic notions (or so I will claim). I will argue that my view is supported by an inference from the best explanation based on these connections.

2. Essence and Necessity

What is essence? A definition of essence is not often offered, presumably because the notion is supposed to be primitive. Yet, there are a few things that are often said to introduce the notion. The essence of a thing, we are told, is its nature, or, following Locke, "the being of any thing whereby it is what it is". Essence is also etymologically linked to being, since it comes from Latin *essentia*, which can be translated as 'being' or 'beingness', and was introduced in the philosophical jargon to translate some Aristotelian expressions also derived from the Greek verb which expresses being.[1] Sometimes essences are also linked to real definitions: the definition of a thing, as opposed to definition of words. There could be some room, if one wished to, for complaining that we do not really understand any of these notions, and even that they do not have meaning outside the Aristotelian context where they originated. I used to make this sort of complaint. But I now think my complaint was, although not entirely unmotivated, short-sighted and, in the end, a little bit dishonest. For, after all, I can quite well understand and use the English word 'essence' and its adverbial form 'essentially'. For example, I can say that I made my complaint because my philosophical outlook is essentially the product of a twentieth-century education. Consider also some claims I found on the internet, like "the essence of true friendship is to make allowance for another's little lapses", or "the essence of government is force", as well as "egg yolks are essential for carbonara" and "water is essential to life". While I am not sure these claims are true, I do not seem to have trouble understanding them. Of course, it is not trivial that the common notion of essence employed in those claims is the same notion philosophers are interested in. But I now think that the relation between the philosophical notion and the everyday notion is not so different from the relation between the everyday (non-epistemic) use of the notions of possibility and necessity and their (non-epistemic) philosophical use.[2] It might be that the notions are identical, or that the philosophical notions are a limiting case of the everyday notions, or some sort of rigorous development or Carnapian "explication" of them. Attention to the ordinary usage of the notion does not imply that the aim of philosophical theory is just, or even mainly, an account of the ordinary notion. Be that as it may, I am only claiming that our understanding of the ordinary usage of the notion of essence is sufficient to provide some grasp of the philosophical usage, and we cannot reject the notion altogether as if it were some obscure technical notion of Aristotelian logic or medieval scholastic philosophy. I will come back at the end of this section to the relation between modality and essence and to the role of ordinary language (and thinking) in theorizing about those notions.

[1] An interesting historical complication is that *essentia* was originally introduced to translate the term we now translate as 'substance' (*ousia*), and only later used to translate the complex expression we now translate as 'essence' (*to ti ên einai).* But both are clearly related to *einai*.

[2] Livingston-Banks (2017) is one of the few authors that I know of to explicitly discuss the issue, and he takes a different view, on which "essence" is a theoretical notion, with a loser relation to ordinary usage. But surely *metaphysical modality* is also a theoretical notion, even if some of our ordinary modal talk expresses metaphysical modality. I think the issue therefore should at least not be prejudged. If a notion of essence close to the ordinary usage can be developed that makes sense of philosophical claims as well, that certainly counts in its favor.

Because my philosophical outlook is, however, essentially the product of my twentieth-century education, the view of essence I will propose does not make it a metaphysical primitive. It also does not support some fairly popular philosophical views about the relationships between essence and some other metaphysical notion. For example, my view of essence does not support the idea that the notion of essence helps us to make sense of other metaphysical notions such as "grounding", "fundamentality", or "ontological dependence". It also does not support the view that essence grounds or explains metaphysical modality; in fact, it seems to be incompatible with the latter claim, because it defines essence using modal (although not only modal) notions.[3] If one however thinks that essence is instead a metaphysical primitive, it should be stressed that most of what I say here about the epistemic role of essence is independent of this issue.

I will start with at least one assumption about the relation between essence and modality, one that is fairly uncontroversial in the contemporary debate. If something has a property essentially, then it has that property necessarily. An essential property, in other words, is one which an object could not fail to have. Importantly, I will also assume (again, this is relatively uncontroversial nowadays) that there are conclusive reasons to think the reverse entailment does not hold: it is not always the case that a necessary property of something is an essential property. Most theorists who write on this matter have been convinced of the latter claim—if they did not accept it already—by Fine (1994). I suppose it is likely that anyone reading this paper is already familiar with the arguments in the by now classic paper by Fine.[4] However, I need to briefly rehearse those arguments, because I will claim that, as well as establishing their intended conclusion, Fine's counterexamples to the identification between essentiality and necessity of properties also support a further conclusion, namely that essential properties are necessary properties which have a special epistemic importance. It is worth pointing out immediately that this is not at all something Fine would want to deny. Anyone who thinks that some, and only some, necessary properties are essential, will probably think that whatever metaphysical feature marks the essential properties also provides them with epistemic interest. If essential properties form "the structure of the world", it is interesting to know what they are, and, presumably, this is potentially the source of much further knowledge. The alternative suggestion I wish to spell out is that the explanatory order can be reversed: necessary properties which present a high epistemic interest get singled out as essential.

Let us consider, to begin with, three counterexamples that Fine provided to the view that all necessary properties are also essential.

(1) Socrates is necessarily a member of the singleton {Socrates}, but he is not essentially a member of that set.

(2) Socrates is necessarily distinct from the Eiffel tower, but he is not essentially distinct from the Eiffel tower.

(3) Socrates is necessarily such that 2+2=4, but he is not essentially such that 2+2=4.

[3] Rayo (2013) on the other hand offers a metaphysical and semantic framework in which an epistemic notion of essence could be employed, or at least housed.

[4] It is not however uncontroversial that Fine's objection cannot be met by some modification of the simple modal view that identifies necessity and essentiality. In fact there is a growing number of accounts that attempt that. See e.g. Wildman 2013, 2016, Torza 2015, De 2020.

These claims seem very plausible. They are even more plausible if we put the second conjunct slightly differently, in terms of the relevant property being part of Socrates' essence. For we see immediately that if properties of the sort mentioned in (1), (2) or (3) were part of his essence, many more of the same kind would be, and his essence would then be a very complex and messy sort of construction. But why are these claims plausible? One distracting feature of the claims is that it is obvious that Socrates has these properties. Consider a variant of case (3): Socrates is such that Fermat's theorem is true. This is also necessarily true, and it is in a sense not at all obvious. But we still do not find plausible that it be part of Socrates' essence to be such that Fermat's theorem is true. A more interesting thought that applies to cases (2) and (3) is that the relevant properties are shared by everything, as in the case of (3), or almost everything, as in the case of (2). However, 1 mentions a property that only Socrates possesses, so this cannot be the crux of the matter.[5] The general feature of the properties involved is that they do not tell us something very interesting about Socrates. This is well illustrated by the asymmetrical relation between Socrates and his singleton. While it is not essential to Socrates that he is a member of {Socrates}, it is essential to {Socrates} that it has Socrates as a member. Being a set that contains Socrates as the sole member is a very good candidate for being the essence of {Socrates}. Not only this feature uniquely identifies the set, but it seems to be pretty much everything there is to know about it. On the other hand, although Socrates is uniquely identified by the property of being the sole member of {Socrates}, there is lot more about him that one could want to know.

A further epicycle of the discussion is worth considering, although I will only be able to scratch the surface. A property such as being human seems a good candidate to be a necessary and essential property of Socrates, or any other human being. But now suppose, as it is standard, that it is possible for Socrates not to exist. Suppose also that if he did not exist, he would not be human (after all, being human seems to imply being a concrete being). He is possibly not human then. One way of solving this problem is neutral with respect to the distinction between necessity and essence, and it allows that Socrates is human even when not existing. In possible worlds language, 'Socrates is human' would be true *at* worlds in which he does not exist, although not true *in* those worlds, while 'Socrates exists' would not be true in or even at those worlds (see Adams 1981 and Fine 1985). A different reply consists in allowing a claim of the form 'a is necessarily F' to be true just in case a is F whenever it exists. This seems to be in line with the intuitive thought that an object has a necessary or essential property just in case the object could not exist without that property. However, while this solution allows us to say that Socrates is necessarily human, it also makes existence a necessary property (assuming it is a property) of Socrates, and anything else at all, because everything exists whenever it exists. Williamson (2013) has defended the view that everything necessarily exists on independent grounds, without denying that being human, or, more generally, having any property, requires existing, and without appealing to the distinction between *true at* and *true in*. On his view, Socrates does not have the property of being human necessarily, but he does

[5] Sometimes properties shared by absolutely everything are called "trivial". Della Rocca (1996) would perhaps count as trivial, in a distinct but related sense, also the property mentioned in (1), as a consequence (for Socrates) of a trivial property in the stricter sense. But the one mentioned in (2) is not trivial in its sense (Della Rocca 1996: 3).

possess a conditional property, that of being human whenever he is concrete—that is, as we might put it, whenever he exists spatio-temporally. So there are some views on which existence is a necessary property of absolutely everything. But, as Fine notes, existence does not seem essential to Socrates. It also does not seem essential to most other things. To make this vivid, consider that even if one holds the Williamsonian view on which existence is a necessary property of everything, one might want to say that there are some things for which it is worth asking whether existence is *also* essential to them, such as God or the whole universe, or in general that it is a separate question whether some things exist essentially (if one allows for this distinction, the view that everything exists necessarily might look less implausible). What is crucial for our present purposes is that, once again, our judgements about essentiality correlate with our judgements about epistemic interest. Supposing that there are no non-existent things, knowing of something that it exists does not tell us anything at all about that thing. It does not allow one to deduce, or otherwise infer, any further property whatsoever of what we are talking about. But if there is a being that has existence among its essential properties, then this is a crucial piece of knowledge about it.

So here is a view about what makes a property essential that seems to be not only compatible with, but indeed suggested by, Fine's arguments:

> Essential-Property-Definition (EPD): a property of an object is essential just in case it is necessary that the object has that property and the fact that the object has that property is a significant source of knowledge about the object.[6]

On this view, because what is a significant source of knowledge depends on what cognitive capacities we have, whether a property is essential partly depends on the nature of human beings; and it could also be argued that the view makes what properties qualify as essential depend on specific contexts.[7] In this sense, the view might be counted as an anti-realist or deflationary conception of essence, although it certainly does not make possible for us to stipulate essences into existence. I am not assuming any precise account of knowledge, but I am assuming that knowledge requires at least true belief and some connection between belief and truth, so a broadly externalist or "anti-Gettier" component.[8] Therefore, there are objective facts about what, given one's epistemic position, is conducive to further knowledge.

EPD might be paired in various ways with a definition of essence, as opposed to essential property. For the sake of this paper, I will work, when needed, with the following:

> Essence Definition (ED): The essence of X is a set of properties such that 1) Each property in the set is essential to X, 2) The set specifies sufficient conditions for being X, 3) Where there is more than one set satisfying 1 and 2, the set has the

[6] What it means for knowledge to be *about* something is a good question. But it is not a problem for my definition, unless we assume we cannot have beliefs about something without knowing its essence, a view I find very implausible, and I argued against in Sgaravatti 2016.

[7] See Paul 2004. Lewis 1986 may also be counted as presenting a contextualist account of essence.

[8] I mean this requirement to be compatible with the "knowledge-first" view in epistemology, although on that view there is no way to spell out the requirement without appealing to the notion of knowledge.

best ratio of simplicity to capacity to provide knowledge about X (where more than a set satisfies the 3 conditions, each of them can be called an essence of X).[9]

Conditions 1 and 2 are, I believe, one natural way to move from essential property to essence. Condition 3 will receive some attention below.

Supposing one is not opposed to the idea that there is an epistemic element in the notion of essence (an idea that will be defended and made more precise in subsequent sections), it could be asked why we need a modal element at all in our notion. In some ordinary contexts, "essential" might seem to mean simply very interesting or very important. It might be that this is one meaning of the term. But first, I believe there clearly is a sense of "essence" in which there is a connection between essence and existence, in ordinary contexts as well. Looking at the examples cited above, if the essence of government is force, then a government completely separated from force cannot exist, and if egg yolks are essential to carbonara, then you cannot cook carbonara without eggs.[10] One could object to a claim like "water is essential to life" that we can imagine alien or artificial forms of life that do not rely on water; but it seems to me to that this is equally an objection to "water is necessary to life" and to "water is essential to life".

In the next section, I will focus on the epistemic role of essence, but the connection between essence and modality will again emerge very clearly.

3. The Epistemic Roles of Essences

In this section I will look at several more specific ways in which grasping essential properties is connected to gaining knowledge about the object possessing the property. As noted above, this is not something defenders of essence as a metaphysical primitive, or anyway defenders of essence as a purely metaphysical notion, want to deny. My strategy in general will be this. To explain a certain epistemic role of essences, my opponent has to postulate a) that there is a metaphysical juncture well represented through essence-talk, and b) that our minds, our cognitive faculties anyway, are attuned to those fundamental metaphysical facts. On the other hand, my view has no extra explanatory work at all, because the view is that we single out necessary properties as essential precisely when they can play an epistemic role.

Here is a list of epistemic roles of essence (I will discuss them in some more detail below) that constitute the evidence my view is supposed to explain:

a) Definition

The connection between definition and essence goes back at least to Aristotle, and Fine sees it as the main alternative to the modal conception of essence in the history of Western philosophy. However, like explanations, definitions are supposed to provide understanding; in the case of real, as opposed to nominal, definitions, understanding of the object or phenomenon defined.

b) Explanation

The essence of a thing, which is also natural to call its "nature" in this connection, is supposed to have the potential to explain some or, together with other facts, all of the thing's other features. Some theorists have put this feature of

[9] I am assuming it is always possible to satisfy condition 2, for we may include being identical to X, or some similar condition, in the essence. If one thinks these are not real properties, or wants to rule them out, condition 2 could be omitted altogether.

[10] This would imply that "vegan" is a non-intersective adjective; cfr. "vegan steak". At any rate, I am not committed to the truth of those ordinary claims, only to their intelligibility.

essence at the center of their accounts of essence (e.g. Gorman 2005, Kment 2014, Sullivan 2017). While the notion of explanation is itself controversial, and one reason for that is precisely that it can be read in a more metaphysical or a more epistemic way, it clearly has a connection with understanding, which is itself an epistemic notion. Grasping the essence of something is supposed to provide understanding, and to explain, something about the object or kind.

c) Recognition/individuation

Lowe has a very interesting "transcendental" argument for the conclusion that we have knowledge, or anyhow some grasp, of essence. Without some such knowledge, he claims, we would not have a capacity for recognition across time, we would not be able to tell whether an object is the same we encountered before (see Lowe 2008: 27-28).[11] For example we could not know whether a certain dog is the same we encountered yesterday if we did not have some grasp of what is essential to an object of its kind.[12] Whether or not the argument is sound, it points to further interesting epistemic role for essence: essences are supposed to help us recognizing things and kinds through time and space.

d) Epistemology of modality

This point will need some more discussion below, because it apparently presents a disadvantage for my view. It might seem that understanding an essential property as an epistemically interesting necessary property makes knowing that a property is necessary a precondition for knowing that the property is essential, and therefore makes it impossible to use the notion of essence in the epistemology of modality. And yet, many authors have claimed that essences have a crucial role in the epistemology of modality (e.g. Lowe 2012, Hale 2013, Kment 2014, Mallozzi 2021).

I will argue however that my view is capable of doing justice to the epistemic role of essences in the epistemology of modality too, and in fact it can do that better than other views.[13]

For reasons of space, I will obviously not be able to cover all topics in detail. I will however go through the list in the order in which I anticipated them.

A. Definition

As noted above, there is a long philosophical tradition that links definitions in this sense, sometimes called *real* definitions, to essences.[14] The real definition of something is not just a description that applies to what is defined, but rather some

[11] Lowe also offered a different argument for the same conclusion, based on transcendental considerations on the possibility of thought. Unfortunately, that argument does not stand scrutiny, or so I have argued in [author's reference removed].

[12] This argument is neutral, as I understand it at least, on the issue whether a judgement of this kind is an identity judgment. Wiggins (1980; 2001), takes that view. I am inclined to believe that those judgements are not, strictly speaking, identity judgements. For a systematic development and defence of this kind of view see for example Fara 2008 and 2012.

[13] A further epistemic role of essence could connect essences and induction, so that an induction is stronger (or even only acceptable at all) when the predicates involved express essential properties. I am not convinced about this strategy, which anyway involves very complex issues. But it is worth noting that my list of epistemic roles of essence is not meant to be exhaustive.

[14] Aristotle, Metaphysics VII, 1031a12: Obviously interpreters have dwelled on Aristotle's account of essence (and definition). See e.g., Kung 1977.

description which captures the nature of the thing, helps us to predict the other properties of the thing and explains, together with other, perhaps contingent, facts why the thing has those properties. Now, supposing this notion of definition makes sense, we could explain it in terms of essence. Real definitions will answer the "What is it?" question about the definiendum, thereby giving its nature or essence. I believe there are several reasons to think this will not work. Before explaining why, let me digress, by looking at some remarks from a time when the notion of real definition was taken to completely hopeless (back in the twentieth century). In what is still, in this author's view, a useful book on the subject of definitions, Robinson (1954) writes in connection to his scepticism about real definitions that the expression "what is x?" is "the vaguest of all forms of question except an inarticulate grunt" (p. 190). I disagree. I see no vagueness at all in the question. However, it is true that "what" is context-sensitive. In different contexts, different answers (or sets of answers) will be admissible. If one thinks attributions of essential properties are not similarly context-sensitive, this is a problem. Plausibly the solution would be to isolate a context, or class of contexts, in which the appropriate answer to the "what is x?" question will be a specification of its essence. This exactly holds for my view, except that the appropriate answer has to specify the essence of x in the epistemic context where the question is asked.[15] Moreover, it has to be noted that to reject real definition as a useful category, one has to withhold the analytic/synthetic distinction. Only if we can isolate facts about the meaning of, say, "water", we can isolate the definition of the term from the more general endeavour of communicating interesting facts about water. I am not arguing for this view here of course, but I do not adhere to the theory that statements can be usefully categorized as synthetic or analytic. I will therefore from now on talk about definitions without qualifications (with the caveat that insofar as the distinction makes sense, I am talking about real definitions).

So I agree that definitions, at least sometimes, succeed by expressing the essence of the definiendum. However, *definition*, although this is not always acknowledged, is an epistemic category, or something near enough. Definitions have the purpose of providing understanding; and understanding is an epistemic category. In other words, definitions essentially have an epistemic function. Real definitions express essences only when they are successful. In order to be successful, however, a definition must not only be extensionally correct. It should also be illuminating, or in other terms it should serve the purpose of allowing someone who grasps it to have some understanding of what is defined. This is often expressed, in scientific contexts, by saying that definitions should be *fruitful*. The fruit they bear is of course a successful scientific discipline, which certainly means (among other things) an increase in our knowledge of its subject matter.

I will consider two ways in which definitions aim to go beyond extensional adequacy. Definitions, among other things, should be simple. To illustrate, consider this example from Lowe:

> (E1) An ellipse is the locus of a point moving continuously in a plane in such a fashion that the sum of the distances between it and two other fixed points remains constant […]

[15] The same answer however can be appropriate in a multitude of contexts.

(E2) An ellipse is the closed curve of intersection between a cone and a plane cutting it at an oblique angle to its axis greater than that of the cone's side (Lowe 2012: 936).

Lowe thinks E1 expresses the essence of an ellipse, while E2 merely expresses a necessary property (so E2 is also a further counterexample to the simple identification of essential properties with necessary ones). I do not disagree. E1 is a much better definition. It is also, not coincidentally, a much better source of knowledge, mostly in virtue of its greater simplicity. The latter comparison, however, holds for beings similar to us with respect to mathematical thinking. We can easily imagine alien beings, or even divine beings, that are extremely different from us in that respect. For a (mathematically) omniscient being, there would presumably be no difference in usefulness between E1 and E2. We may also imagine alien beings for which E2 would be simpler to understand than E1. What would beings of this kind claim about the essentiality of the complex properties expressed by E1 and E2? Of course one could insist that such beings would still believe that E1 is the correct definition, while E2 is not, despite the fact that there is no difference for them in terms of usefulness or epistemic value. I do not see why they should. I know of no ontological theory of geometrical entities that would suggest that, independently of our sense that E1 is a simpler, more fruitful, and more useful definition.

A related issue about definitions is the following: definitions shouldn't be circular. Saying that an ellipse is an ellipse, or water is water, is not a good definition, and in fact one is tempted to say these are not even attempts at a definition. Why is that, however? My view has a very straightforward answer. It is not helpful to tell someone that an ellipse is an ellipse. It does not represent any possible source of further knowledge or understanding. Similarly, it is not helpful, in most contexts at least, to be said that ellipses are elliptical figures, or that water is the watery substance.

At this point I must consider an objection based on the notion of haecceitas, or thisness.[16] Some theorists think that the essence of an individual is constitutively related to its numerical identity, and nothing more. On this view, the essence of Socrates, or at least part of it, is being Socrates, *simpliciter*. So if an exact duplicate of Socrates had been created, it would nonetheless have failed to be Socrates, despite having all his other intrinsic properties, because he would have lacked Socrates' thisness. Now such an essence, or essential property, would seem to be a counterexample to my view. If I expressed Socrates' essence, or his real definition, saying that Socrates is Socrates, this wouldn't lead to any understanding or further knowledge; but I would be nonetheless correct. However, the view on which Socrates' essence just is his haecceitas does not pose a very worrying problem. The property is necessary and sufficient for being Socrates, and there is no more informative property that can play that role, on the view under discussion. So condition 2 in ED is satisfied, and condition 3 also, although vacuously. But would that be an essential property at all? There is a sense in which thisness represents a source of knowledge about Socrates, namely the knowledge of which individuals he is identical or different to in different modal circumstances. If it is true that Socrates could be a fried egg, then I can only know that if I somehow see that the possible circumstances in which this happens are relevant to the

[16] Thanks to Maria Scarpati for pressing me on this issue.

evaluation of the claim, and I can do that only if I have some grasp of the thinness of Socrates' essence.

So my definition of essence (ED) predicts that Socrates' haecceitas is his essence (if there are no further essential properties), despite its lack of epistemic power.[17] Once we see this, it is also easy to see how that property can be part of Socrates' essence even if he has further essential properties in my sense. The essence must be sufficient to individuate Socrates. It gets therefore to be added to the set of properties constituting the essence.

B. Explanation

Many examples we have already seen make it clear that there are links between essence and explanation, links that are also traditionally accepted. There is also little doubt, it seems to me, that explanation is either an epistemic notion, or one that has itself strong ties to epistemic notions. For example, it seems plausible that to have (the) an explanation of a fact *F* is to know (the) an answer to the question of why *F* is the case. I will not dwell on these points here. I will further illustrate instead the connection between explanation and essence taking the chance to compare my account to one that is very similar in spirit (or so I believe), the view proposed in Sullivan 2017 (all quotes in this section are from that paper). Sullivan at some points characterizes her view as a sort of eliminativism about essence, "anti-essentialism", or the view that there are no essential properties; but she also calls it "explanation-relative essentialism" (59-60). I believe the latter is a much better characterization, insofar as her view does not aim to eliminate talk of "essence" and related expressions from our vocabulary and grants the truth (in a context) of some attributions of essential properties.

Explanation-relative essentialism claims that "an essence ascription is true relative to an explanatory framework if and only if an object is ascribed that property in any good explanation of that type, and there are objective norms governing explanatory frameworks in that domain" (56). Physics, metaphysics and economics are offered as three distinct examples of explanatory frameworks that allow true essence ascriptions, while astrology is offered as an example of an explanatory framework that lacks objective norms and therefore does not allow true essence ascriptions. I agree with these judgements (although it should be noted that one the three disciplines cited as positive cases is more dubious than the other two, being often based on extremely abstract esoteric principles that do not clearly relate to our ordinary experience; I am talking about economics of course). The view I sketched above predicts these judgements as well, insofar as explanations relative to each one of these frameworks are useful epistemically. I believe this is an advantage. It provides a basis for our judgements that is more solid, arguably, than the idea of the "objectivity" of the norms involved in an explanatory framework, which is simply (although not unreasonably) assumed by Sullivan. A further obvious difference between her account and the one proposed here is that the latter posits an explicit modal element in the definition of essence, while Sullivan's is, assuming explanation is an epistemic notion, a purely epistemic account. Perhaps Sullivan relies on the idea that *all* explanations in a certain domain have

[17] Haecceitas might also be seen as a limiting case, a sort of zero grade of essence, in which one might equally well say that, in a sense at least, Socrates has no essence. Furthermore, it might be correct in some contexts to indicate the haecceitas as a minimal essence and in other contexts to say that the individual has no essence.

to attribute a property to an object to provide a connection to necessity. I believe this connection to necessity is at risk of being too weak, but I will not discuss the matter here. The usefulness of the modal element in my account is to be discussed shortly, directly in connection to the epistemology of modality.

Leaving aside the comparison between Sullivan's proposal and the present one, we can note (again) that the connection between essence and explanation seems to be rather uncontroversial. Of course essential properties are *also* such that if something possesses one of them it could not exist while failing to possess it. But Fine's counterexamples to the modal view precisely show that this is not all there is to essence. And having some explanatory power seems to be an excellent candidate to supplement the modal profile.

C. Recognition/Individuation and D. Epistemology of Modality

Another traditional, arguably Aristotelian, idea about essences is their connection to the distinction between substances and qualities. Substances, in this philosophical sense of the term, are typically individuals; they fall under countable nouns. If you can talk about two dogs then they are distinct substances in this sense.

As I mentioned above, in the recent literature, the connection between this metaphysical role of essence and an epistemic role has been discussed by Lowe (2008; 2012). Recognizing an object, possibly presenting different properties, as something we encountered at a previous time in perception or thought, seems to require some grasp of what the object is, or at least some grasp of what it takes for an object of that kind to continue existing. I will get back to this point shortly.

Essences are supposed to help us "recognizing" things and kinds not only through time and space, but through the space of possible worlds as well. "Recognizing" is in scare quotes because it might suggest that I am committing to the view that we have a problem of identifying objects across possible worlds. I am friendly instead to the Kripkean view that this is a misleading way to put things. A false possibility claim, such as (suppose) "Socrates could have been a dog", it's still a claim about Socrates. It's not like we are talking about a dog in some possible world and falsely saying he is Socrates. But this is compatible with the claim that essences play a crucial role in allowing us to correctly judge which modal claims are true of an object. David Wiggins puts the point very clearly, I believe:

> The general idea [is] that the essential properties of a thing are part and parcel with what it takes for that very thing to be singled out from the rest of reality, and all of a piece with the necessary conditions for one who conceives the thing under a variety of counterfactual circumstances not to lose hold of that very thing while seeking to conceive it under this or that variation from its actual circumstances (Wiggins 2016: 165).

Here it is important though to stress the distinction between an essence and an essential property. An essential property in my sense is a necessary property with a particular interest. However, from the fact that something has a necessary property, however interesting, not much can be directly inferred about other properties the object possibly has.[18] One might think that an object possibly has all the properties that are not incompatible with any necessary property the same object

[18] It can be inferred of course that the object possibly has the same property, and any other entailed by it.

has, and so, if the necessary properties are those entailed by the essential properties, then we have a way to ascertain the truth of a possibility claim based on a complete list of the object's essential properties. Whatever the merits of this picture from a metaphysical point of view, however, it seems unlikely that we have the cognitive resources to use it, and if we do, it seems unlikely that we employ them in coming to know ordinary possibility claims. Consider a specific knife, call it Kenny, which is distinctively yellow. Could Kenny be red? I judge possible a situation in which the material object which is actually coincident with Kenny is painted red. But this would not be enough, by itself, to reasonably judge that Kenny could be red. After all, I judge possible a situation in which the material object which is actually coincident with Kenny is melted and reshaped as a fork. But that does not lead me to judge that Kenny could be a fork. In the former case, by contrast, my conception of Kenny, applied to the imaginary situation, yields a clear verdict. Spatiotemporal continuity and a continuity in function are sufficient to individuate the object. Something in that situation is Kenny, and it is red. It seems that I will, and should, be inclined to judge so just in case I would be able to recognize the object as being the same knife in case I actually decided to paint it red. My cognitive capacities are, as it were, prepared to track Kenny through various changes, and while the simulation of these changes perhaps requires an additional cognitive capacity for hypothetical thought, the ability to recognize the object seems to work in exactly the same way. Our cognitive capacities are limited. It wouldn't make sense to employ two different sets of criteria to judge that the object could change its colour and to judge that it is the same object, even though it changed its colour.

The foregoing should explain why I am discussing in a single section the roles of essence in connection to recognition and the epistemology of modality. It also should explain why, in my view, we need a modal-epistemic notion, which is what the notion of essence represents in my view. We may well be able to recognize objects through some of their accidental properties, and we often do. Lowe does not give us sufficient reasons to rule this possibility out. However, the properties involved should at least be modally robust enough to track the object through changes that are likely to occur in the actual circumstances. It is therefore natural that the same capacity may be employed, in hypothetical thought ("offline", to use Williamson's (Williamson 2007 expression), to reach modal judgements about that object. In theory, any set of necessary and sufficient conditions for something to be identical to a certain x will be able to play both roles. But, again, our cognitive capacities are limited. Other things being equal, we would like to have a simple and yet informative way to track x. We may name this way to track something our "conception" of that object. I use this term to mean whatever mechanisms guide our application of concepts. Conceptions can consist in explicit beliefs, implicit beliefs, or even non-propositional capacities.[19] A conception is adequate, roughly, when it yields mostly correct judgements.[20] Having adequate conceptions of the objects of our thought is a primary epistemic good.

[19] I take the notion from Millikan 2000. See also Wiggins 1980 (in particular "Preamble" and fn. 2 on p. 79, both also present in his 2001).

[20] Lowe (2012) uses the notion of adequacy of a concept in a similar way (944-47). The view I defend could also be formulated in terms of the idea defended in Vaidya 2010 that our grasp of essence is to be spelled out in terms of *understanding*.

What about necessary properties, one could ask? Surely my account, a possible objection would go, cannot give essences any role in our knowledge of (de re) necessities, since it defines "essential" in terms of necessary (plus something else). This would be too hasty. We must distinguish the epistemic and the metaphysical levels. It is possible that in the cognitive development of the individual the notion of necessity only comes after the notion of essentiality (like we may acquire the concept of a sibling after the concepts of sister and/or brother). We do not learn by definition, but rather by example. It is true, however, that the epistemology of modality in my view cannot at its core have the notion of essence. I believe there are several promising alternatives, but there is no space here to explore the issue further (see Mallozzi et al. 2021). I do accept that the view I am defending has the consequence that there must be some way of knowing modal truths independently of essences.

We now have sufficient material to build the inference to the best explanation in favour of EPD an ED. It is fairly uncontroversial, as I noted numerous times, that essence is supposed to play the epistemic roles I described. If the notion of essence, however, did not contain an epistemic element, explaining this phenomenon would be a rather difficult task, one which I think contemporary defenders of the notion of essence have not even attempted, by and large. We need to assume, or argue, that there are in the structure of the world junctures that individuate different objects, substances and kinds, and our cognitive capacities are capable to track these junctures. We must, in other words, be able to single out among the properties of an object those that are essential, and, moreover, be able to employ our knowledge of these essential properties in extending our knowledge through fruitful definitions, informative explanations, judgements of sameness and difference, and modal judgements. My view is instead, that we search for properties that can play these roles, and we call them essential.

I will close by considering an objection. The objection is closely related to the one I briefly discussed for the epistemology of modality, but it also concerns the metaphysics of modality. I already noted that my view must posit some way of knowing modal truths independently of essences. The view must also assume that modal truths are, so to speak, not grounded in essential truths. As for the case of epistemology, there are of course alternatives. The objection now is, though, that a certain form of essentialism offers a simpler explanation of the epistemic role of essences because it puts them at the center of both the epistemology and the metaphysics of modality. There are I believe, other views that allow for a similar unification (see e.g. Vetter 2015, 2016, 2020). What I am claiming however, is that the advantage of simplicity is not so clear in this case. The alignment of metaphysics and epistemology, so to speak, calls for a further explanation; one that, in the case of essence, really seems not be available outside an Aristotelian framework.

4. Conclusion

My aim in this paper was to clarify some connections which exist, and in my view are crucial, between essence and knowledge.

In the first section, I discussed Fine's objection against purely modal accounts of essence, which are, as far as I know, accepted by all theorist. I did not argue against modified or hybrid modal accounts, but I argue that Fine's counterexamples themselves seem to point toward a hybrid modal-epistemic account.

Having sketched such an account, I moved in the section 2 to argue that it is supported by an inference to the best explanation, where the explanandum is constituted by a number of epistemic roles essence plays.[21]

References

Adams, R.M. 1981, "Actualism and Thisness", *Synthese*, 49, 3-41.

De, M. 2020, "A Modal Account of Essence", *Metaphysics*, 3, 17-32.

Della Rocca, M. 1996, "Essentialism: Part I", *Philosophical Books*, 37, 1-13.

Fara, D.G. 2008, "Relative-Sameness Counterpart Theory", *The Review of Symbolic Logic,* 1, 167-89.

Graff Fara, D. 2012, "Possibility Relative to a Sortal", in Bennett, K. and Zimmerman, D. (eds.), *Oxford Studies in Metaphysics*, Vol. 7, Oxford: Oxford University Press, 3-40.

Fine, K. 1985, "Plantinga on the Reduction of Possibilist Discourse", in *Alvin Plantinga*, Dordrecht: Springer, 145-186, repr. in Fine 2005, 176-213.

Fine, K. 1994, "Essence and Modality", *Philosophical Perspectives*, 8, 1-16.

Fine, K. 2005, *Modality and Tense: Philosophical Papers*, Oxford: Oxford University Press.

Gorman, M. 2005, "The Essential and the Accidental", *Ratio,* 18, 276-289.

Hale, B. 2013, *Necessary Beings*, Oxford: Oxford University Press.

Kment, B. 2014, *Modality and Explanatory Reasoning*, Oxford: Oxford University Press.

Kung, J. 1977, "Aristotle on Essence and Explanation", *Philosophical Studies,* 31, 361-383.

Lewis, D. 1986, *On the Plurality of Worlds*, Oxford: Blackwell.

Livingstone-Banks, J. 2017, "In Defence of Modal Essentialism", *Inquiry*, 60, 8, 816–838.

Lowe, E.J. 2008, "Two Notions of Being: Entity and Essence", in Le Poidevin, R. (ed.), *Being: Developments in Contemporary Metaphysics*, Cambridge: Cambridge University Press, 23-48.

Lowe, E.J. 2012, "What is the Source of Our Knowledge of Modal Truths", *Mind*, 121, 919-950.

Mallozzi, A. 2021, "Superexplanations for Counterfactual Knowledge", *Philosophical Studies*, 178, 1315-1337.

Mallozzi, A., Wallner, M., & Vaidya, A.M. 2021, "The Epistemology of Modality", *The Stanford Encyclopedia of Philosophy* (Fall 2023 Edition), https://plato.stanford.edu/entries/modality-epistemology/

Millikan, R.G. 2000, *On Clear and Confused Ideas: An Essay About Substance Concepts*, Cambridge: Cambridge University Press.

Paul, L.A. 2004, "The Context of Essence", *Australasian Journal of Philosophy*, 82, 170-84.

Rayo, A. 2013, *The Construction of Logical Space*, Oxford: Oxford University Press.

Robinson, R. 1954, *Definition*, Oxford: Oxford University Press.

Sgaravatti, D. 2016, "Is Knowledge of Essence Required to Think about Something?", *Dialectica*, 70, 217-28.

[21] For comments on previous versions of this material I would like to warmly thank the audiences at the *Argumenta* workshop on the Epistemology of Metaphysics in Padova and at the Sixth Italian Conference in Analytic Metaphysics and Ontology in L'Aquila.

Sullivan, M. 2017, "Are There Essential Properties? No.", in Barnes, E. (ed.), *Current Controversies in Metaphysics,* Oxford: Routledge, 45-61.

Torza, A. 2015, "Speaking of Essence", *The Philosophical Quarterly*, 65, 211-231.

Vetter, B. 2015, *Potentiality: From Dispositions to Modality*, Oxford: Oxford University Press.

Vetter, B. 2016, "Williamsonian Modal Epistemology, Possibility-Based", *Canadian Journal of Philosophy*, 46, 766–795.

Vetter, B. 2020, "Perceiving Potentiality: A Metaphysics for Affordances", *Topoi*, 39, 1177–1191.

Wiggins, D. 1980, *Sameness and Substance*, Oxford: Basil Blackwell.

Wiggins, D. 2001, *Sameness and Substance Renewed*, Cambridge: Cambridge University Press.

Wiggins, D. 2016, *Continuants. Their Activity, Their Being and Their Identity. Twelve Essays*, Oxford: Oxford University Press.

Wildman, N. 2013, "Modality, Sparsity, and Essence", *The Philosophical Quarterly*, 63, 253, 760-782.

Wildman, N. 2016, "How (not) to Be a Modalist About Essence", in Jago, M. (ed.), *Reality making*, Oxford: Oxford University Press, 177-196.

Williamson, T. 2007, *The Philosophy of Philosophy*, Oxford: Blackwell Publishing.

Williamson, T. 2013, *Modal Logic as Metaphysics*, Oxford: Oxford University Press.

Advisory Board

Argumenta 10, 1 (2024)
Book Symposium

On Jessica Wilson's *Metaphysical Emergence* OUP 2021

The Journal of the Italian Society for Analytic Philosophy

Précis of *Metaphysical Emergence*

Jessica Wilson

University of Toronto

Introduction

The notion of metaphysical emergence is inspired by certain target cases, whereby—on the face of it, and in ways I'll expand on shortly—'higher-level' entities (objects, events, and the like) and features (properties, relations, behaviours, and the like) cotemporally materially depend on 'lower-level,' ultimately fundamental physical, micro-configurations and features; yet are also to some extent autonomous, ontologically and causally, from dependence base configurations and features. Relatedly, metaphysical emergence is inspired by a conception of natural and artifactual reality as manifesting a kind of leveled structure generally mirrored in the special sciences vis-à-vis the more fundamental physical sciences.

But what is metaphysical emergence, more precisely, and is there more than one variety of such emergence? And is there (really) any metaphysical emergence, in principle and moreover in fact?

In *Metaphysical Emergence* (2021), I aim to provide clear and systematic answers to these questions. I argue that there are two, and only two, forms of metaphysical emergence capable of accommodating the target cases—one 'Weak' (compatible with a physicalist world-view, given that the lower-level goings-on are physical), one 'Strong' (not so compatible). After defending the in-principle viability of each form of emergence, I consider whether complex systems, ordinary objects, consciousness, and free will are actually metaphysically emergent. I argue that some cases of each phenomenon are plausibly Weakly emergent, and I offer a new argument for there being free will of a Strongly emergent variety.

In what follows, I expand upon this rough overview, summarizing each chapter of *Metaphysical Emergence*. In the interest of efficiency, the presentation sometimes mixes prose with features more characteristic of a visually structured outline.[1]

Chapter 1: Key Issues and Questions

In Chapter 1, I begin by canvassing the prima facie motivations for thinking that there is metaphysical emergence (§1.1). To start, scientific orthodoxy takes for

[1] Please keep in mind that this précis necessarily elides what I take to be important dialectical qualifications and content. The book remains the official statement of my view(s).

Argumenta 10, 1 (2024): 191—224
ISSN 2465-2334

DOI 10.14275/2465-2334/202419.wil
First published: 30 November 2024

granted *Physical monism*, understood as contrasting with substance pluralist views such as Cartesian dualism or vitalism:

- *Physical monism*: The only matter or substance is physical matter or substance, such that the matter of a macro-entity at a time is inherited from some micro-configuration of ultimately physical constituents at that time.

Scientific orthodoxy also takes for granted that the features of macro-entities do not float entirely free of features of micro-configurations:

- *Cotemporal dependence*: The features of any macro-entity at a time or over a given temporal interval are at least in part a function of the features of the micro-configuration(s) which materially constitute the macro-entity at that time or during that temporal interval.

Reflecting these commitments, we can say that on the face of it, macro-entities and features *cotemporally materially depend* on micro-configurations and features.

What about autonomy? That macro-entities and features are to some extent both ontologically and causally autonomous from—that is, distinct from and distinctively efficacious as compared to—their underlying micro-configurations and features is motivated by a variety of considerations, including:

- *Distinctive taxonomies*: Special-science entities/features are classified under types which appear to be different from those classifying micro-configurations and features of such configurations (supports distinctness).
- *Distinctive causal laws*: Special-science entities enter into special-science laws describing features and behaviours of, including causal interactions involving, such entities—laws that, on the face of it, are different from those governing physical micro-configurations (supports distinctive efficacy, hence also distinctness).
- *Universal properties and behaviour*: Many special-science entities/features, including thermodynamic complex systems and features, are functionally and causally independent of underlying micro-configurations and features (supports distinctive efficacy, hence also distinctness).
- *Perceptual unity*: Macro-entities such as trees and tables perceptually appear to us as comparatively stable, unified entities, even though (as science tells us) they are materially constituted by complex, constantly changing micro-configurations (supports distinctness).
- *Compositional flexibility*: The existence and persistence of macro-entities/features typically appears to transcend that of underlying micro-configurations, in not depending on any *specific* micro-configuration(s) or features (supports distinctness).
- *Seemingly free will*: It introspectively seems as if we human persons are able to make free choices to produce (or intend to produce) certain effects, where this efficacy appears to be quite different from that associated with the (deterministically or indeterministically) lawfully governed micro-configurations and features upon which we and our mental states cotemporally materially depend (supports distinctive efficacy, hence also distinctness).

On the face of it, then, many macro-entities are *ontologically and causally autonomous* from—that is, distinct from and distinctively efficacious as compared to—

the micro-configurations and features upon which they cotemporally materially depend.

There is thus clear good reason to explore the notion of metaphysical emergence, understood as coupling *cotemporal material dependence* with *ontological and causal autonomy*.

Two key questions are immediately salient (§1.2):

1. Just what is metaphysical emergence, more precisely? How is it, exactly, that macro-entities and features can cotemporally materially depend on micro-configurations and features, while retaining some degree of ontological and causal autonomy? And is there more than one way in which this can be—is there more than one form of metaphysical emergence?
2. Is there actually any metaphysical emergence? To start: are there any insuperable problems with the notion(s) of metaphysical emergence, such that emergence is, at best, an epistemic or representational phenomenon? And supposing that a given variety of metaphysical emergence is in-principle viable, are there any actual cases of such emergence?

Indeed, in past decades there has been an explosion of philosophical and scientific interest in metaphysical emergence; yet the answers to the key questions have remained unclear. In re the first question: a bewildering variety of accounts of metaphysical emergence has been proposed, appealing to different, often incompatible interpretations of the core notions of dependence[2] and autonomy.[3]

[2] Candidate accounts of the dependence at issue in metaphysical emergence include mereological ('part-whole') determination (see Stephan 2002, Gillett 2002), causation or nomological connection (see Searle 1992, O'Connor and Wong 2005), functional realization (see Putnam 1967, Boyd 1980, Poland 1994, Antony and Levine 1997, Melnyk 2003), constitutive mechanism (see Craver 2001, Haug 2010, Gillett 2016), the determinable-determinate relation (see MacDonald and MacDonald 1986, Yablo 1992, Ehring 1996, Wilson 2009), inheritance of causal powers (see Kim 1992, Wilson 1999 and 2015, Shoemaker 2000/2001), and primitive 'Grounding' (see Schaffer 2009, Dasgupta 2014).

[3] Candidate accounts of the ontological and/or causal autonomy at issue in metaphysical emergence include nomological but not metaphysical supervenience (see Cleve 1990, Chalmers 1999, Seager 1999/2016, Noordhof 2010), non-fundamental novelty (of features, powers, laws, entities) (see Humphreys 1996, Wimsatt 1996, Crane 2001, Pereboom 2002, Megill 2013), fundamental novelty (of features, powers, forces/interactions, laws, entities) (see Mill 1843/1973, Alexander 1920, Broad 1925, Kim 1992, O'Connor 1994, Cunningham 2001, Wilson 2002 and 2015, Barnes 2012, Paolini Paoletti 2017), non-additivity/non-linearity (see again Mill, Alexander, and Broad, Newman 1996, Bedau 1997, Silberstein and McGeever 1999, Mitchell 2012), 'downward' causal efficacy (see Morgan 1923, Sperry 1986, Klee 1984, Thompson and Varela 2001, Searle 1992, Schroder 1998, Stephan 2002), multiple realizability/universality/compositional plasticity (see Putnam 1967, Fodor 1974, Boyd 1980, Klee 1984, LePore and Loewer 1989, Wimsatt 1996, Antony and Levine 1997, Aizawa and Gillett 2009, Morrison 2012), causal proportionality/difference-making/counterfactual considerations (see Yablo 1992, LePore and Loewer 1987 and 1989, Bennett 2003), elimination in degrees of freedom (see Wilson 2010 and Lamb 2015), sometimes associated with symmetry breaking (see Morrison 2012), and the holding of a proper subset relation between token powers (see Wilson 1999), sometimes cashed in terms of a proper parthood relation between properties and behaviours (see Shoemaker 2000/2001, Clapp 2001, Rueger and McGivern 2010). Also relevant here are 'epistemic criteria' accounts of ontological and/or causal autonomy, including in-principle failure of deducibility/predictability/explicability (see Broad 1925, Hempel and Oppenheim 1948, Klee 1984, LePore and Loewer 1989), pre-

Indeed, the extent of variability has led many to conclude that there is nothing systematic to be said or discovered about metaphysical emergence. The answer to the second key question has also remained unclear, owing to still-live concerns about whether the appearances of metaphysical emergence are genuine. Among these concerns are that metaphysical emergence is naturalistically unacceptable; that considerations of parsimony push against taking the appearances of metaphysical emergence ontologically seriously; that the notion of metaphysical emergence is either trivially fulfilled or trivially never fulfilled; and—perhaps most problematically—that metaphysically emergent entities or features, were they to exist, would give rise to problematic causal overdetermination of effects already produced by micro-configurations/features. Here the diversity of accounts of emergence again muddies the waters; for while some accounts have resources to respond to some concerns, the absence of any systematic treatment of metaphysical emergence renders it unclear whether the notion can survive all the various attacks.

In light of all this, the point and purpose of my book is to provide clear, compelling, and systematic answers to the two key questions of what, more precisely, metaphysical emergence is, and whether there actually is any such emergence. As discussed in §1.3, I go about this project as follows:

- In Ch. 2, I argue that there are two (and only two) schematic forms of metaphysical emergence which accommodate the target cases. One—'Weak emergence'—is compatible with physicalism, the view that all broadly scientific goings-on are completely metaphysically dependent on lower-level physical goings-on, on the assumption that the lower-level (ultimately compositionally basic) goings-on are physical; the other—'Strong emergence'—is incompatible with physicalism, on that assumption.[4]
- In Ch. 3, I consider and respond to a range of objections to the viability of Weak emergence.
- In Ch. 4, I consider and respond to a range of objections to the viability of Strong emergence.
- In Chs. 5–8, I consider whether complex systems, ordinary objects, consciousness, and free will, respectively, are actually either Weakly or Strongly metaphysically emergent. For each of these phenomena, I argue that some cases of the phenomenon are plausibly Weakly emergent. For most of these phenomena, I argue that existing arguments for the phenomenon's being Strongly emergent don't go through (though in some cases this remains a live empirical possibility). One exception: I argue that there is presently good reason to think that there is libertarian free will of a Strongly emergent variety.
- In Ch. 9, I finish up and point towards work remaining to be done.

dictability, but only by simulation (see Newman 1996, M. Bedau 1997), lack of conceptual or representational entailment (see Chalmers 1996, Van Gulick 2001), and the presence of theoretical/mathematical singularities (see Batterman 2002).

[4] As I observe, although the assumption that the base-level entities and features are physical or physically acceptable is typically operative in what follows, the schemas generalize to characterize emergence of two different varieties, whatever the precise ontological status of the base-level goings-on.

Besides motivating the book project and setting out the chapter structure, in Ch. 1 I expand on certain suppositions and operative notions informing my investigations (§1.4). In brief:

- *Certain core suppositions*. Notwithstanding their diversity, accounts of metaphysical emergence typically agree on the following theses, which are preserved on my account(s):
  - Metaphysical emergence couples cotemporal material dependence (hence, in particular, does not involve any new substance of the sort posited, e.g., by Cartesian dualists) and some degree of autonomy, where the autonomy at issue is causal as well as ontological.[5]
  - The metaphysical emergence of entities can be investigated by attention to the metaphysical emergence of features of the entities, with the supposition being that if some entity is metaphysically emergent, this is due to its having some characteristic metaphysically emergent feature (e.g., *being conscious*, *being in the basin of a strange attractor*) which can be the target of investigation.
  - Metaphysically emergent features 'minimally nomologically supervene' on base features, in that in every world (actual or hypothetical) with the same or relevantly similar laws of nature, the occurrence of an emergent feature *S* requires the occurrence of some or other base feature *P*, and in every such world, the occurrence of any such *P* will be accompanied by the occurrence of such an *S*.
- *The physical*. Discussions of metaphysical emergence as actually instantiated typically suppose that dependence base goings-on are ultimately physical. But what is it for some goings-on to be physical? The account operative here is that I advance in Wilson 2006, according to which the physical goings-on are those which are treated approximately accurately by present or future (in the limit of inquiry, ideal) physics, with the proviso that the physical goings-on are not fundamentally mental—that is, do not individually either have or bestow mentality. Not much turns on the specific details of the account of the physical, however; the main take-home point is that there is at least one physics-based account of the physical up to the task of characterizing the views at issue.
- *The individuation of levels*. It is common to think of metaphysical emergence in the target cases as going hand-in-hand with the suggestion that emergent entities and features are 'higher-level' with respect to the 'lower-level' goings-on upon which they depend.[6] But which entities and features should be taken to exist at a given level? An important constraint here is that levels (or the one level, if anti-realism or reductionism turns out to be correct) be individuated so as to include any combinations or configurations of entities and features to which the anti-realist or reductionist may reasonably

[5] Even with respect to these components there is some dispute; such variations, however, are either subsumable under the core understandings (as I argue is the case for diachronic accounts of metaphysical emergence; see also Wilson forthcoming*b*) or else are not to the point of accommodating the target phenomena (hence I put aside epiphenomenalist approaches to metaphysical emergence).

[6] Note that 'emergent' and 'higher-level' are not synonymous, however, since non-emergentist views (e.g., Cartesian dualism) also aim to accommodate leveled structure.

appeal. For example, if the basic physical entities are atoms and the basic physical relations include spatial relations and pairwise atomic bonding relations, then we should allow as existing, at the atomic level, not just small numbers of atoms standing in atomic relations, but also large numbers of atoms standing in highly complex atomic (including spatial) relations, constituting pluralities or aggregates of the sort that might, if reductionism is correct, be identical with a rock, a plant, or a person, at least at any given time.

Given this constraint, I offer two different approaches to answering the question of which combinations of entities and associated features should be taken to exist at a given level *L* of broadly scientific reality, beyond the entities and features typically taken, by lights of the associated science *S*, to be characteristic of *L*:

a. *The lightweight combination approach*. Here the individuation of levels proceeds by allowing that various ontologically 'lightweight' (including lower-level relational, mereological, and Boolean) combinations of the characteristic entities and features treated by a given science *S* and placed at a level *L* are also appropriately placed at *L*. For example, the goings-on at the atomic level would include not just atoms and pairwise atomic relations, but any configurations of atoms standing in atomic relations, any boolean combinations of such configurations, and so on.

b. *The 'law-consequence' approach*. Here the individuation of levels proceeds by allowing that any consequences of laws operating at a given level *L*, upon which those laws can operate (take as input), are also appropriately placed at *L*. For example, the goings-on at the atomic level would include any atomic configurations which the atomic laws are capable of taking as input (operating on).[7]

- *The fundamental*. Both physicalists and their Strong emergentist rivals suppose that there are fundamental physical goings-on; where they disagree is over whether there are any fundamental non-physical goings-on. But what is it for some goings-on to be fundamental (at a world, here and throughout)? There are three main approaches (see Tahko 2018 for discussion). On independence-based accounts, what makes it the case that some goings-on are fundamental is that those goings-on are (individually) metaphysically independent. On dependence-based accounts, this is a matter of the goings-on being part of a complete minimal dependence basis for everything that exists. And on primitivist accounts, this is a primitive matter, not metaphysically analyzable in any other terms. (Nota bene that it is not the fun-

[7] Note that on a law-consequence approach, only those consequences of laws at a given level *L* preserving the information required for the *L*-level laws to operate are placed at *L*. As such, a law-consequence approach does not automatically rule out Weak emergence, notwithstanding that Weak emergentists typically maintain that Weak emergents are in some sense metaphysical consequences of physical laws and conditions. For (as an empirical matter—so Weak emergentists argue) the metaphysical consequences associated with Weak emergents typically abstract away from certain lower-level details (e.g., quantum spin) such that were these input into the physical laws, the laws would not have all the information needed for them to operate.

damenta themselves, but what makes it the case that some goings-on are fundamental, that is on these accounts taken to be primitive). My own preference is for a primitivist account, as advanced in my 2014 and developed and defended in my forthcoming*a* and under contract. For the most part, which account of fundamentality is at issue won't matter for what follows, with one exception—namely, an independence-based conception on which individual fundamenta are metaphysically independent (see, e.g., Schaffer 2009, 373; Bennett 2017, 138) rules out fundamenta that are partly but not completely metaphysically dependent on other fundamenta, and so rules out a common understanding of Strongly emergent phenomena. That said, a collectivist variation on an independence-based account, on which the fundamental goings-on collectively do not depend on any other goings on, can accommodate Strong emergence, and so (versions of) all three approaches are suitable for present purposes.

- *Causes and powers*. The discussions to come often advert to causal relations and associated powers to produce effects. More specifically, the schemas for metaphysical emergence that I offer encode certain relations between powers of emergent and dependence base features. There are vast literatures on causation and powers, and on how these notions enter, metaphysically and modally, into the characterizations of entities and features. Fortunately, it is possible to remain almost entirely neutral as regards these more specific details.

  To start, the operative notion of 'power' in what follows is metaphysically highly neutral, following the presuppositions operative in my 2015*b*:

  > [T]alk of powers is simply shorthand for talk of what causal contributions possession of a given feature makes (or can make, relative to the same laws of nature) to an entity's bringing about an effect, when in certain circumstances. That features are associated with actual or potential causal contributions ('powers') reflects the uncontroversial fact that what entities do (can do, relative to the same laws of nature) depends on how they are (what features they have). So, for example, a magnet attracts nearby pins in virtue of being magnetic, not massy; a magnet falls to the ground when dropped in virtue of being massy, not magnetic. Moreover, a feature may contribute to diverse effects, given diverse circumstances of its occurrence (which circumstances may be internal or external to the entity possessing the feature). Anyone accepting that what effects a particular causes (can cause, relative to the same laws of nature) is in part a function of what features it has—effectively, all participants to the present debate—is in position to accept powers, in this shorthand, metaphysically neutral and nomologically motivated sense (354).

  The operative notion of causation is also highly metaphysically neutral. By way of proof of concept, I argue that even a contingentist categoricalist Humean—someone who maintains that causation is a matter of regularities, features have their powers contingently, and all features are ultimately categorical—can accept powers and the associated notion of causation in the neutral sense(s) here. For such a Humean, to say that an (ultimately categorical) feature has a certain power would be to say that, were a token of the feature to occur in certain circumstances, a certain (contingent) regularity would be instanced.

More generally, no controversial theses pertaining to the nature of powers, causation, properties, or laws are presupposed in the discussions or the schemas to follow. That said, I do suppose that we can make sense of physical causation. Some (e.g., Russell, 1912, and Field, 2003) claim that this is problematic; but first, the Russell/Field position is an outside view, as is clear from the usual formulations of physicalism as committed to Physical Causal Closure, according to which any physical effect has a sufficient purely physical cause; second, in any case, I argue that the Russell/Field line(s) of thought can be resisted.

- *Methodology*. Following most contemporary metaphysicians, I implement a broadly abductive methodology (i.e., 'inference to the best explanation', per Harman 1965 and Douven 2021), whereby candidate metaphysical accounts of a given phenomenon are assessed by attention to how well they do, overall, at satisfying various theoretical desiderata. To be sure, there is variation in exactly which theoretical desiderata are operative as well as in how these desiderata, which may push in different directions, should be weighted. As I discuss in my 2011, 2016*c*, and 2016*b*, this variation is unsurprising, given the wide purview of metaphysical investigations and our present distance from the end of inquiry. Even in the absence of complete consensus regarding methodological standards, progress can be made, so long as one is suitably explicit about which theoretical desiderata are primarily guiding one's investigations. Two methodological desiderata which I take to be especially important in my theorizing are as follows:
    1. *Criterion of Appropriate Accommodation*: An adequate account of metaphysical emergence should make natural (straightforward, default) and realistic sense of the appearances of metaphysical emergence, in the absence of reasons to think that this cannot be done. Hence while I take it to be part of my burden to show that various purported problems with metaphysical emergence can be addressed, I do not take it to be part of my burden to show that no deflationary (anti-realist or reductionist) account of the appearances of metaphysical emergence is viable. My ultimate goal is not to knock the anti-realist or reductionist off their horse, but to show the metaphysical emergentist who aims to accommodate the appearances at realistic face value how to stay on their own horse. I hope that those with different methodological sensibilities will nonetheless find the ensuing discussion useful, at least as revealing the extent to which the heavy weighting of parsimony considerations, as opposed to any specific problem with the notion of metaphysical emergence itself, may be playing a role in deflationary accounts of such emergence.
    2. *Criterion of Illuminating Accommodation*: An adequate account of metaphysical emergence should provide an illuminating basis for accommodating the appearances of metaphysical emergence in natural (straightforward, default) fashion. Hence it isn't enough to simply stipulate, or take it to be brute or primitive, that some goings-on are both cotemporally materially dependent and suitably autonomous; what is desired is one or more intelligible, explanatory account(s) of how there can be metaphysical emergence in this sense.

Chapter 2: "Two Schemas for Metaphysical Emergence"

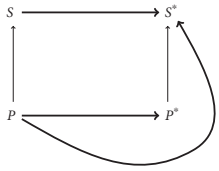
In Chapter 2, I motivate my two schemas for metaphysical emergence by attention to what is seen by many as the most pressing challenge to taking the appearances of metaphysical emergence as genuine—namely, the problem of higher-level causation, made salient by Kim in his 1989, 1993a, 1998, and elsewhere. I argue, following discussions in Wilson 1999, 2001, 2011*b*, and elsewhere, that there are two and only two strategies of response to this problem that make sense of seemingly higher-level entities and features' being metaphysically emergent as above. One strategy provides a schematic basis for 'Weak' (physically acceptable) emergence; the other provides a schematic basis for 'Strong' (physically unacceptable) emergence.[8] For each of these strategies and associated schemas, I show that a representative range of seemingly diverse accounts of metaphysical emergence are plausibly seen as satisfying the conditions in one or the other schema, and thus are more unified than they appear.

I start by presenting Kim's problem of higher-level causation (§2.1). The general concern is that any purported effects of higher-level features are already produced by the lower-level features upon which they minimally nomologically supervene, such that the metaphysical emergentist is committed to such effects' being problematically causally overdetermined—that is, problematically caused twice over. More specifically, the problem is usefully seen as involving the following six premises:

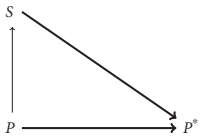
1. *Dependence*. Special science features cotemporally materially depend on lower-level physical features ('base features').
2. *Reality*. Both special science features and their base features are real.
3. *Efficacy*. Special science features are causally efficacious.
4. *Distinctness*. Special science features are distinct from their base features.
5. *Physical Causal Closure*. Every lower-level physical effect has a purely lower-level physical cause.
6. *Non-overdetermination*. With the exception of cases of the double-rock-throw variety, effects are not causally overdetermined by distinct individually sufficient cotemporal causes.

There are two cases to consider, reflecting two sorts of effect. In Kim's presentation, *S* is a mental state (e.g., *being thirsty*); *P* is a base state upon which *S* depends; and *S* is taken to cause either another mental state *S** (e.g., a desire to quench one's thirst) or a base state *P** (e.g., a physical reaching for a glass of water). But the challenge more generally concerns how any real, distinct, dependent higher-level feature might be unproblematically efficacious. The two cases are as follows (bold lines = causation, thin lines = cotemporal material dependence):

[8] Again, the schemas more generally operate to characterize emergence of two different varieties, whatever the precise ontological status of the base-level goings-on.



Case 1 of the problem of higher-level causation: *S* causes *S**



Case 2 of the problem of higher-level causation: *S* causes *P**

Kim rejects *Distinctness*, favouring reductive physicalism. But more generally (see Wilson 2015), rejection of each premise is associated with certain prominent views. To start:

1. *Substance dualism*. Deny *Dependence*: avoid overdetermination by denying that *S* and *S** cotemporally materially depend on base features *P* and *P**, respectively.
2. *Eliminativism*. Deny *Reality*: avoid overdetermination by denying that *S* and *S** are real.
3. *Epiphenomenalism*. Deny *Efficacy*: avoid overdetermination by denying that *S* is efficacious.
4. *Reductive physicalism*. Deny *Distinctness*: avoid overdetermination by identifying *S* with *P*.

These strategies avoid overdetermination, but don't make sense of higher-level features as metaphysically emergent—that is, as real, dependent, distinct, and distinctively efficacious.

There are, however, two strategies of response to Kim which do accommodate metaphysical emergence:

5. *Strong emergentism*. Deny *Physical Causal Closure*: avoid overdetermination by denying that every lower-level physical effect has a purely lower-level physical cause. This is the strategy encoded in 'British Emergentist' accounts.
6. *Weak emergentism*. Deny *Non-overdetermination*: allow that effects caused by *S* are also caused by *P*, but maintain that the overdetermination here is of an unproblematic *non*-double-rock-throw variety. This is the strategy encoded in non-reductive physicalist accounts (e.g., functional realization, determinable-determinate, and constitutive mechanism accounts).

As I argue in the next two sections, these two strategies and associated positions are perspicuously seen as motivated by two conditions on the powers of a given special-science feature, where satisfaction of one or other condition provides a prima facie plausible and principled (i.e., appropriate and illuminating) basis for taking the feature to be emergent, in ways that standard proponents of the strategy/position would endorse. In each of these sections, treating Strong emer-

gence and Weak emergence, respectively, I start by motivating the associated condition on powers by attention to standard versions of the position; I then show how satisfaction of the condition dovetails with the associated strategy for responding to the problem of higher-level causation; I then provide prima facie reasons for thinking that satisfaction of the condition provides an appropriate and illuminating basis for taking special-science features to be both cotemporally materially dependent and ontologically and causally autonomous; finally, I use the condition to formulate the associated schema for metaphysical emergence.

The Schema for Strong Emergence

I start with the Strong emergentist strategy, as implemented most saliently by British emergentists (§2.2). The conception of higher-level efficacy at issue in Strong emergentism is, as above, one which denies *Physical Causal Closure*, and is correspondingly incompatible with physicalism. And while different accounts of Strong emergentism emphasize different aspects of this distinctive efficacy as located in fundamentally novel features, laws, effects, forces, or interactions, core and common to these accounts is that Strongly emergent features have fundamentally novel powers—powers to produce effects entailing the violation, in particular, of *Physical Causal Closure*, as per the following condition:

> *New Power Condition*: Token feature *S* has, on a given occasion, at least one token power not identical with any token power of the token feature *P* upon which *S* cotemporally materially depends, on that occasion.

This is true, to start, on British emergentism, as endorsed most systematically by Mill (1843/1973), Alexander (1920), Lewes (1875), and Broad (1925). Hence in his classic survey, McLaughlin (1992) describes British emergentism as

> [T]he doctrine that there are fundamental powers to influence motion associated with types of structures of particles that compose certain chemical, biological, and psychological kinds" (52), where the powers at issue are typically taken to be "powers to generate fundamental forces not generated by any pairs of elementary particles. (71)

Contemporary accounts of Strong emergence also typically agree in taking emergent features to have or bestow fundamentally novel powers, not had (or had only in derivative fashion) by base features or associated microconfigurations. For example, O'Connor and Wong (2005) characterize emergent features as "fundamentally new", not just in being (perhaps epiphenomenally) different, but more specifically in having fundamentally novel causal capacities:

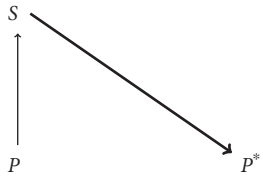
> [A]s a fundamentally new kind of feature, [an emergent feature] will confer causal capacities on the object that go beyond the summation of capacities directly conferred by the objects microstructure. (665)

(See also, e.g., Silberstein and McGeever 1999, Wilson 1999, and Van Gulick 2001.)

Given that higher-level feature *S* has a (fundamentally novel) power to cause a given effect—a power that its dependence base feature *P* does not

have—the Strong emergentist's responses to Kim's cases can be represented as follows:

The Strong emergentist's response to case 1



The Strong emergentist's response to case 2

Prima facie, satisfaction of the New Power Condition by a special-science feature *S* which cotemporally materially depends on a base feature *P* provides an appropriate and illuminating basis for avoiding overdetermination while guaranteeing that *S* is both ontologically and causally autonomous with respect to *P*. We have thus arrived at our first schema for metaphysical emergence:

> *Strong Emergence*: What it is for token feature *S* to be Strongly metaphysically emergent from token feature *P* on a given occasion is for it to be the case, on that occasion, (i) that *S* cotemporally materially depends on *P*, and (ii) that *S* has at least one token power not identical with any token power of *P*.

Here the locution 'what it is for' is intended to flag that Strong Emergence provides a schematic metaphysical basis for a given case of such emergence, encoding what is core and crucial to that notion. Some clarifications:

- The notion of 'power' operative in the schema is metaphysically highly neutral.
- The base feature *P* in the schema is a feature of a micro-configuration (not of an individual component of the configuration), and the conditions should be understood accordingly.
- The first condition encodes substance monism and minimal nomological supervenience.
- The second condition ensures ontological and causal autonomy (distinctness and distinctive efficacy). For Strong emergence, distinctive efficacy involves the higher-level feature's having *a new power*—a power not had, or not had in same way, by the base feature:
  - Note that the novel token power is fundamentally novel, since non-fundamentally novel powers (powers had just in virtue of aggregation) are had by base feature *P*.
  - In having a novel token power, *S* can cause an effect that *P* can't cause, or that *P* can't cause in the same (non-derivative) way as *S*;

hence *S* is causally autonomous—that is, distinctively efficacious—with respect to *P*.

– That a Strong emergent has a token power not had by its base feature *P* entails that *S* is distinct from *P*, by Leibniz's Law.

- The schema is relativized to occasions (times or temporal intervals), but it would be reasonable to suppose that it suffices for the Strong emergence of *S*, simpliciter, that the condition is ever satisfied, and to suppose that it suffices for the Strong emergence of the feature type (of which *S* is a token), simpliciter, that any token feature *S* on any occasion satisfies (or would satisfy) the condition.

The Schema for Weak Emergence

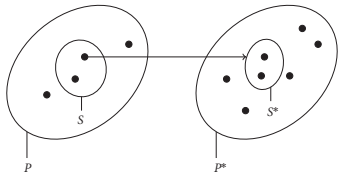
I focus next on the Weak emergentist strategy, as implemented most saliently by non-reductive physicalists (§2.3). Like Strong emergentists, non-reductive physicalists maintain that (some) higher-level features are real, cotemporally materially dependent, distinct, and distinctively efficacious with respect to their base features. But as physicalists, their response to the problem of higher-level causation cannot entail the rejection of *Physical Causal Closure*, which is core to the physicalist view that the physical goings-on are an existential and causal basis for all other broadly scientific phenomena. Rather, non-reductive physicalists reject *Non-overdetermination*, maintaining that distinct special science and base features can each be sufficient causes of a single effect, in virtue of standing in a relation that, while not identity, is intimate enough both to avoid overdetermination of the problematic (since implausible, for the cases at issue) double-rock-throw variety and to retain compatibility with *Physical Causal Closure*, hence with physicalism.

Non-reductive physicalists posit a variety of relations as showing how it can be that a higher-level feature can be completely metaphysically dependent on, yet distinct and distinctively efficacious with respect to, lower-level dependence base features. These include functional realization (Putnam 1967, Fodor 1974, Papineau 1993, Antony and Levine 1997, Melnyk 2003, Witmer 2003, Polger 2007, Yates 2012), the determinable-determinate relation (MacDonald and MacDonald 1986, Yablo 1992, Wilson 1999 and 2009), constitutional mechanism (Cummins 1975, Craver 2001, Haug 2010), mereological realization (Shoemaker 2000/2001, Clapp 2001, Rueger and McGivern 2010), and many others. Though there are interesting differences between these accounts of non-reductive realization, I argue that they have in common that each is plausibly such as to satisfy the following condition on token powers of realized and realizing features:

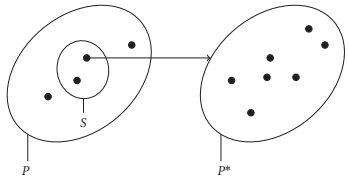
Proper Subset of Powers Condition: Token feature *S* has, on a given occasion, a non-empty proper subset of the token powers of the token feature *P* on which *S* cotemporally materially depends, on that occasion.[9]

Representing the features at issue as having overlapping sets of powers, with each power represented as a dot, the non-reductive physicalist's responses to Kim's cases are as follows:

[9] The requirement that the proper subset of powers be non-empty reflects the rejection of epiphenomenal features as metaphysically emergent, in the relevant sense.



The Weak emergentist's response to case 1



The Weak emergentist's response to case 2

Prima facie, satisfaction of the Proper Subset of Powers Condition by a special-science feature *S* which cotemporally materially depends on a base feature *P* provides an appropriate and illuminating basis for avoiding overdetermination while guaranteeing that *S* is both ontologically and causally autonomous with respect to *P*. We have thus arrived at our second schema for metaphysical emergence:

> *Weak Emergence*: What it is for token feature *S* to be Weakly metaphysically emergent from token feature *P* on a given occasion is for it to be the case, on that occasion, (i) that *S* cotemporally materially depends on *P*, and (ii) that *S* has a non-empty proper subset of the token powers had by *P*.

Here again, the locution 'what it is for' is intended to flag that Weak Emergence provides a schematic metaphysical basis for a given case of such emergence, encoding what is core and crucial to that notion. Some clarifications:

- The notion of 'power' operative in the schema is metaphysically highly neutral, as is the supposition that one can make sense of the identity (non-identity) of powers (see my reply to Bennett for further discussion).
- The base feature *P* in the schema is a feature of a micro-configuration (not of an individual component of the configuration), and the conditions should be understood accordingly.
- The first condition encodes substance monism and minimal nomological supervenience.
- The second condition ensures ontological and causal autonomy (distinctness and distinctive efficacy. For Weak emergence, distinctive efficacy involves the higher-level feature's having *strictly fewer* powers than are had by the base feature, and hence having a distinctive power profile:
  - Here the response to Kim proceeds by maintaining—contra what Kim assumes—that distinctive efficacy of a higher-level feature does not require that it have a new power.
  - It suffices for distinctive efficacy that the feature have a distinctive power profile, tracking difference-making considerations (if my thirst had been differently physically realized, I would still have reached for

the Fresca), or comparatively abstract levels of causal or nomological grain.

- That a Weak emergent has a distinctive power profile entails that it is distinct from its base feature, by Leibniz's Law.
- Again, the schema is relativized to occasions (times or temporal intervals), but it is reasonable to suppose that (given that S's type is not Strongly emergent) it suffices for the Weak emergence of S, simpliciter, that the condition is ever satisfied, and to suppose that it suffices for the Weak emergence of the feature type (of which S is a token), simpliciter, that any token feature S on any occasion satisfies (or would satisfy) the condition.

I close the chapter by observing that attention to the problem of higher-level causation makes clear the limited ways in which a cotemporally materially dependent higher-level feature can be causally, hence ontologically, autonomous with respect to its base feature, as the operative conception of metaphysical emergence requires (§2.4). First, the feature may have *more* powers than its base feature, as in Strong emergence;[10] second, the feature may have fewer powers than its base feature, as in Weak emergence. In terms of effects: the higher-level feature may be distinctively efficacious in potentially contributing to causing at least one different effect than its base feature (Strong emergence), or it may be distinctively efficacious in potentially contributing to fewer effects than its base feature (Weak emergence). Since complete coincidence of token powers doesn't make room for causal autonomy (distinctive efficacy), these routes to metaphysical emergence exhaust the available options.

I conclude that satisfaction of the conditions in either schema is, as I put it, 'core and crucial' to metaphysical emergence of the sort relevant to realistically vindicating the seeming appearances of emergence as pertaining to special-scientific and artifactual entities and features. Modulo the supposition that the schemas are sensibly filled in, the results of this chapter can be seen as providing prima facie reason to think that the conditions in the schemas are both necessary and sufficient for (appropriate and illuminating accommodation of) metaphysical emergence of both physically acceptable and physically unacceptable varieties—a bold claim, but one that, as I argue in ensuing chapters, is surprisingly robust.

Chapter 3: "The Viability of Weak Emergence"

In Chapter 3, I consider and respond to a representative range of objections to the viability of Weak emergence, understood as per the associated schema:

> *Weak Emergence:* What it is for token feature S to be Weakly metaphysically emergent from token feature P on a given occasion is for it to be the case, on that occasion, (i) that S cotemporally materially depends on P, and (ii) that S has a non-empty proper subset of the token powers had by P.

These objections fall into four main categories, according to which satisfaction of the conditions in Weak Emergence is …

[10] By 'more' I just mean that a Strong emergent must have at least one power not had by the base feature; pace Ney (2022), I do not suppose (and nor does satisfaction of the conditions in the schema require) that a Strong emergent have all the powers of the base feature, and then some.

- compatible with anti-realism about higher-level features (§3.1);
- compatible with reductionism about higher-level features (§3.2);
- compatible with the emergent feature's being physically unacceptable (§3.3); or
- not necessary for metaphysical emergence of a physically acceptable variety (§3.4).

The primary focus of many of the objections is on condition (ii) in the schema—i.e., the Proper Subset of Powers Condition. These diverse challenges can, I argue, be answered. Each of these objections admits of at least one response that could be endorsed by any proponent of Weak emergence, whatever their preferred implementation of the schema. Upon occasion, however, I offer certain attractive responses appealing to either a determinable-based account of Weak emergence (per my 1999 and 2009, developing the proposals in MacDonald and MacDonald 1986 and Yablo 1992), or an account of Weak emergence as involving an elimination in degrees of freedom (per my 2010, developing the proposal in Batterman 1998 and elsewhere).

Here, by way of partial illustration, I sketch certain representative lines of response to each of the four categories of concern.

According to the first concern (see, e.g., Heil 2003, Ney 2010, and Morris 2018), "nothing has been said to rule out" (as Ney puts it) an abstractionist or pragmatist line on seeming satisfaction of the Proper Subset of Powers Condition. I grant that this is the case, but deny that the viability of Weak emergence hinges on accomplishing such a 'ruling out.' Given the many prima facie reasons for thinking that there is metaphysical emergence, the burden is on the anti-realist to provide reasons for not taking the appearances at face value; but so far anti-realists have not provided any such good reason—in particular, as telling against a Weak emergentist treatment of the appearances. For example, Heil suggests that predicates such as 'red' should be understood not as referring to higher-level features, but rather as tracking inexact similarities between lower-level features, especially in light of Kim-style overdetermination concerns; but even granting that the predicates at issue are tracking inexact similarities among lower-level features, this would not show that the higher-level features did not exist, unless it was antecedently clear that the inexact similarities at issue were not themselves higher-level, which it isn't; and as above, the Weak emergentist has a response to Kim's overdetermination concerns, which makes clear how Weak emergents can be causally efficacious in spite of not having any new powers, in virtue of having a distinctive power profile, tracking difference-making considerations and comparatively abstract levels of causal grain.

According to the second concern, even granting that feature *S*'s satisfying the conditions in Weak emergence physical feature *P* ensures that *S* is real and distinct from *P*, this much is compatible with *S*'s being ontologically reducible to—that is, identical with—some *other* lower-level physically acceptable feature P' (see Yates 2012, 6, for discussion of the general concern). There are diverse reductive strategies here, according to which *S* is reducible to …

- a conjunct of a lower-level conjunction (§3.2.1);
- a disjunction of lower-level disjuncts (§3.2.2); or
- a metaphysical consequence of lower-level laws (§3.2.3).

To each strategy I offer one or more responses that any Weak emergentist might accept. In the case of the first strategy (see Shoemaker 2000/2001 for discussion), one might stipulatively rule out conjunctive realization (as Shoemaker does), or implement Baysan's suggestion that, on the supposition that conjunct features are more fundamental than associated conjunctive features, a conjunct feature *S* would not be appropriately taken to satisfy the relevant condition on dependence in the schema for Weak emergence. I additionally note that an appeal to a determinable-based implementation of Weak emergence will suffice to non-stipulatively rule out conjunctive realization, since it is definitive of the determinable/determinate relation that it is not properly metaphysically characterized in terms of anything like the conjunct/conjunction (or relatedly, genus/species) relations (see Wilson 2022/2017 for discussion). In the case of the second 'disjunctive' strategy (see, e.g., Fodor 1987, Jaworski 2002, and Dosanjh 2014 and 2019), I argue that on the usual understanding according to which what it is for a disjunctive type to be tokened on a given occasion is for one of the disjunct types to be tokened on that occasion, the disjunctive strategy is incompatible with satisfaction of the Proper Subset of Powers Condition. And in the case of the third strategy (see, e.g., Nagel 1961, Klee 1984, Kim 2010, and Morris 2018), I observe (see note 7 of this précis) that a proper understanding of how laws enter into the individuation of levels enables the Weak emergentist to maintain that, notwithstanding that special scientific goings-on are, on their view, metaphysical consequences of lower-level physical goings-on, it does not follow that the former are identical with any of the latter, since the former do not contain all the information needed for the lower-level physical laws to operate. I additionally note that a DOF-based implementation of Weak emergence develops this idea, in that on this implementation special-science goings-on may be metaphysical (and even deductive, so to speak) consequences of lower-level physical goings-on, yet be distinct from any lower-level physical goings-on, in failing to have all the DOF that are needed for the lower-level physical laws to operate (as first discussed in Wilson 2010).

According to the third line of concern, that a feature *S* satisfies the conditions in Weak emergence vis-à-vis a given physical feature *P* is compatible with *S*'s being physically unacceptable. Again, there are several variations of the theme of the concern, according to which satisfaction of the Proper Subset Condition on Powers, in particular, is compatible with *S*'s being 'over and above' *P* in virtue of …

- *S*'s having a non-causal quiddity (§3.3.1);
- *S*'s having a phenomenal aspect (§3.3.2);
- S's failing to be entailed by *P* (§3.3.3);
- *S*'s having a fundamentally mental power (§3.3.4); or
- *S*'s being associated with physically unacceptable constraints (§3.3.5).

In re non-causal quiddities (per Melnyk 2006, Morris 2018), I argue that the Weak emergentist can reasonably maintain that whether *S* and/or *P* have quiddities, shared or not, is irrelevant to whether *S* is physically acceptable, since the occurrence of scientific features, and any truths about such features, does not depend on or otherwise track whether such features have quiddities, much less track how the noncausal quiddities of seemingly distinct features are related; and similarly for artifactual features satisfying the conditions in Weak Emer-

gence. In re phenomenal aspects (per, e.g., Walter 2010), I argue that the common supposition that phenomenal aspects (of mental features, in particular) cannot be characterized in terms of causal roles or associated powers is incorrect; rather, as per what I call the 'Phenomenal Incorporation Thesis,' phenomenal aspects of mental features are fully incorporated into the powers of these features (compatible with powers' being contingently associated with features, relative to a given set of laws), reflecting that differences in phenomenality give rise to causal differences. In re a supposed failure of *S* to be entailed or necessitated by *P* (per Melnyk 2006, McLaughlin 2007), I observe (among other responses) that the cases usually offered as showing that *S* would be 'over and above' *P* in not even being nomologically entailed or necessitated by *P* fail to take the cotemporal material dependence condition in Weak emergence into account. In re fundamentally mental powers (per Baltimore 2013), I observe that while the Proper Subset Condition on Powers itself does not rule out *P*, hence *S*, from having fundamentally mental powers, the operative 'no fundamental mentality' account of the physical (per my 2006) does so. Finally, in re physically unacceptable constraints (per Melnyk 2006), I grant that when the Proper Subset Condition is satisfied as a result of constraints being imposed on lower-level goings-on, the constraints themselves need to be physically acceptable, and that it might be worth adding this requirement to the schema for Weak emergence (as I explicitly do in my DOF-based implementation of Weak emergence).

According to the fourth line of concern, satisfaction of the conditions in Weak emergence is not necessary for physically acceptable emergence; rather, one or other account in terms of token identity (per Davidson 1970, Macdonald and Macdonald 1995, Ehring 2003, and Robb 1997) (§3.4.1), constitutive mechanism (per Gillett 2002*a*, 2002*b*, 2016) (§3.4.2), constitution (per Pereboom 2002) (§3.4.3), or primitive Grounding (per Schaffer 2009, Rosen 2010, and Dasgupta 2014) (§3.4.4) will do the job. Considerations of space prevent my discussing these alternatives in any detail here; I can say, however, that a common theme is that the views at issue either fail to establish the ontological and causal autonomy of higher-level features, and so are not really accounts of physically acceptable emergence; or else are plausibly seen as imposing the Proper Subset of Powers Condition, and so are not really competitors to my view.

Chapter 4: "The Viability of Strong Emergence"

In Chapter 4, I consider and respond to a representative range of objections to Strong emergence, understood as per the associated schema:

> *Strong Emergence:* What it is for token feature S to be Strongly metaphysically emergent from token feature P on a given occasion is for it to be the case, on that occasion, (i) that S cotemporally materially depends on P, and (ii) that S has at least one token power not identical with any token power of P.

These objections fall into four main categories, according to which satisfaction of the conditions in Strong Emergence is …

- incompatible with scientific theory or practice (§4.1);
- impossible, since any purportedly novel powers of Strongly emergent features are inherited by (or 'collapse' into) base features (§4.2);
- compatible with physical acceptability (§4.3); or

- not necessary for emergence of a physically unacceptable variety (§4.4).

Here again, I argue that these diverse challenges can be answered. And here again, each objection admits of at least one response that any proponent of Strong emergence could endorse, whatever their preferred implementation of the schema. Upon occasion, however, responses draw on features of my preferred 'fundamental interaction-relative' account of Strong emergence (as per my 2002), according to which a Strongly emergent entity (feature) has at least one power that is grounded, at least in part, in a novel (nonphysical) fundamental interaction.

Here, by way of partial illustration, I sketch certain representative lines of response to each of the four categories of concern.

According to the first commonly voiced concern, Strong emergence is naturalistically or scientifically unacceptable. In response, I start by observing, following McLaughlin 1992, that Strong emergence would not be incompatible with laws such as $F = ma$ or Schrödinger's equation, but would rather just involve adding another force or energy to the mix of those input into these laws of nature. I moreover argue, following Wilson 2002, that reflecting that scientific practice suggests that powers are plausibly grounded, one way or another, in fundamental forces or interactions (as when the power of a magnet to attract a pin is grounded in the electromagnetic interaction), naturalistic good sense can be made of the Strong emergentist posit of fundamentally novel powers, as reflecting novel fundamental interactions that come into play only at certain levels of compositional complexity, such that Strong emergentism "is committed to there being at least one other fundamental force beyond those fundamental forces currently posited" (74). Indeed, the case of the weak nuclear interaction, posited in response to apparent conservation law violations in beta decay, supports the naturalistic/scientific respectability of Strong emergence: since a nucleus is a complex entity, evidently scientists have no problem with positing fundamental configurational interactions and associated powers. Similar experiments could provide an empirical basis for Strong emergence, in principle.

Finally, I observe that claims that there is "not a scintilla of evidence" in favor of there being Strongly emergent features (McLaughlin 1992; see also Ladyman and Ross 2007) are overstated, especially in light of the result forthcoming in Ch. 8 (see also my response to McLaughlin, this volume).

According to the second concern, Strong emergence is impossible, due to the base feature's inheriting any purportedly novel power, as per what Taylor (2015) evocatively calls the 'collapse' objection (see Cleve 1990, Kim 1999, O'Connor 1994, Wilson 2002, Francescotti 2007, Howell 2009, Taylor 2015, and Carruth 2018). Drawing on Baysan and Wilson 2017, I offer four strategies for avoiding collapse. Three might be implemented by any account of Strong emergence; these involve (i) distinguishing between direct and indirect having of powers, (ii) distinguishing between lightweight and heavyweight dispositions, and (iii) taking Strongly emergent features to be 'new object entailing,' in ways that block lower-level inheritance of powers. The fourth strategy draws on my fundamental interaction-relative account of Strong emergence. On this account, to start, powers are grounded (I make some specific suggestions as to how) in fundamental interactions: as above, magnets have the power to attract pins in virtue of the electromagnetic, not the gravitational, interaction; and so on. One can understand the New Power Condition accordingly. Relative to the set of

purely physical fundamental interactions, a cotemporally materially dependent feature S can have a fundamentally novel power p, as per the schema for Strong emergence; relative to the set of any and all fundamental interactions, p will be inherited by the lower-level physical features P upon which S cotemporally materially depends.

According to the third concern (due to Yates 2016), satisfaction by a feature S of the conditions in Strong emergence is compatible with S's being physically realized, hence physically acceptable. By way of illustrative motivation Yates argues that the molecular geometry G of a water molecule is a mathematically specified, physically realized feature which bestows certain powers upon its bearer—in particular, those, including hydrogen bonding in water, associated with the molecule's dipole moment—not had/bestowed by G's realizers. Here I argue that Yate's reasons for thinking that the powers had by G are not had by the base feature F that 'qualitatively' realizes G on a given occasion do not go through. In particular, he supposes that if such power inheritance were in place, references to G could be eliminated in broadly deductive explanations of the dipole moment and associated powers, yet such references can't be eliminated; but (I observe) nothing in physicalism or in the physicalist supposition that higher-level features inherit their powers from physical base features requires that elements of higher-level explanations, deductive or otherwise, be 'dischargeable' in terms referring only to lower-level physical goings-on. Moreover, Yates maintains that G can be deduced from lower-level physical goings-on, as an "intermediary step"; but then why think that the need to appeal to G indicates that G has new powers, as opposed to thinking that this need simply reflects that the explanation of the existence and powers of the dipole moment has to proceed in steps, compatible with the physicalist assumption that any powers of deducible features such as G are inherited? More generally, I argue that Yates does not establish that the relation of qualitative realization is (like functional and other forms of realization) also a relation of causal power bestowal.

According to the fourth concern, satisfaction of the conditions in Strong Emergence is not necessary for physically unacceptable emergence. There are four main alternative approaches on offer, in terms of …

- epiphenomenalism (§4.4.1);
- supervenience (§4.4.2);
- primitivism (§4.4.3); or
- epistemic criteria (§4.4.4).

In response, I provide reasons for thinking that each of these alternative approaches to physically unacceptable emergence is unsatisfactory. Again, considerations of space prevent my discussing these alternatives in any detail; here I briefly register some lines of argument.

In re epiphenomenalism (per, e.g., Chalmers 1996): the motivations for making room for an epiphenomenalist conception of emergence rest on there being phenomenal properties, along with the assumption that such properties cannot be characterized in terms of causal roles or associated powers; but as per the 'Phenomenal Incorporation Thesis,' discussed above, this is incorrect. In re supervenience (per, e.g., Chalmers 2006, Witmer 2001): I first canvass reasons for thinking that Strong emergence cannot be characterized as involving nomological but not metaphysical necessity of emergent on base features, since (per sce-

narios highlighted in, e.g., Horgan 1993 and Wilson 2005) Strongly emergent features might supervene with metaphysical necessity on base features. I then offer several responses to Howell's 2009 argument that such scenarios pose no threat to a supervenience-based characterization of such emergence, since metaphysically necessitated features would 'pollute' the dependence base features in such a way that the latter would no longer be properly considered physical, including one according to which (as in the case of a fundamental interaction-based response to the collapse objection) fundamental interactions provide a basis for distinguishing lower-level physical from Strongly emergent goings-on, even when these are deeply dispositionally connected. In re a view on which Strongly emergent goings-on are those which are both fundamental and dependent, and where the notions of fundamentality and dependence are each taken to be primitive (per Barnes 2012): I argue that such a view is too abstract to satisfy the criteria of appropriate and illuminating accommodation; relatedly, it does not provide any clear means of engaging with or addressing either Kim's problem of higher-level causation or the collapse objection, or of ensuring that Strongly emergent goings-on properly contrast with views such as substance dualism. Finally, in re epistemic criteria: I argue that while accounts of Strong emergence as involving one or other epistemic failure have been historically common—per, e.g., appeals to failures of deducibility (Broad 1925), explainability (Horgan 1993), or conceptual entailment (Chalmers 2006), such accounts should be rejected, both because it is clear that the proponents offer the epistemic criteria in service of tracking a metaphysical distinction—in particular, one conforming to the conditions in Strong emergence, and because in any case such epistemic failures are not distinctive of physically acceptable emergence, but can attach to phenomena (e.g., the behaviour of artificial complex systems; see below) for which Strong emergence is clearly not at issue.

Chapter 5: "Complex Systems"

Having established the in-principle viability of both Weak and Strong conceptions of metaphysical emergence, I go on to consider whether certain phenomena are plausibly seen as actually either Weakly or Strongly emergent. I start in Chapter 5 with complex systems, as perhaps the phenomena that have been most often offered as emergent, by scientists as well as philosophers. Complex systems take many forms, both natural (e.g., turbulent water flows, phase transitions, and weather patterns) and artificial (e.g., Conway's 'Game of Life'). And among the distinctive characteristics of complex systems are non-linearity (whereby certain features or behaviours cannot be seen as linear or other broadly additive combinations of features of the system's composing entities), unpredictability (and relatedly, extreme sensitivity to initial conditions), algorithmic incompressibility (whereby the operative equations of motion do not admit of analytic or 'closed' solutions'), 'universality' (whereby certain features are common across diverse micro-structures, especially as associated with asymptotic singularities near critical points), and self-organization (whereby coherent 'system-wide' patterns arise as a result of interactions between parts).

I first consider whether any complex systems might be Strongly emergent (§5.1). I start with a compressed historical discussion of why the British Emergentists (Mill and Broad, among others) took nonlinearity and in-principle failures of predictability to suffice for fundamental novelty (§5.1.1)—a view that,

while reasonable at the time, was undermined by the discovery and creation of complex systems clearly not involving any fundamentally novel powers/interactions/laws. This discussion is useful for appreciating how nonlinearity moved from being a criterion of Strong emergence to being a criterion of Weak emergence (though in ways leaving open, as I argue in §5.1.3, the possibility that some complex systems are Strongly emergent), and for seeing how a recognizable descendant of nonlinearity as a criterion of Strong emergence is present in the aforementioned motivation for new fundamental interactions, reflecting seeming violations of conservation laws. By lights of the latter criterion, I observe, there is presently little support for taking non-mental complex systems to be Strongly emergent (§5.1.4)—though the case is less clear for certain mental phenomena, a topic to which I return in later chapters.

I next consider whether any complex systems might be Weakly emergent (§5.2), focusing on three existing cases for such emergence as involving one or other characteristic of such systems: Bedau's (1997 and 2008) appeal to algorithmic incompressibility (§5.2.1), Mitchell's (2012) appeal to self-organization (§5.2.2), and Batterman's (2000 and 2002) appeal to asymptotic singularities (§5.2.3). I argue that the cases made in these discussions fall short of establishing that complex systems are Weakly emergent, in failing to rule out certain reductionist strategies for accommodating the characteristics at issue. That said, I go on to argue that the prospects for developing these cases in a way that reveals an associated satisfaction of the conditions in Weak Emergence are good (§5.2.4). In particular, after expanding a bit on my (2010) degree-of-freedom (DOF)-based account of Weak emergence, and responding to the concern, due to Morrison (2012) and Lamb (2015), that complex systems involve not fewer but more DOF than base systems (associated with 'order parameters' that emerge near critical points), I argue that complex systems exhibiting universality of the sort Batterman focuses on also have (as he observes) DOF that are eliminated relative to the systems of their composing lower-level entities, and so are Weakly emergent by lights of a DOF-based account. And I go on to offer reasons for thinking that certain other complex systems (Bedau's gliders in Conway's Game of Life; Mitchell's flocks of birds) may also be seen as Weakly emergent by these lights.

Chapter 6: "Ordinary Objects"

In Chapter 6, I turn to the question of whether ordinary objects are either Strongly or Weakly metaphysically emergent. By 'ordinary' objects I have in mind objects which are uncontroversially inanimate (as Thomasson, 2007, puts it) or nonliving (as Merricks, 2003, puts it), and of the sort with which creatures like us are or may be perceptually acquainted. Such objects might be either natural (rocks, feathers, mountains, planets) or artifactual (tables, baseballs, statues). My discussion is broadly neutral on which metaphysical account of objects is correct, so long as a given such account does not rule out of court the possibility that ordinary objects are metaphysically emergent.

I start by considering whether any ordinary objects are either Weakly emergent or (as I will sometimes put it) are 'at least' Weakly emergent, in having at least one feature satisfying the conditions in the schema for Weak emergence (§6.1). I offer three routes to an affirmative answer. First, I argue that ordinary objects of the sort appropriately treated by classical (or 'Newtonian') me-

chanics are Weakly emergent by lights of a DOF-based account, thanks to the elimination of quantum DOF in the classical limit (§6.1.1); second, I argue that a common conception of artifacts as associated with sortal properties and distinctive functional roles, and the associated compositionally flexible persistence conditions typically encoded in these sortal features, supports thinking of artifacts as being at least Weakly emergent by lights of a functional realization account (§6.1.2); third, I argue that ordinary objects typically have metaphysically indeterminate boundaries, which when coupled with an attractive determinable-based account of such indeterminacy (advanced in my 2013 and 2016a), indicates that such ordinary objects are at least Weakly emergent, by lights of a determinable-based account of such emergence (§6.1.3).

I next consider whether any ordinary objects are Strongly emergent (§6.2). I argue that the best case for this stems from the role mentality plays in both the specification and the constitution of the functional roles (typically encoding social practices involving normative or aesthetic goings-on) which are typically associated with artifacts. The ultimate status of such objects as Strongly or rather just Weakly emergent hinges, like the status of certain complex systems involving mentality, on the status as Weakly or Strongly emergent of the associated mental features of persons, of the sort to be discussed in the next chapters.

I close by observing that the results of this chapter undercut the motivations for Thomasson's meta-ontological view, as discussed in her (2010) and elsewhere, according to which investigations into the ontological status of artifactual ordinary objects should proceed differently from investigations into the ontological status of special-science entities (§6.3). Thomasson's suggestion is primarily motivated by thinking, first, that the usually stated concerns with ordinary objects (e.g., Kim-style causal overdetermination concerns) arise from trying to give scientific and ordinary objects (including artifacts) a unified treatment, and second, that the concerns as attaching to scientific goings-on do not admit of any good answers. But as I have argued, there are good responses to the concerns at issue, whether natural or artifactual ordinary objects are at issue. Nothing stands in the way of a systematic treatment of natural and artifactual ordinary objects as at least Weakly emergent, and—contingent upon future empirical results and the import of mentality to be next considered—perhaps even Strongly emergent.

Chapter 7: "Consciousness"

In Chapter 7, I turn to considering whether consciousness or conscious experience of the sort that we and other creatures enjoy is either Weakly or Strongly emergent. There are many forms or species of consciousness, including perceptual awareness of the external world, conscious awareness of internal states (e.g., pain), and self-consciousness (i.e., consciousness of ourselves as conscious beings). Little in this chapter hinges on differences between these forms of consciousness, so I speak generically of consciousness or conscious awareness (or associated mental features), which may have as its seeming object the external world, one's internal states, or (as a special case of the latter) consciousness itself.

I start by considering whether consciousness is Strongly emergent (§7.1). Arguments for consciousness's being Strongly emergent (or in any case physically unacceptable, in a way compatible with being Strongly emergent) typically

rest on the commonly accepted failure of consciousness to be predictable from or explainable in terms of lower-level physical phenomena. Although for reasons mentioned previously, even in-principle epistemic failures can't be the whole story, proponents of these arguments offer reasons for thinking that the explanatory gaps are taken to be metaphysically significant, in reflecting not just mathematical barriers to explanation (e.g., non-linearity), but rather that the subjective or qualitative aspects of conscious experience depart so greatly from lower-level physical features that no physicalist account of consciousness can be correct. I consider the two most promising forms of explanatory gap argument, however, and argue that neither goes through.

I first address knowledge arguments (per Nagel 1974 and Jackson 1982 and 1986) aiming to show that one could have complete physical knowledge of some entity or subject matter, but nonetheless fail to know certain facts pertaining to conscious states associated with the entity or subject matter (§7.1.1). I focus on Jackson's case-based argument, whereby Mary, a scientist confined to a black and white room, comes to possess complete physical knowledge about human color vision; but upon being released and seeing a ripe tomato, learns something new—such that, the conclusion goes, physicalism is thereby revealed to be false. Much physicalist ink has been spilled on responding to Jackson's argument; here I advance a response not much on the books, which proceeds by denying that Mary has complete physical knowledge about human color vision before her release, per what I call the 'Incomplete Physical Knowledge' strategy. I motivate this strategy by observing that a physicalist need not agree that physical knowledge must be 'objective' in the sense of failing to be of subjective or qualitative aspects of reality, since such a view is in tension with physicalism—which maintains, after all, that some sufficiently complex physical goings-on are identical with or realize conscious mental states and associated subjective/qualitative features. Relatedly, I maintain, the physicalist can and arguably should simply grant that acquaintance is a necessary condition for knowing certain physical facts—namely, those providing a constitutive basis for any subjective or qualitative aspects of consciousness there may be. I note certain advantages that the Incomplete Physical Knowledge strategy has over other responses, and diagnose the failure for this strategy to be properly appreciated as reflecting a mistaken characterization of the physical goings-on in overly representational, insufficiently expansive (i.e., appropriately complex), and qualitatively etiolated terms. The upshot is that the knowledge arguments do not provide compelling reason to think that consciousness and its associated subjective and qualitative aspects are actually physically unacceptable, much less actually Strongly emergent.

I next address the conceivability argument advanced and developed by Chalmers (in his 1996, 1999, 2009, and elsewhere), according to which the conceivability of zombies—creatures which are functional and physical duplicates of creatures like us, but which are lacking in any conscious mentality—is taken, in combination with certain other commitments, to establish the Strong emergence of consciousness (§7.1.2). Chalmers's argument goes beyond previous explanatory gap arguments in that the conceivability of zombies is situated in an independently motivated framework—'epistemic two-dimensionalism' (E2D)—according to which certain facts about meaning, which are taken to be a priori accessible, can be used to identify or establish certain facts about modality, expressing or encoding what is genuinely metaphysically possible (necessary, contingent, impossible). It is commonly assumed that the mode of a priori access to

meanings that enters into the E2D strategy proceeds by way of conceiving. Consequently, commitment to the E2D strategy for gaining (much) access to modal truth, and to implementing this strategy via a conceiving-based epistemology of meanings, provides an independent basis for taking the conceivability of zombies to have anti-physicalist metaphysical import, as reflecting a systematic connection between conceivability and metaphysical possibility. The conceivability argument then proceeds as follows:

1. It is conceivable that there is a world which is physically exactly like our world, but in which there is no consciousness.
2. If the world described in (1) is conceivable, then it is metaphysically possible. (E2D)
3. If the world described in (1) is metaphysically possible, then physicalism is false.
4. Physicalism is false.
5. In particular, consciousness is physically unacceptable (and moreover might be Strongly emergent).

The focus of my critical attention here is on the second premise. Drawing on Biggs and Wilson 2017*a* and 2019, I suggest that there is an alternative, and superior, way in which the E2D strategy might be implemented—namely, by appeal to an abduction-based rather a conceiving-based epistemology of the meanings entering into this strategy. I then argue that it is far from clear that the genuine possibility of zombies, or the associated Strong emergence of consciousness, is output from E2D, when this framework is implemented using abduction rather than conceiving. One might wonder, as against this line of thought, whether abduction is apt for purposes of implementing E2D, given that (as above) the access to the meanings which are in turn supposed to provide a basis for access to modal truths is supposed to proceed in a priori fashion. Here again, I draw on joint work with Biggs (Biggs and Wilson 2017*b*), where we argue that, contra common assumption, abduction is an a priori mode of inference—as a priori as conceiving, in particular.[11] The upshot is that, like the knowledge arguments, Chalmers's two-dimensional argument fails to establish that consciousness is actually physically unacceptable, much less Strongly emergent.

I go on to consider whether consciousness is Weakly emergent (§7.2). Here I argue for an affirmative answer, based in the fact that qualitative conscious states—e.g., states of conscious awareness of colors or pains—are typically determinable rather than (maximally) determinate, in a way that defensibly renders them suitable (again, assuming that they are not Strongly emergent) for being realized in determinable-based fashion, and hence Weakly emergent. I first provide two reasons for thinking that various of our perceptions are determinable (§7.2.1), the first being that qualitative mental states are susceptible to Sorites phenomena, and the second reflecting that our perception of macro-entities and

[11] Such a view is not as unusual as it might first appear. To start, the view has precursors in Kant (via the notion of the synthetic a priori) and Carnap (and his appeal to conceptual analysis as involving 'explication,' which proceeds abductively). Moreover, the view reflects the underappreciated fact that the ceteris paribus clauses in abductive principles (e.g., one or other principle of parsimony) effectively operate to shield them from disconfirmation. See our papers for further details.

their features typically fails to register micro-determinate details. Now, as previously, one implementation of the schema for Weak emergence is a determinable-based account of realization, according to which it suffices for the realization of a feature that the feature be a determinable of lower-level physical determinates. So, if the determinable qualitative conscious states at issue can be seen as having lower-level physical determinates, we will be in position to conclude that such conscious features are Weakly emergent.

I then present arguments, due to Ehring (1996), Funkhouser (2006), and Walter (2006), according to which this does not make sense; here the common line is that while the determinable/determinate relation has some feature *F*, the relation between qualitative conscious states and lower-level physical states does not have *F* (§7.2.2). For example, Ehring argues that taking qualitative conscious features to be determinables of lower-level physical determinates is incompatible with the intuitive possibility of there being qualitative mental superdeterminates (e.g., a maximally specific pain), since implying, falsely, that these could be further determined. Drawing on my (2009), I respond to Ehring's and the other concerns by noting, first, that different sciences may treat a single determinable as having different determination dimensions (hence mental features may be superdeterminate relative to a purely psychological science, while being further determined relative to a lower-level physical science), and second, arguing that a proper understanding of the determinable/determinate relation, per

> *Powers-based Determination*: feature P is a determinate of feature Q iff Q is associated with a proper subset of the powers associated with P, and the set of powers had by P but not by Q is not associated with any property,

provides a comprehensible metaphysical basis for accommodating the phenomenon of science-relative determination dimensions. To wit: relative to one set of determination dimensions, reflecting sensitivity to powers associated with the determinable set, a given qualitative conscious state might be characterized as a superdeterminate; but relative to a finer-grained set of determination dimensions (reflecting sensitivity to powers in relevant supersets of the determinable set) that same feature might not be appropriately characterized as a superdeterminate (§ 7.2.3).

Chapter 8: "Free Will"

Free will (or free agency), if such there be, involves the ability to mentally choose an outcome (an intention to ϕ, or a ϕ-ing), where the outcome is 'free' in being, in some substantive sense, up to the agent of the choice. In Chapter 8, I consider whether free will of the sort that we appear to have and to exercise is either Weakly or Strongly emergent.

I start by drawing on Bernstein and Wilson 2016 in order to set up a useful framework for investigating into whether free will is metaphysically emergent (§8.1). Recall that the schemas for Weak and Strong emergence were initially motivated as associated with two specific responses to the problem of higher-level causation. Mental features are a common focus of this problem, but in the usual case the mental features at issue are qualitative or intentional features, for which free choice is supposed not to be at issue. More generally, debates over the status of free will have tended to proceed in relative independence from debates over the status of mental features whose governance by natural law is taken for granted. As Bernstein and I argue, however, the problematics underlying

the free will and the mental causation debates are appropriately seen as special cases of a more general problem, concerning whether and how mental features of a given type may be efficacious, qua the types of feature they are (qualitative, intentional, freely deliberative), given their apparent causal irrelevance—i.e., apparent failure of distinctive efficacy—for effects of the type in question. That the free will and mental causation debates can be seen as special cases of a more general problem serves to suggest certain parallels between positions in the respective debates, which parallels are useful for purposes of assessing whether free will is either Weakly or Strongly emergent.

In the next two sections I develop these parallels for compatibilism and libertarianism, respectively. Again drawing on Bernstein and Wilson 2016, I first argue that a representative range of compatibilist accounts, including accounts of freedom as underdetermination (per, e.g., Ayer 1954), freedom as ownership (per, e.g., Davidson 1963), and freedom as responsibility (per, e.g., Strawson 1962), implement a structurally similar 'proper subset' strategy for responding to the problem of free will (§8.2). Effectively, the general compatibilist strategy is to identify a proper subset of the total causal antecedents of a given outcome (effect) of a mental choosing, as that which is relevant for the choosing's being efficacious qua free; different compatibilists then differ about which proper subsets of the total causal antecedents are those which are so relevant. I then extend this result, arguing that the compatibilist strategy can be more specifically understood as entailing the holding of a proper subset relation between token powers associated with two complex, cotemporal events, corresponding to, first, the mental choosing M in combination with the relevant causal antecedents of M (call this complex event C'), and second, the mental choosing M in combination with the total causal antecedents of M (call this complex event C). I next argue that a representative range of libertarian accounts, including event-causal accounts (per, e.g., Kane 1996 and Merricks 2003), agent-causal accounts (per, e.g., O'Connor 2005), and 'non-causal'[12] accounts (per, e.g., Ginet 1990, McCann 1998, and Stump 1999) are reasonably seen as committed to free will's being associated with a fundamentally novel power—namely, the power to freely choose to ϕ—not had by lower-level physical goings-on, of the sort that satisfaction of the schema for Strong emergence requires (§8.3).

Parallels established, I turn to considering whether (some cases of) free will might be Weakly emergent (§8.4.1). The prospects are good, I argue. Though free choices are not taken to be part of a higher-level system of laws on either compatibilist or libertarian accounts, a compatibilist account is one manifesting the usual Weak emergentist characterization of special-science goings-on as comparatively insensitive to lower-level physical details, in the sense that an agent's reasons for action in a given case float free of many such details (and in particular, are sensitive only to facts about 'relevant' causal antecedents). Since our deliberations and associated acts of choice clearly are insensitive to many microphysical details, then given that free will is understood along compatibilist (Weak emergentist) lines, there is good reason to think that such free will actually exists, and moreover is abundant.

[12] Note that non-causal accounts of libertarian free will only require that the choice not be antecedently caused; they are compatible with, and indeed require, that the choice itself be efficacious (hence have powers).

Notwithstanding that there is presumably plenty of what compatibilists count as free will, is there actually free will of a libertarian, nomologically transcendent variety (§8.4.2)? I offer a new argument for an affirmative answer, as follows:

1. We experience ourselves as seeming to freely choose, in ways transcending any nomological (deterministic or indeterministic) goings-on.
2. In the absence of good reasons to think that our experience of nomologically transcendent free will cannot be taken at face value, we are entitled to take this experience at face value.
3. There are no good reasons to think that our experience cannot be taken at face value.
4. We are entitled to take our experience of nomologically transcendent free will at face value.

The argument is valid, and premise (1) is clearly true (even non-libertarians agree). Premise (2) also seems reasonable: if we have clear experience of some seeming phenomenon, we need good reason not to take that experience at face value. I focus on defending premise (3) against the 'Libet cases' which pose the most serious challenge to taking our experience at face value.

Recall that Libet (1999) determined that when a subject is asked to move their finger and track exactly when the urge to do so occurs, an unconscious 'Readiness Potential' RP precedes the "experience of will" by around 400 milliseconds. Libet and others concluded that conscious will is not the initiator of voluntary action, but instead a consequence of an unconscious physical process that triggers the action. In response, I first canvass certain alternative interpretations of the data, due to Mele (2009) and O'Connor (2005), which are compatible with nomologically transcendent free will. I then offer a new interpretation of my own, which is also so compatible, and which takes advantage of the cotemporal material dependence condition in Strong emergence. On my interpretation, the intention to choose and the associated brain activity are cotemporally initiated, but it takes a bit of time for this fact to consciously register as a complete thought in the agent's mind. Thinking takes time—more time, perhaps, than a choice. A very small lag—less than half a second—would be a natural concomitant of our mental decision-making processes, compatible with transcendent free will. Correspondingly, Libet's assumption that "In the traditional view […], one would expect conscious will to appear before, or at the onset, of the RP, and thus command the brain to perform the intended act" (1999, 49) reflects an overly simplistic account of how nomologically transcendent free will would actually work.

Chapter 9: "Closing Remarks"

In Chapter 9, I summarize the results of the book and call attention to some phenomena whose status as metaphysically emergent deserves further attention, including quantum entanglement, molecular structure, biological systems, brain dynamics, and spacetime. I close with some methodological observations pointing towards other ways in which attention to broadly mereological relationships between sets of powers might serve to shed light on other aspects of higher-level reality, beyond metaphysical emergence.

References

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

Alexander, S., 1920. *Space, time, and deity*. London: Macmillan.

Antony, L.M. and Levine, J.M., 1997. Reduction with autonomy. *Philosophical perspectives*, 11, 83–105.

Ayer, A.J., 1954. Freedom and necessity. *In*: A.J. Ayer, *Philosophical Essays*. London: Macmillan, 271–284.

Baltimore, J.A. 2013. Careful, physicalists: Mind–body supervenience can be too superduper. *Theoria*, 79, 8–21.

Barnes, E., 2012. Emergence and fundamentality. *Mind*, 121, 873–901.

Batterman, R.W., 1998. Why equilibrium statistical mechanics works: Universality and the renormalization group. *Philosophy of science*, 65, 183–208.

Batterman, R.W., 2000. Multiple realizability and universality. B*ritish journal for the philosophy of science*, 51, 115–45.

Batterman, R.W., 2002. *The devil in the details: Asymptotic reasoning in explanation, reduction, and emergence*. Oxford: Oxford University Press.

Baysan, U. and Wilson, J., 2017. Must strong emergence collapse? *Philosophica*, 91, 49–104.

Bedau, M.A., 1997. Weak Emergence. *In*: J. Tomberlin, ed. *Mind, causation and world, philosophical perspectives annual volume 11*. Oxford: Wiley-Blackwell, 375–399.

Bedau, M.A., 2008. Is weak emergence just in the mind? *Minds and machines*, 18, 443–459.

Bennett, K., 2003. Why the exclusion problem seems intractable and how, just maybe, to tract it. *Noûs*, 37, 471–497.

Bennett, K., 2017. *Making things up*. Oxford: Oxford University Press.

Bernstein, S. and Wilson, J.M., 2016. Free will and mental quausation. *Journal of the American philosophical association*, 2, 310–331.

Biggs, S. and Wilson, J.M., 2017a. Abductive two-dimensionalism: A new route to the a priori identification of necessary truths. *Synthese*, 197, 59–93.

Biggs, S. and Wilson, J.M., 2017b. The a priority of abduction. *Philosophical studies*, 174, 735–758.

Biggs, S. and Wilson, J.M., 2019. Abduction vs. conceiving in modal epistemology. *Synthese*, 198, 2045–2076.

Boyd, R., 1980. Materialism without reduction: What physicalism does not entail. *In*: N. Block, ed. *Readings in the philosophy of psychology*. Cambridge (Mass.): Harvard University Press, 67–106.

Broad, C.D., 1925. *Mind and its place in nature*. Cambridge: Kegan Paul.

Carruth, A., 2018. Emergence, reduction and the identity and individuation of powers. *Topoi*. 39, 1021–1030.

Chalmers, D.J., 1996. *The conscious mind*. Oxford: Oxford University Press.

Chalmers, D.J., 1999. Materialism and the metaphysics of modality. *Philosophy and phenomenological research*, 59, 473–96.

Chalmers, D.J., 2006. Strong and weak emergence. *In*: P. Clayton and P. Davies, eds. *The re-emergence of emergence*. Oxford: Oxford University Press, 244–256.

Chalmers, D.J., 2009. The two-dimensional argument against materialism. *In*: B.P. McLaughlin and S. Walter, eds. *Oxford Handbook to the Philosophy of Mind*. Oxford: Clarendon Press.

Clapp, L., 2001. Disjunctive properties: Multiple realizations. *Journal of philosophy,* 98, 111–136.

Cleve, J. van., 1990. Mind-dust or magic? Panpsychism versus emergence. *Philosophical perspectives*, 4, 215–226.

Crane, T., 2001. The significance of emergence. *In*: C. Gillett and B. Loewer, eds. *Physicalism and its discontents*. Cambridge: Cambridge University Press, 207–224.

Craver, C.F., 2001. Role functions, mechanisms, and hierarchy. *Philosophy of science*, 68, 53–74.

Cummins, R. 1975. Functional analysis. *Journal of philosophy*. 72, 741–764.

Cunningham, B., 2001. The reemergence of emergence. *Philosophy of science*, 68, S62–S75.

Dasgupta, S., 2014. The possibility of physicalism. *Journal of philosophy*, 111, 557–592.

Davidson, D., 1963. Actions, reasons, and causes. *Journal of philosophy*, 60, 685–700.

Dosanjh, R., 2014. *A defense of reductive physicalism*. Thesis (PhD). University of Toronto.

Dosanjh, R., 2019. Token-distinctness and the disjunctive strategy. *Erkenntnis*, 86, 715–732.

Douven, I., 2021. Abduction. *In*: E. Zalta, ed. *The Stanford encyclopedia of philosophy*, Stanford University, summer 2021 edition.

Ehring, D., 1996. Mental causation, determinables, and property instances. *Noûs*, 30, 461–480.

Field, H., 2003. Causation in a physical world. *In*: M. Loux and D. Zimmerman, eds. *The oxford handbook of metaphysics*. Oxford: Oxford University Press.

Fodor, J., 1974. Special sciences (or, the disunity of science as a working hypothesis). *Synthese*, 28, 77–115.

Fodor, J., 1987. *Psychosemantics*. Cambridge, MA: MIT Press.

Francescotti, R., 2007. Emergence. *Erkenntnis*, 67, 47–63.

Funkhouser, E., 2006. The determinable-determinate relation. *Noûs*, 40, 548–569.

Gillett, C., 2002. The dimensions of realization: A critique of the standard view. *Analysis*, 62, 316–323.

Gillett, C., 2016. *Reduction and emergence in science and philosophy*. Cambridge: Cambridge University Press.

Ginet, C., 1990. *On action*. Cambridge: Cambridge University Press.

Harman, G.H., 1965. The inference to the best explanation. *Philosophical review*, 74, 88–95.

Haug, M.C., 2010. Realization, determination, and mechanisms. *Philosophical studies*, 150, 313–330.

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

Hempel, C. and Oppenheim, P., 1948. Studies in the logic of explanation. *Philosophy of science*, 15, 135–175.

Horgan, T., 1993. From supervenience to superdupervenience: Meeting the demands of a material world. *Mind*, 102, 555–586.

Howell, R.J., 2009. Emergentism and supervenience physicalism. *Australasian Journal of philosophy*, 87, 83–98.

Humphreys, P., 1996. Aspects of emergence. *Philosophical topics*, 24, 53–70.

Jackson, F., 1982. Epiphenomenal qualia. *Philosophical quarterly*, 32, 127–136.

Jackson, F., 1986. What Mary didn't know. *The journal of philosophy*, 83, 291–295.

Jaworski, W., 2002. Multiple-realizability, explanation and the disjunctive move. *Philosophical studies*, 108, 289–308.

Kane, R., 1996. *The significance of free will*. New York: Oxford University Press.

Kim, J., 1992. 'Downward causation' in emergentism and nonreductive physicalism'. *In*: A. Beckermann, H. Flohr, and J. Kim, eds. *Emergence or reduction? Prospects for nonreductive physicalism*, Berlin: De Gruyter, 119–138.

Kim, J., 1999. Making sense of emergence. *Philosophical studies*, 95, 3–36.

Kim, J., 2010. Thoughts on Sydney Shoemaker's physical realization. *Philosophical studies*, 148, 101–112.

Klee, R., 1984. Micro-determinism and concepts of emergence. *Philosophy of science,* 51, 44–63.

Ladyman, J. and Ross, D., 2007. *Every thing must go: Metaphysics naturalized*. Oxford: Oxford University Press.

Lamb, M., 2015. *Characteristics of non-reductive explanations in complex dynamical systems research*. Thesis (PhD). University of Cincinnati.

LePore, E. and Loewer, B., 1987. Mind matters. *The journal of philosophy*, 84, 630–642.

LePore, E. and Loewer, B., 1989. More on making mind matter. *Philosophical topics*, 17, 175–191.

Lewes, G.H., 1875. *Problems of life and mind*. London: Kegan Paul, Trench, Turbner & Co.

Libet, B.W., 1999. Do we have free will? *Journal of consciousness studies*, 6 (8-9), 47–57.

MacDonald, C. and MacDonald, G., 1986. Mental causes and explanation of action. *In*: L. Stevenson, R. Squires, and J. Haldane, eds. *Mind, causation, and action*. Oxford: Basil Blackwell, 145–158.

McCann, H., 1998. *The works of agency: On human action, will, and freedom*. Ithaca: Cornell University Press.

McLaughlin, B., 1992. The rise and fall of British emergentism. *In*: A. Beckerman, H. Flohr, and J. Kim, eds. *Emergence or reduction? Essays on the prospects of nonreductive physicalism*. Berlin: De Gruyter, 49–93.

McLaughlin, B.P., 2007. Mental causation and Shoemaker-realization. *Erkenntnis.* 67, 149–172.

Megill, J., 2013. A defense of emergence. *Axiomathes*, 23, 597–615.

Mele, A. 2009. *Effective intentions: The power of conscious will*. Oxford: Oxford University Press.

Melnyk, A., 2003. *A physicalist manifesto: Thoroughly modern materialism*. New York: Cambridge University Press.

Melnyk, A., 2006. Realization-based formulations of physicalism. *Philosophical studies*, 131, 127–155.

Merricks, T. 2003. *Objects and persons*. Oxford: Clarendon Press.

Mill, J.S., [1843]1973. *A system of logic*. Toronto: University of Toronto Press.

Mitchell, S.D. 2012. Emergence: Logical, functional and dynamical. *Synthese*, 185, 171–186.

Morgan, C.L., 1923. *Emergent evolution*. London: Williams & Norgate.

Morris, K., 2018. *Physicalism deconstructed: Levels of reality and the mind–body problem*. Cambridge: Cambridge University Press.

Morrison, M., 2012. Emergent physics and micro-ontology. *Philosophy of science*, 79, 141–166.

Nagel, E., 1961. *The structure of science*. London: Routledge & Kegan Paul.

Nagel, T., 1974. What is it like to be a bat? *The philosophical review*, 83, 435–450.

Newman, D., 1996. Emergence and strange attractors. *Philosophy of science*, 63, 245–261.

Ney, A., 2010. Convergence on the problem of mental causation: Shoemaker's strategy for (nonreductive?) physicalists. *Philosophical issues*, 20, 438–445.

Ney, A., 2022. Review of *Metaphysical emergence* by Jessica Wilson. *Notre Dame philosophical reviews.*

Noordhof, P., 2010. Emergent causation and property causation. *In*: C. Macdonald and G. Macdonald, eds. *Emergence in mind*. Oxford: Oxford University Press, 69–99.

O'Connor, T., 1994. Emergent properties. *American philosophical quarterly*, 31, 91–104.

O'Connor, T., 2005. Free will. *In*: J.M. Fischer, ed. *Free will: Critical concepts in philosophy*. New York: Routledge, 7–32.

O'Connor, T. and Wong, H.Y., 2005. The metaphysics of emergence. *Noûs*, 39, 658–678.

Paolini Paoletti, M., 2017. *The quest for emergence*. Munich: Philosophia.

Papineau, D., 1993. *Philosophical naturalism*. Oxford: Basil Blackwell.

Pereboom, D., 2002. Robust non-reductive materialism. *Journal of philosophy*, 99, 499–531.

Poland, J., 1994. *Physicalism: The philosophical foundations*. Oxford: Clarendon Press.

Polger, T.W. 2007. Realization and the metaphysics of mind. *Australasian journal of philosophy*, 85, 233–259.

Putnam, H., 1967. Psychological predicates. *In*: D.D. Merrill and W. Capitan, eds. *Art, mind, and religion*. Pittsburgh: University of Pittsburgh Press, 37–48.

Rueger, A. and McGivern, P., 2010. Hierarchies and levels of reality. *Synthese*, 176, 379–397.

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

Schaffer, J., 2009. On what grounds what. *In*: D. Manley, D. Chalmers, and R. Wasserman, eds. *Metametaphysics: New essays on the foundations of ontology*, Oxford: Oxford University Press, 347–383.

Schroder, J., 1998. Emergence: non-deducibility or downward causation? *The philosophical quarterly*, 48, 433–452.

Seager, W., 1999/2016. *Theories of consciousness: an introduction and assessment*. London: Routledge.

Searle, J.R., 1992. *The rediscovery of the mind*. Cambridge (Mass.): MIT Press.

Shoemaker, S., 2000/2001. Realization and mental causation. *In*: Eds. *Proceedings of the 20th world congress in philosophy*. Cambridge: Philosophy Documentation Center, 23–33.

Silberstein, M. and McGeever, J., 1999. The search for ontological emergence. *Philosophical quarterly*, 50, 182–200.

Sperry, R., 1986. Discussion: macro- versus micro-determination. *Philosophy of science*, 53, 265–270.

Stephan, A., 2002. Emergentism, irreducibility, and downward causation. *Grazer philosophische studien*, 65, 77–93.

Strawson, P.F., 1962. Freedom and resentment. *Proceedings of the British academy*, 48, 1–25.

Stump, E., 1999. Dust, determinism, and Frankfurt. *Faith and philosophy*, 16, 413–422.

Tahko, T.E., 2018. Fundamentality. *In*: E. Zalta, ed. *The Stanford encyclopedia of philosophy*. Stanford University, Fall 2020 edition.

Taylor, E., 2015. Collapsing emergence. *Philosophical quarterly*, 65, 732–753.

Thomasson, A.L., 2007. *Ordinary objects*. Oxford: Oxford University Press.

Thomasson, A.L., 2010. The controversy over the existence of ordinary objects. *Philosophy compass*, 5, 591–601.

Thompson, E. and Varela, F.J., 2001. Radical embodiment: neural dynamics and consciousness. *Trends in cognitive sciences*, 5, 418–425.

Van Gulick, R., 2001. Reduction, emergence and other recent options on the mind/body problem: a philosophic overview. *Synthese*, 8, 1–34.

Walter, S., 2006. Determinates, determinables, and causal relevance. T*he Canadian journal of philosophy*, 37, 217–243.

Walter, S., 2010. Taking realization seriously: no cure for epiphobia. *Philosophical studies*, 151, 207–226.

Wilson, J.M., 1999. How superduper does a physicalist supervenience need to be?' *The philosophical quarterly*, 49, 33–52.

Wilson, J.M., 2002. Causal powers, forces, and superdupervenience. *Grazer philosophische-studien*, 63, 53–78.

Wilson, J.M., 2005. Supervenience-based formulations of physicalism. *Noûs,* 39, 426–459.

Wilson, J.M., 2006. On characterizing the physical. *Philosophical studies*, 131, 61–99.

Wilson, J.M., 2009. Determination, realization, and mental causation. *Philosophical studies*, 145, 149–169.

Wilson, J.M., 2010. Non-reductive physicalism and degrees of freedom. *British journal for the philosophy of science*, 61, 279–311.

Wilson, J.M., 2011. Much ado about 'something': Critical notice of Chalmers, Manley, Wasserman, *Metametaphysics*. *Analysis*, 71, 172–188.

Wilson, J.M., 2013. A determinable-based account of metaphysical indeterminacy. *Inquiry*, 56, 359–385.

Wilson, J.M., 2014. No work for a theory of grounding. *Inquiry*, 57, 1–45.

Wilson, J.M., 2015. Metaphysical emergence: Weak and strong. *In*: T. Bigaj and C. Wüthrich, eds. *Metaphysical emergence in contemporary physics; Poznan studies in the philosophy of the sciences and the humanities*, Amsterdam/New York: Brill, 251–306.

Wilson, J.M., 2016a. Are there indeterminate states of affairs? Yes. *In*: E. Barnes, ed. *Current controversy in metaphysics,* London: Routledge, 105–119.

Wilson, J.M., 2016b. The question of metaphysics. *The philosophers' magazine*, 74, 90–96.

Wilson, J.M., 2016c. Three barriers to philosophical progress. *In*: D. Broderick and R., Blackford, eds. *Philosophy's future: the problem of philosophical progress*, Oxford: Wiley-Blackwell, 91–104.

Wilson, J.M., 2021. *Metaphysical emergence*. Oxford: Oxford University Press.

Wilson, J.M., 2022/2017. Determinables and determinates. In E. Zalta, ed. *The Stanford encyclopedia of philosophy*. Stanford University, Spring 2023 edition.

Wilson, J.M., forthcoming*a*. The fundamentality first approach to metaphysical structure. A*ustralasian journal of philosophy*.

Wilson, J.M., forthcoming*b*. On the notion of diachronic emergence. In: D. Yates, ed. *Rethinking emergence*.

Wilson, J.M., under contract. *Fundamentality and metaphysical dependence*. Oxford: Oxford University Press.

Wimsatt, W., 1996. Aggregativity: Reductive heuristics for finding emergence. *Philosophy of science*, 64, 372–384.

Witmer, D.G., 2003. Functionalism and causal exclusion'. *Pacific philosophical quarterly*, 84, 198–214.

Yablo, S., 1992. Mental causation. *The philosophical review*, 101, 245–280.

Yates, D., 2012. Functionalism and the metaphysics of causal exclusion. *Philosophers' imprint*. 12, 1–25.

Yates, D., 2016. Demystifying emergence. *Ergo: an open access journal of philosophy*, 3, 809–841.

Biochemical Functions as Weakly Emergent

Francesca Bellazzi

University of Birmingham

Abstract

This paper will consider how the account of weak emergence presented by Wilson in the book *Metaphysical emergence* (2021) can be used to explore the relation between biochemical functions and chemical structure in biochemical molecules, as vitamin B12. The structure of the paper is the following. Section 2 will introduce why biochemical functions are interesting from a philosophical perspective and why their relation to molecular structure can be seen as problematic. In doing so, it will consider the definition of biochemical functions as in Bellazzi (2022) for which they can be seen as sets of chemical dispositional properties that contribute to biological processes. Section 3 will explore how, given this definition of biochemical functions, we can interpret the relation between chemical structure and biochemical structure via weak emergence. Section 4 concludes.

Keywords: Metaphysical emergence, Biochemical functions, Biochemical kinds.

1. Introduction

This paper will consider how the account of weak emergence presented by Wilson in the book *Metaphysical Emergence* (2021) can be used to explore the relation between biochemical functions and chemical structure in biochemical molecules, as vitamin B12. The discussion of the relation between biochemical function and chemical structure is relevant to the debate concerning inter-level relations together with being a foundational topic for biochemistry (Santos et al. 2020).[1] Moreover, the results of this paper provide a novel application of Wilson's account of weak emergence, enriching the case studies that can fit with the framework and offering new insights into the understanding of weak emergence in non-yet-considered cases.

The structure of the paper is the following. Section 2 will introduce why biochemical functions are interesting from a philosophical perspective and why their relation to molecular structure can be seen as problematic. In doing so, it will

[1] This paper draws on some of the results in Bellazzi 2023.

Argumenta 10, 1 (2024): 225—235
ISSN 2465-2334

DOI 10.14275/2465-2334/202418.bel
First published: 30 November 2024

consider the definition of biochemical functions as in Bellazzi 2022 for which they can be seen as sets of chemical dispositional properties that contribute to biological processes. Section 3 will explore how, given this definition of biochemical functions, we can interpret the relation between chemical structure and biochemical structure via weak emergence. Section 4 concludes.

2. Structure and Function in Biochemical Kinds

Chemistry is often taken to be the domain of chemical structure and kinds characterized in micro-structural terms, such as constituent atomic properties.[2] Biology, instead, is the domain of evolutionary functions, etiological classifications and pluralism (Slater 2009; Bartol 2016). Biochemistry stands as an hybrid domain between the two. While it is not easy to provide a set of necessary and sufficient conditions for a kind to be biochemical, the literature on the topic agrees that biochemical kinds need to exhibit at least two kinds of properties: structural ones and functional ones (Slater 2009; Bartol 2016; Havstad 2016, 2018; Kistler 2018; Tahko 2020). Proteins, for example, are characterised in terms of structure, the amino-acid chain that composes them, and in terms of the functional roles that they play within biological systems.

Prima facie, this definition or the combination of these two sets of properties might not be particularly problematic, however the exact relation between structural and functional properties still posits questions (Bartol 2016; Tahko 2020). One of the reasons why this is so is based on the complexity of the relations between structure and function, as they often take the form of multiple realisability and multiple determinability. Multiple realisability (MR) refers to a phenomenon in which the same entity or property can be realised by different ones.[3] For example, the property of being an eye can be realised by different organs in different animals. Multiple determinability (MD) refers to the opposite phenomenon: when the same entity can determine different properties or other entities. For example, the same chemical compound can enter into different chemical reactions, realising different properties.

In the biochemical case, MR and MD are particularly relevant because the same biochemical function can be realised by multiple microstructures and the same microstructure can realise multiple biochemical functions (Tahko 2020). Two relevant examples in this regard are haemoglobin for MR and the crystalline proteins for MD. As discussed and presented by Tahko (2020, 2021), haemoglobin is a protein with the function of binding and releasing oxygen and can be constituted by at least two different polypeptide chains (or more). The biochemical function of haemoglobin can be considered an instance of MR, as the function of binding and releasing oxygen is realised by at least two distinct macromolecules (chains of polypeptides) that present some micro-structural differences. This can challenge the identification of an identity reductive relation between the chemical structural properties and the functional ones. Multifunctional proteins or "moonlighting" proteins, such as crystallines, represent instances of MD instead. Crystallines are structural proteins present in all vertebrates' eye lenses, having a function in allowing sight, but they can also have an enzymatic role in digestive

[2] Even if this has been challenged as in Tobin 2010, Havstad 2016, 2018.

[3] Realization can be defined as a "synchronic ontological dependence relation, distinct from identity, and that transmits physical legitimacy from physical realizers to what is realized" (Polger and Shapiro 2016: II, 4).

processes. In these cases, we notice a form of MD, as the same chemical structure can lead to very different functions in sight and digestion mechanisms (Tobin 2010; Bartol 2016; Tahko 2020). This again challenges a direct identification of the relation between structure and function, as a strict identity relation between the some underlying structural properties and functional properties does not hold. Moreover, both MR and MD generate issues of taxonomy or classification. If we follow a micro-structuralist approach, then we should favour structure over function and have either many kinds that have the same function (in the case of MR) or one unique kind that has different functions (in the case of MD). If we follow a functional approach, then we have two or three—or as many as the functions—different kinds (in the case of MD) or one kind (in the case of MR).

As a reaction to these tensions, Bartol argues that we should bite the bullet and simply embrace the duality of the two sets of properties: there are chemical structural ones and the biological functional ones (2016). However this approach does not really do justice to the features of biochemical macromolecules that display *both* chemical structure and biological function. These two features are strongly entangled, as supported by some more complex relations between the functions and the chemical structure (see also Goodwin 2011). For instance, Tahko suggests that some cases of MD can be explained or derived from the amphoteric nature[4] of some microstructures (2020). In the cases of some moonlighting proteins for instance, their dual-functions nature can be seen as rooted in some chemical properties of the molecule (Goodwin 2011; Tahko 2020), or at least this can be an option to be analysed in detail.[5] The scientific successes of biochemistry in predicting, manipulating and explaining phenomena encourages instead the exploration of the relation between structure and function, despite its complexity. This is so because this discipline combines chemical and physical model systems to explain and predict biological phenomena.[6]

3. The Double Problem of Biochemical Functions

In order to explore the relation between the chemical structure and biochemical functions one should clarify what are the terms under discussion. Chemical structure comprises both the characterisation of the electronic structure and the molecular geometry of the molecule. What about functional properties? Functional properties in the biochemical context generate what we can call the double problem of biochemical function: the "relation problem" and the "function problem". The "relation problem" asks about the relationship between the chemical structure and the function of a biochemical molecule: how a chemical structure can realise a given biochemical function. As briefly introduced in the previous section, the relation problem is generated by the fact that functional properties in the

[4] An amphoteric chemical substance is one that can react both as a base or as an acid.

[5] The reducibility of the dual nature of moonlighting proteins has been challenged by Santos et al. (2020). This article stresses the importance of analysing the "dynamical interplay between the micro-level of the parts and the macro-level of the relational structures of their systems" in order to understand these proteins (2020: 1). Here I am not supporting the reducibility of biochemical functions to chemical structural properties but rather the relation between functional and structural properties.

[6] The Biochemical Society defines biochemistry as "the branch of science that explores the chemical processes within and related to living organisms" (https://biochemistry.org/education/careers/becoming-a-bioscientist/what-is-biochemistry/).

biochemical domain are often multiply realised, and because biochemical molecules manifest multiple determinability (see Slater 2009; Bartol 2016; Tahko 2020). Furthermore, it is difficult to understand which of the two components, the functional or the structural, has ontological priority in the taxonomy and identification of the biochemical kinds (Slater 2009; Bartol 2016; Tahko 2020). The "function problem" instead asks what biochemical functions are and how they relate to biological functions and the biological component of the kind (Tahko 2020, Bellazzi 2022). Let us consider these problems in more detail with the main case study of this paper, vitamin B12 (as in Bellazzi 2022).

3.1 Vitamin B12

Vitamins B12 are cobalamin chemical compounds that can act as coen-zymes in specific biological processes—specifically, propionate metabolism and methionine biosynthesis. This vitamin comes in four forms—or vitamers—that display similar but different chemical structures: cyanocobalamin, methylcobalamin, hydroxocobalamin, adenosylcobalamin (Combs 2012: 377; Fang et al. 2017).[7] They share a cobalt-corrin complex and the coenzyme function in humans for various biochemical processes such as hematopoiesis, DNA and RNA production, neural metabolism, and carbohydrate, fat, and protein metabolism.[8] Accordingly, these chemical compounds are classified under the same category, 'B12 vitamin', because they display a combination of stable microstructure, a cobalt-corrin complex, and physiological functions.

Vitamin B12 represents an interesting case study relevant to discussing the relation between structure and function because it displays both MR and MD. First, it presents a form of MR in that the biochemical functions of vitamin B12 can be realised by each of the four vitamers recognised in scientific practice.[9] Second, vitamin B12 plays various roles in human physiology, acting in different biological processes, from DNA and RNA production to hematopoiesis, displaying a form of MD too. The combination of MR and MD challenges the identification of simple relations between structure and function. For instance, it makes forms of identity-based reduction, in which the functions of vitamin B12 would be identical to some of the properties of the microstructure, difficult to hold (Tahko 2020). For the sake of the example, let me focus on the function "being a coenzyme in hematopoiesis (the production of blood cells)" (**Coenz-Blood**). B12 vitamers have a biochemical function in the proliferation of erythroblasts (red blood cells) during their differentiation (Koury and Ponka 2004). This happens because vitamin B12 acts as a coenzyme in the reaction involved in regenerating methionine, which is required in normal erythropoiesis. This function is a definitionally important part of the four vitamers of B12: it distinguishes generic cobalt-corrin

[7] A more detailed description is the following: vitamin B12 is "the generic descriptor for all corrinoids (compounds containing the cobalt-centered corrin nucleus) exhibiting qualitatively the biological activity of cyanocobalamin".

[8] Reference for chemical structure and function of vitamin B12 (https://pubchem.ncbi.nlm.nih.gov/compound/Cobalamin). Also, Chapter 12 "B12 Vitamin" in Combs' *The Vitamins: Fundamental Aspects in Nutrition and Health* (2012).

[9] This might represent an instance of multiple constitution of the kind B12, where this kind can be constituted by different chemical compounds that share some functional properties (Kistler 2018). In Kistler, a kind is multiply constituted when it can be constituted by two or more microscopic structures (2018: 18). See also Gillet 2013.

complexes from B12 vitamers, and this shows that, even if it might not be necessary and sufficient on its own to define B12, the functional component is nevertheless important.

Let us go back to the double problem of biochemical functions and elucidate them with the example. First, the "relation problem": **Coenz-Blood** is realised in four different ways via the four vitamers of vitamin B12 and, as such, the relation between the chemical properties of the vitamin B12 and one of its functions should be further explored. The MR of **Coenz-Blood** means that it is at least challenging or not straightforward to map a 1:1 correspondence between it and the possible underlying physicochemical properties. The realisation of this function should be further explored. Second, the "problem of function": what does it mean that vitamin B12 has **Coenz-Blood**} as a *biochemical function*?

The combination of these two problems of biochemical functions might support the suggestion that structure and function could be considered independently. The realisation problem challenges the unification or reduction between the biochemical functions of B12 and its chemical structure. The function problem supports a separation between the chemical and the biological component of biochemical kinds because the nature of biochemical functions could be subsumed under some biological characteristics, which do not relate straightforwardly to the chemical. However, the successes of biochemistry itself seem to provide reasons for the opposite: if we can explain, predict and manipulate biochemical kinds in terms of their function and composition, the two aspects need to be related and, to some extent, ontologically unified.

In order to do so, we should, first, offer a definition of biochemical functions that considers the relation between chemical powers and properties and being dependent on biological context. In this regard, the analysis will start from the following characterisation of biochemical functions (as in Bellazzi 2022):

> **BC-function**: Biochemical functions are associated with a set of chemical powers to bring out a specific effect within biological processes. These biological processes are a product of evolution and, as such, the relevant chemical powers are indirectly evolutionary selected [Fig. 1].

This account of biochemical functions is in line with the general characterisation of biochemistry as the science that considers the behaviour and effects of chemical processes in biological systems (Santos et al. 2020). Moreover, this approach to biochemical functions allows us to answer the function problem, telling us what these properties are, while maintaining the autonomy of the two properties. This provides a starting point to explore the relation between structure and function.

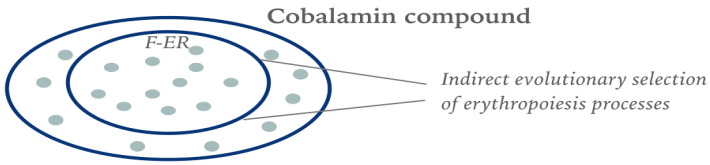


Fig. 1 – The evolutionary selection of the relevant dispositional properties or chemical powers for biochemical functions. In the example, F-ER is **Coenz-Blood** as the function to contribute to erythropoiesis for vitamin B12 as the relevant cobalamin compound.

4. Biochemical Functions as Weakly Emergent

As mentioned in the previous sections, a straightforward form of identity reduction is challenged by the widespread cases of MD and MR in the biochemical domain. Moreover, the set of dispositions relevant to biochemical functions are not any arbitrary chemical powers of the considered molecule or compound but some very specific ones. The relevant powers are those contributing to biological processes and have undergone at least an indirect selection process. The consideration of the biological process they contribute to and—indirectly—evolution that has selected such specific chemical powers is necessary to understand the relevant set of powers (Santos et al. 2020; Bellazzi 2022). Moreover, the causal efficacy of biochemical molecules is distinctive in that it should bring about specific effects within biological processes. Accordingly, an answer to the relation problem should take into account the specificity of biochemical functions together with the relation with structure. In order to provide such an answer, I will consider weak emergence via the proper subset strategy, as in Wilson (2011, 2015, 2021) and as suggested by Tahko (2020). This account, I will suggest, provides an answer to the relation problem and allows for the specificity of biochemical functions.

4.1. Weak Emergence and the "Proper Subset of Powers Strategy"

Weak emergence is a form of emergence compatible with non-reductive physicalism: there is only one broader kind of properties, physical properties. According to non-reductive physicalism, higher-level entities are real and constitute a novel level of reality, being distinctively causally efficacious; at the same time, their causal actions operate in a way respecting physical causal closure and hence in line with physicalism.[10] This combination of distinctiveness and causal efficacy, together with a sense of dependence, can be maintained by defending a form of weak emergence based on the "Proper Subset of Powers strategy" (Wilson 2011, 2021; Tahko 2020).[11] This strategy comprises two steps: i) accepting the *Token Identity of Powers Condition*; ii) accepting the *Proper Subset of Powers Condition*. The first states that every token power of a given token feature H on an occasion *t* is identical with a token power of the token feature L on which H co-temporally materially depends at *t*.[12] The second states that the token feature H has at *t* a non-empty proper subset of the token powers of the token feature L on which H co-temporally materially depends on at *t* (as formulated in Wilson 2021, 57-58). The combination of these two conditions constitutes the basis for a weak emergence relation between the higher and the lower-level entities or features:

[10] The principle of causal closure is often taken as a condition for forms of physicalism and claims that "all physical effects have sufficient physical causes", avoiding cases of problematic overdetermination.

[11] This strategy presupposes a very simple ontology of objects, properties, and powers. Properties are instantiated by objects and are identified by a range of causal powers (Shapiro 2020). In this case, a biochemical molecule instantiates the property "having a given biochemical function", individuated by a specific set of causal powers.

[12] Material dependence implies a form of substance monism, in line with physicalism, and a form of minimal nomological supervenience of the emergent features *type* H on the base features *type* L (Wilson 2021: 73). This means that supervenience should happen with at least nomological necessity.

> **WE**: "What is it for token feature H to be **Weakly Metaphysically Emergent** from token feature L on a given occasion is for it to be the case, on that occasion, i) that H co-temporally materially depends on L, and ii) that H has a non-empty proper subset of the token powers had by L" (Wilson 2021: 75; variables modified, emphasis added).

The first condition i) allows for a form of dependence as there is a token identity of the powers associated with the two features; the second condition ii) allows for a form of distinctiveness. In particular, this account allows for a form of relation between the features because the token powers of the realised feature H are nothing more than a subset of the token powers of a realising feature L, and the two features can be unified as the two sets of powers are both physically acceptable and the token powers of both sets are identical (as also in Shapiro 2020). At the same time, H is ontologically autonomous from L because H has a *proper subset* of the token powers of L and by Leibniz's laws and via set-theory principle, a proper subset of token powers is different from its set of token powers. This permits to maintain the type difference between H and L. The proper subset strategy also allows for a form of causal autonomy, as discussed by Wilson (2011, 2021). Specifically, H has a distinctive causal profile compared to L because it possesses a distinctive set of causal powers or distinctive causal profile compared to L. H's causal autonomy is based on the fact that H has a distinctive set of powers compared to the feature from which it emerges. One of the advantages of this account is that it allows for the relation between the higher and the lower level features, but the higher level ones can still be maintained as ontologically autonomous (Wilson 2011).

Moreover, as will be further detailed in 4.3, the proper subset strategy and weak emergence are able to deal with MR and MD. In the case of MR, it can be possible to identify more than one district token power subset of the lower-level L that can be associated with the higher-level feature H. While in the case of MD, the token set of powers of a given lower-level feature L could present different proper subsets of token powers associated with different higher-level emergent feature H. This allows the account to tackle with some of the issues concerning the relation between structure and function.

4.2 Biochemical Functions Are Weakly Emergent

Let us now consider the interface between biochemical functions and chemical properties and the answer to the relation problem in the light of weak emergence. As in the provided definition, a biochemical function is associated with a set of chemical token powers to bring in a given effect within biological processes (Bellazzi 2022). More precisely, the relation between the token powers associated with the biochemical functions and the correspondent chemical powers can be interpreted with the proper subset view. A biochemical function (BF) has in a given *t* a proper subset of token powers of the set of chemical token powers of the chemical molecule. This proper token subset is individuated via the evolutionary history of the biological process to which BF contributes. Accordingly, following the aforementioned account, we can state the weak emergence of the BF:

> **WE$_{BF}$**: A biochemical function BF weakly emerges from the chemical compound (C) under consideration at a given *t* because: i) BF co-temporally

materially depends on C at *t*; ii) BF has an identifiable and non-empty proper subset of token powers of C at *t*.

At a given *t*, it is possible to identify the biochemical functions as being associated with a proper subset of the chemical powers, with the powers associated with BF being token identical at *t* to powers in C. This makes the biochemical function BF *type* different from C, while it also allows us to maintain that the biochemical functions are co-temporally materially dependent on the chemical ones. Biochemical functions can then be considered weakly emergent from the chemical powers of the molecule and this provides an answer to the relation problem: the relation between the chemical properties of a biochemical kind and the functions is weak emergence. This also allows the identification of a relation between structural and functional properties, given by the token identity of the instances of the biochemical functions and the chemical properties, while at the same time maintaining a type difference and the related causal efficacy. Moreover, as will be elucidated in the next subsection, this view is also compatible with MR and MD.

In the case of vitamin B12, **Coenz-Blood** has a specific proper subset of the chemical powers of cobalamin, the ones relevant to the regeneration of erythroblasts in hematopoiesis. Those powers are those involved in the relevant co-enzymatic action that the vitamin plays: the token of the powers of **Coenz-Blood** are the same token powers of the cobalamin compound involved in the process, however the causal contribution is distinctive. The function **Coenz-Blood** emerges from the chemical compound in that it has a proper specific subset of causal powers. Specifically, in this specific case, it amounts to those chemical properties that allow for the regeneration of methionine via "the transfer of a methyl group from 5-methyl-THF to homocysteine via methylcobalamin" (Koury and Ponka 2004: 109). This set is not arbitrarily chosen, but it is identifiable thanks to the evolutionary history of the different biological processes in which B12 acts as a co-enzyme [see Figure 1]. The causal contributions are those relevant to the given environment and the given process. The biochemical functions of B12 vitamins can be considered weakly emergent from the chemical dispositional properties of cobalamin compounds at a given time *t*. This makes the causal profile of vitamin B12 distinctive, as recognised in scientific practice and in the functional characterisation of B12. At the same time, this emergence is only weak as it does not presupposes any stronger forms of ontological novelty, as the one of a strong form of emergence of a physically unacceptable variety. The identity of the token powers associated with both the emergent feature and the lower basis allows us to maintain a relation between structural and functional properties. The proper subset view and weak emergence allow us then to answer to the relation problem.

4.3 Multiple Realisability and Multiple Determination

As previously presented, biochemical functions are multiply realisable, and in some biochemical cases, such as in the crystallin protein, the same chemical features can be determined into many biochemical functions. This is often taken as a challenge to the identification of a relation between structure and function. Here, we have presented the proper subset view and weak emergence as an answer to the relation problem. However, more must be said on how this view can be compatible with MD and MR.

MR and MD are "type issues": it is the realised *type* that can be multiple realisable or be one of the determinations of a given lower-level feature. How are they compatible with weak emergence as defined above? Starting with MR, it is the type function **Coenz-Blood** that is multiply realisable by the four vitamers of B12. However, in a given moment, such as during a specific instance of hematopoiesis, a token instance of **Coenz-Blood** will be realised by a specific token instance of the four vitamers of B12. At the time *t*, *only* the token powers of a proper subset of the lower-level entity are identical to the token powers of the emergent feature **Coenz-Blood**. This implies that despite MR at the type level, at *t* the token entity is realised by one lower-level set of features. In the case of MD instead, there is only one token subset of powers that in a given time *t* realises the biochemical functions under discussion. A token biochemical function is emergent in that it has a proper subset of the token powers of chemical features. This makes the proper subset view straightforwardly compatible with multiple realisation and multiple determination, as discussed by Tahko (2020, 2021). Let us consider these them in more detail.

For MR, there may be several distinct token proper subsets of powers of the chemical features that can be associated with the biochemical function. In the case of **Coenz-Blood**, there are several distinct token proper subsets of the B12 vitamers that can be associated with the function and, as such, can realise the biochemical function under consideration. This is possible because, while the type is multiply realised, the token is always realised by a specific subset of token powers. For MD, two aspects should be considered. From the perspective of the token realised feature, one identifiable proper subset of chemical powers is associated with the higher-level feature, and, as such, MD is not problematic. From the multiply determinable feature perspective, instead, the token set of powers of a given chemical feature could present different proper subsets of token powers associated with different biochemical functions. Or, as suggested by Tahko 2020, there could be one proper subset of powers associated with two distinct type features, bringing in different effects in the relevant biological context. Accordingly, the token powers of the functional properties are a subset of those of a single chemical kind [Fig. 2].

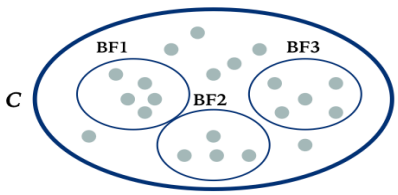


Fig. 2 – Multiple determinability of the cobalamin molecule, for which only one subset of powers is realised at a given *t*.

Moreover, the proper subset view is also able to deal with the reductionist view of MR for which it can be explicated in terms of a closed disjunction. This would make the biochemical functions reducible, and not emergent, to a closed disjunction of chemical structural powers. In this respect, Wilson discusses how the proper subset view ensures a form of ontological autonomy *contra* the disjunctive

strategy (2021). In the case of MR, when the entity H is weakly emergent, the token powers of H are a proper subset of the token powers of either L1 or L2. This makes H type different from the disjunction of Ls because of Leibniz's law: there are some powers of L that are not of H. Moreover, the nature of biochemical functions as defined here also allows to see how the defended view is compatible with MR and MD. The BF is associated with a set of powers whose selection is at least indirectly a result of evolution, and their causal efficacy is embedded in biological systems that are currently evolving. This has an impact on the fact that the types of realisers of the biochemical functions can change or increase in time. In addition to this, there could be a biologically possible world in which the biochemical function is realised by another chemical molecule yet unknown, or that does not play the function in current systems, but could have the function. This would make the disjunction an open disjunction, and, as such, challenges a straightforward reductionist approach.

In conclusion, the proper subset view and an account of weak emergence seem to be compatible with accounting for forms of MR and MD.

5. Conclusion

In this paper, I have considered how biochemical functions can be linked to chemical structure by using Wilson's account of weak emergence (2011, 2015, 2021). Section 2 introduced why the relation between structure and functions in biochemistry is interesting from a philosophical perspective and why can be seen as problematic. Section 3 focused on the double problem of biochemical function, the "function problem" and the "relation problem" offering further context to this debate. Section 4 then explored how, given a definition of biochemical functions, we can interpret the relation between chemical structure and biochemical structure via weak emergence. In doing so, I have considered how this framework offers us a way to think about the relation between structure and function that is compatible with multiple realisability and multiple determinability.

This paper has a series of interesting results. First, it enriches the case studies compatible with Wilson's account of weak emergence. This can bring in new insights relating to the emergence between entities that we would associate to the same level (Bellazzi, 2023). Second, it relates to one of the main research topics of biochemistry, the relation between biochemical functions and chemical structure. The account presented allows us to maintain a form of autonomy for biochemical functions while being compatible with the identification of the relation between structure and function. Third, the results of this paper contributes to the debates on unity of science and reductionism. In particular, they could be further explored to develop the our understanding of the interface between chemistry and biology, if we can establish a relation between the functional and chemical aspects of biochemical kinds.

References

Bartol, J. 2016, "Biochemical Kinds", *British Journal for the Philosophy of Science*, 67, 531-51.

Bellazzi, F. 2022, "Biochemical Functions", *British Journal for the Philosophy of Science* (waiting for an issue).

Bellazzi, F. 2023, *Biochemical Kinds and the Unity of Science*, PhD in Philosophy Thesis, University of Bristol, UK.

Combs, G.F. 2012, *The Vitamins: Fundamental Aspects in Nutrition and Health*, San Diego: Elsevier Science-Technology.

Fang, H., Kang, J., and Dawei, Z. 2017, "Microbial Production of Vitamin B12: A Review and Future Perspectives", *Microbial Cell Factories*, 16, 15, DOI: 10.1186/s 12934-017-0631-y

Gillett, C. 2013, "Constitution, and Multiple Constitution, in the Sciences: Using the Neuron to Construct a Starting Framework", *Minds & Machines*, 23, 309-37.

Goodwin, W. 2011, "Structure, Function and Protein Taxonomy", *Biology and Philosophy*, 26, 533-45.

Havstad, J.C. 2016, "Proteins: Tokens, Types and Taxa", in Kendig, C. (ed.), *Natural Kinds and Classification in Scientific Practice*, New York: Routledge, 74-86.

Havstad, J.C. 2018, "Messy Chemical Kinds", *British Journal for Philosophy of Science*, 69, 719-43.

Koury, M.J. and Ponka P. 2004, "New Insights into Erythropoiesis: The Roles of Folate, Vitamin B12, and Iron", *Annual Review. Nutrition*, 24, 105-31, DOI: 10.1146/annurev.nutr.24.012003.132306

Kistler, M. 2018, "Natural Kinds, Causal Profile and Multiple Constitution", *Metaphysica*, I, 19, 1, DOI: https://doi.org/10.1515/mp-2018-0006

Polger, T.W. and Shapiro, L.A. 2016, *The Multiple Realization Book*, Shapiro, L.A. (ed.), Oxford: Oxford University Press.

Santos, G., Vallejos, G., and Vecchi, D. 2020, "A Relational-Constructionist Account of Protein Macrostructure and Function", *Foundations of Chemistry*, 22, 3, 363-82.

Shapiro, L. 2020, "Theories of Multiple Realisation", *American Philosophical Quarterly*, 57, 1, 17-30.

Slater, M.H. 2009, "Macromolecular Pluralism", *Philosophy of Science*, 76, 851-63.

Tahko, T.E. 2020, "Where Do You Get Your Protein? Or: Biochemical Realization", *British Journal for the Philosophy of Science*, 71, 3, 799-825.

Tahko, T.E. 2021, *Unity of Science*, Cambridge Elements in Philosophy of Science, Cambridge: Cambridge University Press.

Tobin, E. 2010, "Microstructuralism and Macromolecules: The Case of Moonlighting Proteins", *Foundations of Chemistry*, 12, 1, 41-54.

Wilson, J. 2011, "Non-Reductive Realization and the Powers-Based Subset Strategy", *The Monist*, 94, 1, 121-54.

Wilson, J. 2015, "Metaphysical Emergence: Weak and Strong", in Wuthrich, T. (ed.), *Metaphysics in Contemporary Physics: Poznan Studies in the Philosophy of the Sciences and the Humanities*, Schöningh: Brill, 251-306.

Wilson, J. 2021, *Metaphysical Emergence*, Oxford: Oxford University Press.

Emergence, Exclusion, and the Proper Subset of Powers Strategy

Karen Bennett

Rutgers University

Abstract

Wilson characterizes weak and strong emergence partly based on their differing solutions to the exclusion problem. The weak emergentist should claim that emergent phenomena and their bases can both cause the same effect without overdetermining it, because they literally share causal powers. I compare this strategy with a different but related strategy also available to the weak emergentist, and argue that the virtues of the former cost more than it appears.

Keywords: Causal powers, Dependence, Emergence, Exclusion problem, Mental causation, Nonreductive physicalism, Overdetermination.

1. Introduction

Jessica Wilson's *Metaphysical Emergence* (2021) is an excellent and important book that brings together roughly twenty years of work on the ways in which one set of phenomena could be dependent on, and yet to some degree autonomous from, another set of phenomena. Wilson identifies the core shared ideas in the sea of mushy and contradictory usages of the term 'emergence', and articulates notions of 'weak' and 'strong' emergence that (in the philosophy of mind case) correspond to nonreductive physicalism and dualism respectively. She distinguishes these positions, in part, by how they approach the well-known exclusion problem for mental causation. Wilson's discussion of emergence and exclusion will be my focus in this commentary. What exactly does solving the exclusion problem require, and how exactly does her version of weak emergentism pull it off?

Before getting started in earnest, however, I would like to briefly call attention to a particular virtue of Wilson's book: its engagement with, and reliance upon, classic older work in the metaphysics of mind. She engages with a lot of material by people like Terence Horgan, Jaegwon Kim, Andrew Melnyk, Sydney Shoemaker, and Stephen Yablo. This is both appropriate and important, because a lot of excellent work in this area has been somewhat neglected of late. Both Wilson and I began our careers thinking about the mind-body problem, and are therefore well aware that the question of how some things give rise to other things

Argumenta 10, 1 (2024): 237—246
ISSN 2465-2334

DOI 10.14275/2465-2334/202418.ben
First published: 30 November 2024

is not exactly a new topic in metaphysics, as those in the contemporary grounding literature sometimes seem to suggest.

2. Weak and Strong Emergentism, Characterized by How They Handle the Exclusion Problem

Although terms like 'emergence' and 'emergentism' are used in many slightly different ways, Wilson argues that the most basic commitment of philosophical positions worthy of these labels is that emergent properties and states of affairs involve 'autonomy with dependence'. They are synchronically and non-causally dependent on their base, and yet *somehow or other* are autonomous from it: they have different causal powers, figure in different laws, or something along those lines.

That 'somehow or other' is, of course, crucial. Wilson distinguishes two primary forms of emergentism as meaning quite different things by the claim that emergent phenomena have 'different causal powers'. *Weakly* emergent features—if there are any—have fewer causal powers than the bases from which they arise, and *strongly* emergent features—ditto—have more causal powers then their bases. Wilson draws this distinction in the course of exploring available emergentist answers to the exclusion problem. It's a rather neat methodological trick: she simultaneously explains how these two kinds of emergence have different available responses to the exclusion problem, and uses their responses to the exclusion problem to shed light on the difference between them (Chapter 2).

Here's a simple version[1] of the exclusion problem, formulated as a set of five inconsistent claims:

Distinctness: Mental properties (and perhaps events) are distinct from physical properties (events).

Efficacy: mental events cause things, including physical things, and at least sometimes do so in virtue of their mental properties.

Completeness:[2] every physical effect has a sufficient physical cause.

Exclusion: all events that have multiple sufficient causes (that are not themselves causally related)[3] are overdetermined.

Nonoverdetermination: the effects of mental causes are not routinely and systematically overdetermined.

So, the physical effects of mental causes both are and are not systematically overdetermined. No bueno.

[1] The main way in which this version is simplified is that I merely gesture at how it can be run in either or both a property (type) or event version (token). Further, this is not how Wilson presents it. While the differences do not matter to anything of substance, footnotes 4 and 6 are worth reading.

[2] Most people, including Wilson, call this 'closure'. I prefer the label 'completeness', because the term 'closure' suggests that physical effects have *only* physical causes. That is an excessively strong premise that blocks the weak emergentist solution from the start.

[3] The parenthetical clause is there because the proper formulation of Exclusion ought not say that the outcome of a single, non-branching causal chain is overdetermined. If $c_1 \rightarrow c_2 \rightarrow c_3 \rightarrow e$, then e has multiple distinct sufficient causes but is not overdetermined by anyone's lights. An alternate way to circumvent this issue is to instead stipulate that the multiple sufficient causes be direct/unmediated.

One can of course dissolve the exclusion problem by denying that there are any mental phenomena, or claiming that they are epiphenomenal, or insisting that they are to be identified with the physical after all. But, as Wilson points out, these are not *emergentist* responses. They do not respect the core commitments that a) the mental is in some sense emergent (and thus exists), and b) emergent phenomena are in some sense causally autonomous (so mental events/properties are neither epiphenomenal nor identical to their physical bases).

So how should emergentists respond to the exclusion problem? Wilson claims that there are two and only two properly emergentist moves that can be made. The first is to deny Completeness, and claim that mental phenomena have genuinely novel causal powers that are neither determined by nor dependent on their physical bases. This strategy is non-physicalist, and is the distinctively strong emergentist position. The second solution is to deny Exclusion, and say that mental phenomena are causally efficacious and yet their effects are not overdetermined, or at least not overdetermined in the two-kids-simultaneously-throwing-two-rocks-at-a-window variety.[4] This is the weak emergentist or nonreductive physicalist (henceforth 'WE/NP') strategy.

The key WE/NP move is to appeal to an intimate relation short of identity, such as—to borrow Wilson's list (55-57)—functional realization, constitutive mechanism, mereological realization, the determinate-determinable relation, or 'superdupervenience'. (Though Wilson herself would wince (2014, 2018), we might replace some or all of those relations with grounding.)

I have long been fond of the WE/NP response to the exclusion problem, which I once called 'compatibilism'[5] (Bennett 2003). It will be the focus of the rest of the paper.

3. A "Deeper Unity of Strategy"? The Proper Subset Condition and the Counterfactual Condition

Wilson suggests that the fact that different WE/NPs appeal to different intimate non-identity relations is relatively unimportant as far as the exclusion problem is concerned, because

> underlying the seeming diversity in these and many other accounts of nonreductive physicalism hides a deeper unity of strategy (57).

[4] It is just a terminological matter whether we describe this move as saying that the effects of mental causes are not overdetermined at all, or as saying that they are not overdetermined in the bad 'double-rock' way. Discussions and defenses of the strategy take both forms in the literature. (See, e.g., Bennett 2003 and Sider 2003.) Wilson herself frames the strategy in the latter way, as "allowing that [the effects of mental causes] are overdetermined […] but maintain[ing] that the overdetermination here is of an unproblematic *non*-double-rock-throw variety" (44). Characterized like that, the move denies Nonoverdetermination rather than Exclusion: the effects of distinct causes are always overdetermined, but it turns out that overdetermination is more widespread and less troublesome than usually thought.

I prefer the characterization in the main text, which reserves the word 'overdetermination' for the double-rock-style cases. It is also the better characterization for Wilson herself. See note 6.

[5] I called it that because it says that the non-overdeterministic causal efficacy of the mental is compatible with the conjunction of Completeness and Distinctness.

I agree that there is a deeper unity of strategy here. Indeed, I have argued that there *must* be a deeper story, in the sense that the WE/NP ought not simply name an intimate non-identity relation, and announce that events related in that way do not overdetermine their effects. That is not good enough. What is required is a story about how and why that relation has that kind of impact:

> the burden is on the compatibilist here. She needs to be able to *argue* that the effects of mental causes are not overdetermined, and to explain *why* they are not (2003: 474).

That is, in essence, what Wilson is after when she claims a "deeper unity of strategy". She is saying that all of the tight relations postulated by the WE/NP lend themselves to a particular sort of explanation: what I hereby dub the "Proper Subset Strategy".

While I clearly agree about the need for *some* kind of deeper explanation, I am not convinced that the Proper Subset Strategy is the right one. An alternative is available whose relative merits must also be investigated. After sketching both Wilson's story and this alternative, I will explore the relation between them, and argue that the apparent virtues of the Proper Subset Strategy cost more than it seems.

4. Wilson's Proposed Underlying Idea: The Proper Subset Strategy

Wilson claims that whenever one phenomenon *E* is weakly emergent from a base phenomenon *B*, *E*'s causal powers will be a non-empty proper subset of *B*'s. In particular, when mental and physical phenomena stand in any of the close relations posited by the WE/NP, it will be the case that mental phenomena have a non-empty proper subset of the causal powers of the physical phenomena from which they weakly emerge (58-66). Thus the various particular mechanisms for securing weak emergence "are unified in each [endorsing the Proper Subset Strategy] as a means of avoiding problematic overdetermination" (66).

The Proper Subset Strategy certainly sounds good. Indeed, it sounds like it decisively solves the exclusion problem. The picture is that mental and physical causes do not overdetermine their effects because there is a *literal shared core* of causal juice: to say that mental phenomenon *M* and its physical base *P* overdetermine their effects would be wrong in the same way that it would be wrong to say that our two favorite hooligans, Billy and Suzy, overdetermine the breaking of the window by holding hands and jointly throwing one single mutually-owned rock. It is wrong in the same way that it would be to say that you and I double-pay the bridge toll by together tossing in one $5 bill from our shared piggybank, or that it would be to say that there are two winners of the local 5K, the Johnson family and the García family, because Inez García-Johnson won it. In none of these cases is there any genuine doubling. The window's breaking has just one proximate cause; the 5K has just one winner; the bridge toll has been paid only once. Exclusion begone![6]

[6] Now it can be seen that it is less than optimal for Wilson to characterize the WE/NP solution to the exclusion problem as saying that the effects of mental causes *are* overdetermined, but not in the bad double-rock way—that is, as denying Nonoverdetermination

Unfortunately, this is all a bit of legerdemain. But before I explain why, I need to put the alternative on the table.

5. An Alternative Underlying Idea: The Counterfactual Strategy

Talk of overlapping sets of causal powers is not the only way to explain how various intimate relations between the causes defuse the threat of overdetermination. In a 2003 paper, I offered a different explanation. I provided a necessary condition on overdetermination (genuine, 'double-rock' overdetermination), and argued that it is not met by pairs of causes related in any of the ways WE/NPs think that mental and physical phenomena are.[7]

The necessary condition is simply that two causes overdetermine an effect only if had either happened without the other, the effect would still have occurred.[8] That is, causes c_1 and c_2 overdetermine e only if both of the following counterfactuals are nonvacuously true:

$(c_1 \,\&\, \sim c_2) \,\Box\!\!\rightarrow e$

$(c_2 \,\&\, \sim c_1) \,\Box\!\!\rightarrow e$

This is a very intuitive test for overdetermination. We implicitly rely on it whenever we distinguish between overdetermination and joint causation. Indeed, note that those who would appeal to modal fragility to claim that all apparent overdetermination is really joint causation implicitly rely on these counterfactuals.[9]

Yet if the test is legitimate, the WE/NR is again in good shape. At least one of these counterfactuals will be vacuous or false when (2003) and only when (2008) the mental and physical causes stand in one of the WE/NR's favored relations. Though the details get too complicated to revisit here, the basic idea is that on any such relation, the physical base necessitates the weakly emergent mental phenomena, rendering one of the counterfactuals vacuous.

6. The Relation Between the Two Strategies

Two ways of explaining why the existence of certain tight relations falsifies Exclusion are now on the table. Each strategy offers a necessary condition on overdetermination—one, that certain counterfactuals be nonvacuously true; the other, that the two potential causes not be such that one's set of causal powers is a proper subset of the other's—and claims that weakly emergent phenomena and their

rather than as denying Exclusion. (See note 4). Given the Proper Subset of Powers strategy, she should not think that the effects of mental causes are overdetermined *at all*. For an effect to be overdetermined, it must have at least two distinct causes. But the only sense in which Wilson's WE/NP thinks there are two distinct causes is that there are two distinct phenomena that literally share the efficacious part.

[7] Really, in any of the ways *any* physicalist thinks they are: identity works too.

[8] This is not supposed to be an analysis of overdetermination in noncausal terms, just a condition on which causes count as overdeterminers.

[9] Billy and Suzy throw separate rocks, apparently overdetermining the breaking of the window. The fan of the fragility treatment of such cases (Lewis 1986, 2000) would say, "look I know it seems like the window would still have broken if only Billy threw his rock, or only Suzy threw hers. But that's not actually true, because the precise time and manner of the breaking are essential to it. If only one of them had thrown, it would not have been the very same break. So you're wrong about those counterfactuals. The particular window-breaking that actually happened required both Billy and Suzy to throw their rocks".

bases do not meet the condition, and thus do not overdetermine their effects. Here is a bit more about the relation between these two conditions.

First, the failure of the causal powers to nest in a subset relation does not entail that the overdetermination counterfactuals are nonvacuously true. There are at least two reasons for this. One is that someone who denies that there are any such things as causal powers, or that (foreshadowing!) they are the kinds of countable things that can form sets, will deny that any pairs of events are such that their causal powers nest in the relevant way. But such a person is not committed to thinking that all overdetermination counterfactuals, formulated with whatever pair of events you like, are nonvacuously true. Another reason is the case in which c_1 and c_2 share a lot of causal powers, but not all of them; the two sets overlap but neither is a subset of the other. It could still be the case that one or both of the overdetermination counterfactuals is false or vacuous, for example if the non-shared causal powers are irrelevant to the particular effect in question.

What about the other direction? Does the nonvacuous truth of the overdetermination counterfactuals entail that the causal powers fail to nest in a subset relation? Equivalently, does the subset-nesting of the causal powers entail that at least one of the corresponding overdetermination counterfactuals is false or vacuous? It is tempting to say yes, but matters are somewhat tricky.

Suppose that c_1's causal powers are a proper subset of c_2's, and that c_1 and c_2 are both actual causes of e. It is likely nonvacuously true that if c_1 had happened without the 'larger' c_2, the effect would still have happened. The interesting question is whether e would still have happened if c_2 had happened without the 'contained' c_1. The difficulty in assessing the counterfactual is that the mere claim that c_1's causal powers are a proper subset of c_2's says nothing about the modal status of that inclusion, nor about whether either event has any or all of those token causal powers essentially. The whole shebang could be contingent. And that makes it difficult to mount a decisive case for the falseness or vacuity of the overdetermination counterfactual $(c_2 \,\&\, \sim c_1) \,\Box\!\!\rightarrow e$. The options are that a) c_2 cannot happen without c_1, in which case the counterfactual is vacuous, b) c_2 *can* happen without c_1, and indeed with c_1 and all its causal powers deleted completely, in which case the same counterfactual is probably false, and c) c_2 can happen without c_1 in particular, but only if c_1's causal powers are replaced by numerically different but qualitatively similar ones (in the way that an object might survive the replacement but not complete loss of a part). In *that* case, the counterfactual is probably nonvacuously true, despite the 'subsetting'. And this is the most likely case in the situation at hand: where c_2 weakly emerges from c_1, via any of the standard WE/NP relations. Maybe this mental state could happen without this *particular* physical state that underwrites it, but it cannot happen without *any* physical basis.

Now, I do not want to rest a lot of weight on this. I myself have argued that these kind of 'replacement' interpretations of counterfactuals are problematic (2003: 482), and David Lewis seems to agree (2000: 190). My only point here is that the path from causal-power-subsethood to the falseness or vacuity of the overdetermination counterfactuals is neither obvious nor straightforward. Given the entailment failure in the other direction, it is probably best to think of the two strategies as independent. Two events that vacuify or falsify the counterfactuals need not meet the Proper Subset Condition, and it may well be that two events that meet the Proper Subset Condition can fail to vacuify or falsify the counterfactuals.

7. The Proper Subset of Strategy Is Not More Powerful than the Counterfactual Strategy

I have sometimes thought that the Proper Subset Strategy is a more powerful (groan) implementation of the Counterfactual Strategy. (Both appeared in print at roughly the same time: e.g. Wilson, 1999, 2002; Shoemaker 2001, 2003; Bennett, 2003.) I have come to think that this is wrong. The previous section shows that it isn't clearly right to think of the Proper Subset Strategy as an implementation of the Counterfactual Strategy. And although there is a clear case to be made for the claim that it is more powerful, in two specific senses, this advantage is an illusion.

The first sense in which the Proper Subset Strategy seems to be more powerful than the Counterfactual Strategy is that it appears to provide a deeper, more convincing explanation of why there is no overdetermination. Recall the examples of the bridge toll, the 5k, and the hand-holding hooligans: the weak emergentist gets to similarly claim literally shared causal power. In contrast, the Counterfactual Strategy just says something kind of wishy-washy about the truth-values of certain counterfactuals, while remaining silent about *why* those counterfactuals have the truth-values they do.

The second way in which the Proper Subset Strategy seems to be more powerful than the Counterfactual Strategy is that it not only shows that the weakly emergent entities and their bases can both be causally efficacious without overdetermining their effects, but also shows that weakly emergent phenomena are causally efficacious in the first place. If such phenomena have a nonempty proper subset of the causal powers of their bases, then *a fortiori* they have causal powers.[10] The Counterfactual Strategy, in contrast, does not do this. It simply *assumes* that the mental is causally efficacious, and shows that this (together with Distinctness and Completeness) does not entail that the effects of mental causes are systematically overdetermined.

Unfortunately, these two seeming advantages are just that: mere seemings. There is little substance to either point, which I will address in reverse order.

First, a solution to the exclusion problem that establishes the causal efficacy of the mental, or the weakly emergent more generally, is actually not superior to one that does not—at least, not *qua* solution to the exclusion problem. The exclusion problem is an attempt to undermine the causal efficacy of the mental (the emergent), not because of any intrinsic defect, but rather because there is no causal work for it to do.[11] An adequate response to the exclusion problem is simply one that undercuts this reasoning. My point here is just the elementary one that objecting to an argument that $\sim p$ does not require showing that p is *true*. Thus the fact that the Proper Subset Strategy secures the causal efficacy of the mental does not add anything *qua response to the exclusion problem*.

[10] Wilson admits that nothing she says gives the weakly emergent phenomena *novel* efficacy (58, 67-69), but she is right to accept this consequence. It's what makes weak emergence different from strong emergence. No nonreductive physicalist, for example, should grant causal powers to the mental that aren't possessed by its physical base.

[11] Contrast, for example, Princess Elisabeth-style complaints about substance dualism, where the problem is that the mental is not spatially located, has no mass, has no chemical structure, and so forth.

Of course, this does not mean that it is no advantage at all to the Proper Subset Strategy. It could solve the exclusion problem *and* secure the causal efficacy of the mental. But I am still skeptical; I do not think the strategy actually does secure that. All the work is done by Wilson's claim that weakly emergent entities have a *non-empty* proper subset of the causal powers of their bases. This is the only reason we are guaranteed that weakly emergent entities have causal powers. But Wilson never argues that any particular thing or kind of thing has a non-empty set of causal powers; that is just part of her definition of weak emergence. So those who are inclined to be worried about the causal efficacy of the kinds of phenomena she takes to be weakly emergent—like the mental—will simply deny that they are weakly emergent in her sense.

Second, I also doubt that the Proper Subset Strategy truly provides a deeper, more convincing explanation of why there is no overdetermination—no "causal competition" as Yablo puts it (1992). It looks like it does, yes, but, well, that is the nature of prestidigitation.

The problem is that the deeper explanation requires being quite literal about something that it is not so easy to take literally. The way the Proper Subset Strategy so cleanly escapes overdetermination is by *identifying* each and every causal power of the weakly emergent phenomenon with a causal power of the base phenomenon. As Wilson has emphasized since she began defending the view (1999, 2002), it is crucial that each individual causal power of the emergent thing be possessed by both.

To bring this out clearly, consider two similar but hopeless positions that result from removing the 'subset' part from the Proper Subset Strategy. One position simply says that weakly emergent phenomena have fewer causal powers than their bases. This is no help with exclusion at all; a rock presumably has fewer causal powers than a similarly sized iPhone—for example, only the latter can call an Uber—but throwing both can certainly overdetermine the breaking of a window. The second hopeless position says not only that weakly emergent phenomena have *fewer* causal powers than their bases, but also that their causal powers are *qualitatively indiscernible* from those of their bases. But this again is no help with the exclusion problem. Events with non-identical but qualitatively indiscernible causal powers can absolutely overdetermine things. Consider a scenario in which Billy and Suzy stand 5 feet from each other and throw two indiscernible rocks in indiscernible ways at the window, hitting *almost* the same spot with the same force, at the same angle, at the same time. Their rock-throwings share almost all their causal powers at the type level. (That is, the vast majority of the causal powers belonging to Billy's throw are qualitatively indiscernible from those belonging to Suzy's throw.) But the causal powers of the two events are not numerically identical, and their breaking the window is, again, an uncontroversial case of overdetermination.[12]

In short, the success of the Proper Subset Strategy entirely depends on the idea that the causal powers of the emergent phenomena are numerically identical to the causal powers of the base. And this in turn requires that token causal powers

[12] At this point, one might move to the idea that the causal powers of the base *constitute* or *realize* the distinct but qualitatively indiscernible causal powers of the weakly emergent phenomena. This is basically Derk Pereboom's view (2002, 2011). Whatever its merits, it does not avail itself of the Wilson-Shoemaker idea that there is a shared core of causal power.

are the sort of thing that can not only be *counted* but also *individuated*. Indeed, it is very, very hard not to imagine them as pebbles in a bucket—and Wilson's diagrams on page 70 suggest that she cannot resist this picture either. But this is a serious and rather discombobulating ontological commitment. I will not argue here that causal powers are not like that, but I suspect others will share my reticence. Even Wilson takes pains to insist that her causal powers are nothing dubious or creepy:

> Talk of powers is simply shorthand for talk of what causal contributions possession of a given feature makes […] to an entity's bringing about an effect, when in certain circumstances […] no controversial theses pertaining to the nature of powers, causation, properties, or laws are here presupposed (32-33; also 45).

But the question is, can she really make good on this neutrality? More precisely, can she assuage my ontological qualms while retaining the nice claim that strictly speaking, there is really only one cause of an effect caused both by a weakly emergent phenomenon and its base? That is the challenge I lay before her.

Let me be crystal clear: I have not argued that she cannot meet this challenge. I have simply *raised* the challenge. My real point here is that one cannot have the Proper Subset Strategy on the cheap; the cost-benefit analysis must be made. We can shoulder the ontological commitment to trackable, countable causal powers and accept the benefits, or we can be squeamish and reject the whole picture. What we cannot do is help ourselves to the lovely solution to the exclusion problem while acting as though it costs no more than simply believing in causation. When I accuse the Proper Subset Strategy of sleight of hand, that is what I really mean: not that it cannot fulfill its promise at all, but rather that it hides the expensive machinery required to do so. Regardless, I have appreciated the opportunity to drill deeper into it than I previously have, and discover its secrets.[13]

References

Bennett, K. 2003, "Why the Exclusion Problem Seems Intractable, and How, Just Maybe, To Tract It", Noûs, 37, 471–497.

Bennett, K. 2008, "Exclusion again". In J. Howhy & J. Kallestrup (eds.). Being Reduced, 280–305, Oxford: Oxford University Press.

Lewis, D. 1986, "Postscripts to 'Causation'". In his Philosophical Papers, Volume II, 172–213.

Lewis, D. 2000, "Causation as Influence", The Journal of Philosophy, 97, 182–197.

Pereboom, D. 2002, "Robust Nonreductive Materialism", The Journal of Philosophy, 99, 499–531.

Pereboom, D. 2011, Consciousness and the Prospects for Physicalism, Oxford: Oxford University Press.

Shoemaker, S. 2001, "Realization and Mental Causation", in Identity, Cause, and Mind: Philosophical Essays, Oxford: Oxford University Press, 427-451.

[13] Thanks to Jessica Wilson and the audience at the 2023 Eastern APA Author-Meets-Critics for discussion. I was tempted to somehow work the phrase "metaphysical emergency" into the paper, but I resisted. You're welcome.

Shoemaker, S. 2003, "Realization, Microrealization, and Coincidence", Philosophy and Phenomenological Research, 67, 1-23.

Sider, T. 2003, "What's so Bad About Overdetermination?", Philosophy and Phenomenological Research, 67, 719-726.

Wilson, J. 1999, "How Superduper Does a Physicalist Supervenience Need to Be?", The Philosophical Quarterly, 49, 33-52.

Wilson, J. 2002, "Causal Powers, Forces, and Superdupervenience", Grazer Philosophische Studien, 63, 53-78.

Wilson, J. 2014, "No Work for a Theory of Grounding", Inquiry: An Interdisciplinary Journal of Philosophy, 57, 535-579.

Wilson, J. 2018, "Grounding-Based Formulations of Physicalism", Topoi, 373, 495-512.

Wilson, J. 2021, The Metaphysics of Emergence, Oxford: Oxford University Press.

Yablo, S. 1992, "Mental Causation", The Philosophical Review, 101, 245-280.

A Mereology for Emergence

Claudio Calosi

University of Venice

Abstract

The paper first investigates the tension between reductive accounts of mereological structure and emergence as characterized in Jessica Wilson's seminal work. It then suggests a new mereology for emergence. Finally, the resulting account is applied to a paradigmatic case of an emergent whole.

Keywords: Emergence, Mereological structure, Mereological sum, Matter.

To my partner in crime, J.W.

1. Emergence and Mereological Reductionism

There are several broadly "reductive" accounts of mereological structure. They all try to capture rigorously the somewhat vague intuition that "wholes are nothing over and above their parts". The most radical view in the reductive camp holds that mereological composition is strict numerical identity, in that wholes are numerically identical to their parts considered collectively. The view is known as *Strong Composition as Identity*. Using double signs (such as xx), for plural terms:[1]

Strong Composition as Identity (CAI): If the xx compose y, then $xx = y$.

There is a famous argument in the literature against CAI from the possibility of emergence.[2] It goes roughly as follows. If CAI is true, then wholes cannot have properties that the plurality of their proper parts do not have. Emergent properties are exactly an example of such properties. Hence, if (possibly) there is emergence, CAI is false. Whatever one thinks of the argument, CAI is indeed a radical option. For example it might require substantive changes in the logic of identity and/or comprehension principles of plural logic. Hence, it is important to realize that the tension between reductive accounts of mereological structure and (the possibility of) emergence cuts a little deeper. As Wilson (2021) puts it,

[1] For an introduction see Baxter and Cotnoir 2013.
[2] See e.g., McDaniel 2008, Schaffer 2010, Sider 2013, and Calosi 2016.

Argumenta 10, 1 (2024): 247—255
ISSN 2465-2334

DOI 10.14275/2465-2334/202418.cal
First published: 30 November 2024

> It is the *coupling of cotemporal material dependence* with *ontological* and causal *autonomy* which is most *basically definitive of the notion of emergence*, at least as suggested by the central cases of special-science entities with respect to the physical micro-configurations which are their constant companions (Wilson 2021: 1; italics added).

In the light of this, the general threat coming from emergence to reductive accounts of mereological structure is the following. If emergent wholes are *ontologically autonomous* from their (microscopic) constituents,[3] then they are indeed "something over and above" those constituents, contra the spirit, not just the letter, of reductive accounts. It is not my purpose here to respond to the threat, nor to dissect its presuppositions. Rather, it is to take such a threat at face value and propose a new mereological system that vindicates the claim that "wholes are something over and above their parts"—as seems to be required by metaphysical emergence. This is by no means an easy task. Indeed, many think that mereology alone is not enough to account for complex, highly structured, emergent wholes. This is why they recommend different forms of hylomorphism.[4] Others think that we need to revisit the very mereological framework we use, for example adopting a so-called slot-mereology,[5] or rejecting mereological monism, roughly the view that there is only one notion of (mereological) part.[6] I am going to suggest a mereological account that uses only one notion of parthood. In a nutshell, I am going to suggest that we can define a notion of mereological sum that is not equivalent to extant ones in the literature. Given anti-symmetry of parthood, it turns out that sums are unique. I then define the notion of the matter of an entity as *the* sum of its proper parts. This helps me draw a distinction between *Reducible Wholes*, wholes that are nothing over and above their matter, and *Irreducible Wholes*, wholes that are distinct from their matter. Finally, I suggest that if a whole is an emergent whole, then it is an irreducible whole—as previously defined.[7]

2. A New Mereology

There are three notions of mereological sum in extant literature.[8] I will use $<$ for parthood, $\ll$ for proper parthood, $\circ$ for overlap, defined as usual, and $\prec$ for the plural logic relation of "being one of".[9] For the sake of readability "$xx < y$" abbreviates "$\forall x(x \prec xx \rightarrow x < y)$", and "$x \circ xx$" abbreviates $\exists y(y \prec xx \wedge x \circ y)$". Then the usual notions of sum are defined as follows:

D. 1 $Sum_1(xx, y) \equiv \forall x(x \circ y \leftrightarrow x \circ xx)$ SUM_1

D. 2 $Sum_2(xx, y) \equiv xx < y \wedge \forall x(x < y \rightarrow x \circ xx)$ SUM_2

D.3 $Sum_3(xx, y) \equiv xx < y \wedge \forall x(xx < x \rightarrow y < x)$ SUM_3

[3] I follow Wilson (2021: 10) here. Roughly, an emergent whole is a whole with an emergent feature.
[4] See e.g., Koslicki 2008, Fine 2010, and Sattig 2015.
[5] See e.g., Bennet 2013 and Sattig 2021.
[6] See e.g., Canavotto and Giordani 2020.
[7] I developed the technical work on the new mereological system together with Alessandro Giordani. See Calosi and Giordani 2023a, and Calosi and Giordani 2023b.
[8] See Cotnoir and Varzi 2021.
[9] That is, $x \ll y \equiv x < y \wedge x \neq y$, and $x \circ y \equiv \exists z(z < x \wedge z < y)$.

In plain English, y is a Sum_1 of the xx iff it overlaps all and only the things that the xx overlap, y is Sum_2 if every xx is part of y and every part of y overlaps the xx, and finally, y is a Sum_3 iff every xx is part of y, and everything that includes the xx includes y. It is well-known that in mereologies that are weaker than classical mereology, the three notions are not equivalent.[10] Do they exhaust the notions of Sum definable in terms of $<$ and $\prec$? Hardly so. Consider the following:

D.4 $Sum(xx, y) \equiv xx < y \wedge \forall x(\neg xx \circ x \rightarrow \neg x \circ y)$
$\wedge\ \forall x(xx < x \rightarrow y < x)$ SUM

Definition **D.4** simply says that y is the Sum of the xx iff (i) the xx are part of y, (ii) whatever is disjoint from the xx is disjoint from y, and (iii) everything that includes the xx includes y. In other words, according to (i), the mereological sum of a plurality should be inclusive enough to count every member of xx as a part. According to (ii), it should be no more inclusive than that. Finally, according to (iii), a mereological sum should be minimal, in that it has to be part of everything that includes the original plurality. It is easily seen that, in the absence of strong mereological principles we have (1) and (2) below, where i ranges over the three notions of sum in **D.1-D.3**:

(1) $Sum(xx, , y) \rightarrow Sum_i(xx, y)$
(2) $Sum_i(xx, y) \nrightarrow Sum(xx, , y)$

Thus, Sum is strictly stronger than any Sum_i. Once we have such a stronger notion of Sum, we can put forward an explicit mereological system based on that notion.[11] For the sake of simplicity, I am going to require a very strong principle for the existence of Sum-s. In particular I am going to require a counterpart of the *unrestricted composition* principle of classical mereology.[12] It should be noted however that weaker principles will do as well. I will return to this shortly. Here is the system:

P. 1 $x < y \wedge y < x \rightarrow x = y$ ANTISYMMETRY
P. 2 $x < y \wedge y < z \rightarrow x < z$ TRANSITIVITY
P. 3 $x \ll y \rightarrow \exists w \exists z(w \ll y \wedge z \ll y \wedge \neg w \circ z)$ QUASI-SUPPLEMENTATION
P. 4 $x \prec xx \rightarrow \exists y\big(Sum(xx, y)\big)$ UNRESTRICTED SUM

Let us define "being mereologically simple" and being "mereologically composite" as usual:

D. 5 $S(x) \equiv \neg \exists y(y \ll x)$ SIMPLE
D. 6 $C(x) \equiv \neg S(x)$ COMPOSITE

It is an interesting feature of the system, and one that is crucial for the present argument, that we have extensionality of Sum, in that Sum-s are unique, but we do not have extensionality of proper parthood. That is, (3) below is a theorem but (4) is not:

(3) $Sum(xx, y) \wedge Sum(xx, z) \rightarrow y = z$
(4) $C(x) \vee C(y) \rightarrow \big((z \ll x \leftrightarrow z \ll y) \rightarrow x = y\big)$

It remains to be seen how this relates to emergence. I now turn to that.

[10] See Cotnoir and Varzi 2021.
[11] This is the system we analyze in detail in Calosi and Giordani 2023b.
[12] Note that REFLEXIVITY ($x < x$) follows.

3. The Account

Given UNRESTRICTED SUM and theorem (3) we can define a total function over the domain of concrete objects that assign to each concrete object its *matter*.[13] More precisely, letting xx be the plurality of proper parts of x, we define the matter of x, $m(x)$ as x if x is simple, and as the Sum of the xx if x is composite:

D. 7 $S(x) \rightarrow m(x) = x$ SIMPLE-MATTER

D. 8 $C(x) \rightarrow m(x) = \iota z\big(Sum(xx, z)\big)$ COMPOSITE-MATTER

Now we can distinguish those objects that are identical to their matter and those that are not. I call the first REDUCIBLE WHOLES, the second IRREDUCIBLE WHOLES:[14]

D. 9 $R(x) \equiv x = m(x)$ REDUCIBLE-WHOLE

D. 10 $I(x) \equiv x \neq m(x)$ IRREDUCIBLE-WHOLE

Intuitively, this distinction corresponds to the distinction between objects that are nothing over and above their parts, such as e.g., heaps of sands, and objects that are something over and above their parts, e.g., complex structured objects such as table, trees, organisms, statues. The following are immediate consequences:

(5) $S(x) \rightarrow R(x)$

(6) $I(x) \rightarrow C(x)$

None of the converses hold. As a way of illustration, consider the following model, where $\oplus$ is simply "binary Sum":[15]

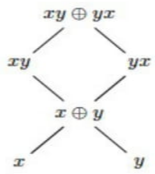


Figure 1: A Model with Reducible and Irreducible Wholes

In the model above $x \oplus y$ is a reducible whole, which is the matter of two irreducible wholes with reducible proper parts, namely xy, and yx, and the matter of a reducible whole with irreducible parts, namely $xy \oplus yx$. It should be clear why the present proposal has a chance to provide a mereology for emergent wholes: it allows for irreducible wholes that are something over and above

[13] As I pointed out before, I require **P.4** only for the sake of simplicity, but it is unnecessarily strong. All the following arguments require is an existence axiom for *Sum*-s that guarantees that the matter of every entity exists. There are different principles that are (i) are compatible with this requirement, and (ii), weaker than **P.4**.

[14] This mirrors the distinction between *unstructured* and *structured* entities in Calosi and Giordani 2023a.

[15] In Calosi and Giordani 2023a we suggest this is how to account for the infamous case of the composition of a syllable in Aristotle's *Met. Z.*

their proper parts, i.e., their matter. Indeed, I suggest that, faced with cases of emergent wholes (E) we should endorse the following conditional:

(7) $E(x) \rightarrow I(x)$

IRREDUCIBILITY as defined above is a necessary condition for emergence. I want to stay neutral as to whether the converse holds. Indeed, I am more hesitant to subscribe to irreducibility being sufficient for emergence. Perhaps there are other "grounds" for irreducibility. Why should one hold that emergent wholes are irreducible in the precise way I defined them? To answer this question, note that we can extract different broad conditions a mereology for emergent wholes needs to meet from the account of emergence in Wilson 2021. Irreducibility in this precise sense helps meeting this requirement. We saw the first (conjunctive) requirement already:

> *Dependence and Autonomy*: Emergent wholes are somewhat dependent on their parts, but at the same time somehow ontologically autonomous from them.[16]

In Wilson's words:

> Summing up: many considerations, drawn from science, perception, language, our practices of individuation, and introspective experience, provide prima-facie support for thinking that many broadly natural entities are co-temporally materially dependent on micro-configurations of fundamental physical entities, yet are also ontologically and causally autonomous with respect to these underlying micro-configurations (Wilson 2021: 6-7).

> *Compositional Flexibility*: The existence of an emergent whole depends on the existence of its parts but does not depend on the existence of any specific plurality of proper parts.[17] In effect, the emergent whole is usually taken to be capable of surviving (some) changes in mereological structure—see e.g., Wilson 2021: 6.
>
> *Sortal Properties of Ordinary Objects*: Some emergent wholes, in particular *ordinary objects*, fall under "sortal features" that do not apply to any collection of proper parts of said wholes and are responsible for their persistence conditions.[18]

To quote Wilson again:

> Candidate sortal features for ordinary objects of the varieties at issue here would be feature expressing membership in the category at issue, such as 'being a table' or 'being a statute' (Wilson 2021: 197).

[16] Wilson (2021) discusses several suggestions to cash out precisely both the *dependence* and the *autonomy* aspects. I will not enter these details here.

[17] It is an interesting question whether this distinction Wilson draws parallels the one in e.g., Simons 1987 between *generic* and *rigid* dependence. My inclination is that both Simons and Wilson are after the same distinction. But the devil is in the details, and I am not sure Wilson would buy the *analysis* of dependence that Simons (1987) puts forward.

[18] Wilson dedicates the entire Chapter 6 to such objects, arguing that they provide an example of Weak Emergence. For Weak Emergence, see Wilson 2021, especially Chapter 3.

As I pointed out already, I want to make a case for the following claim: the mereological system I proposed helps in satisfying all the desiderata above. Consider *dependence*. According to (7), every emergent whole is an irreducible whole, that is, a whole that is distinct from its matter. But note that the *matter* of an irreducible whole is a very *sui-generis* proper part of that whole. In particular it its *only maximal, unsupplemented proper part*. By this I simply mean that every other proper part of the emergent whole is a proper part of its matter, and therefore overlaps its matter. This captures an important sense in which every irreducible whole *depends* on its matter: were we to annihilate its matter, it is unclear that anything would remain of the whole. Note that it is exactly this kind of considerations that are usually taken to be a litmus test for dependence. At the same time, an irreducible whole is *distinct* from its matter. Now, I grant that numerical distinctness is not sufficient for *autonomy*, but I submit, it is at least *necessary*. What about *compositional flexibility*? There is a raging debate over whether mereological sums can undergo mereological changes. But irreducible wholes are exactly those wholes that are *not Sum*-s. Whatever stance one takes on the possibility of Sums of surviving mereological changes, this does not affect the possibility of irreducible wholes to survive such changes. Indeed, the model in Figure 1 shows that different irreducible wholes, such as xy and yx, can have the same matter. Granted, this does not show that the same irreducible whole can have a different matter at different times. Unfortunately, to provide a detailed account of such possibility, one would need to dive deep into the metaphysics of persistence. I cannot do justice to such a project here. I rest content at pointing out that the very distinction between irreducible and reducible wholes provides a leeway to account for both compositional dependence and compositional flexibility. Finally, *sortal properties*. The thought here is that once the distinction between an irreducible whole and its matter is in place, one can simply claim that the relevant sortal property such as e.g., "being a statue" applies to the irreducible whole but not to its matter. The case of the statue is indeed instructive. Let me contrast here the analysis provided by the account I put forward in the paper with another account, that is more familiar in the mereological literature. My contention is that the new account is a better fit with metaphysical emergence.

As we saw in §1 emergent wholes seem to be "something over and above their parts" in virtue of their ontological autonomy. The familiar way of cashing out this proposal in the mereological literature is to endorse a *non-extensional* mereological system, that is, a mereological system that does not have (4) among its axioms or theorems. The system we are investigating is one example. But there are others. Arguably, the most popular one since at least Simons 1987 is the one that endorses Sum_1 as its notion of sum, has **P.1** and **P.2** as its axioms, and replaces **P.3** and **P4** with the following respectively:[19]

P. 5 $x \ll y \rightarrow \exists z(z < y \wedge \neg x \circ z)$ WEAK SUPPLEMENTATION

P. 6 $\exists x(x \prec xx) \rightarrow \exists z\big(Sum_1(z, xx) \leftrightarrow \varphi(xx)\big)$ RESTRICTED-COMPOSITION

[19] But there are many others. For an introduction see Cotnoir 2013. One needs restricted composition because Weak Supplementation and Transitivity, together with Sum_1, yield (3) as a theorem. See Varzi 2009.

Importantly, in this mereological system Sum_1-s are not unique. That is, (3) is not a theorem of the system. Now, suppose we have a statue, call it $Statue$, that is made out of a lump of clay, call it $Lump$, that has two parts, $Lefty$ and $Righty$. According to the more familiar mereological account $Lefty$ and $Righty$ have two Sum_1-s, namely $Statue$ and $Lump$, as in Figure 2 below:

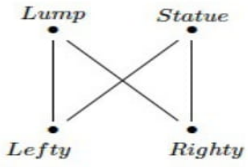


Figure 2: Statue and Lump: Part I

The thought here is that wholes are something over and above their parts in that the existence of proper parts does not determine the identity of the whole. Indeed, different wholes can share the same proper parts. But note that, from a purely mereological perspective, both $Lump$ and $Statue$ are Sum_1 of $Lefty$ and $Righty$. And yet, in the present context, only one of them is an (alleged) emergent whole with a distinguished sortal property such as "being a statue". It seems clear that the mereological structure of the Sum-s cannot account for the difference of the metaphysical status of the wholes with respect to emergence. The mereological system I discussed handles things much differently—and, I contend, better. In the case at hand, there will be only one Sum of $Lefty$ and $Righty$, namely $Lump$ which is a reducible, hence non-emergent whole. $Lump$ is the *matter* of $Statue$ which is a distinct, irreducible emergent whole, as per Figure 3:

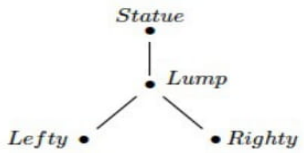


Figure 3: Statue and Lump: Part II

Here, the difference between the composite objects $Lump$ and $Statue$ is reflected in the mereology so to speak. $Lump$ is a Sum, and therefore a reducible object. By contrast $Statue$ is not a Sum. It is something over and above its matter—$Lump$—and this is why the emergent sortal property "being a statue" only applies to *it*. This is reason enough to prefer the mereological system I suggested to the one that is more familiar from the literature, at least if one maintains that statues are emergent wholes distinguished by their emergent (sortal) properties.

4. An Application

Beside ordinary objects and artifacts, Wilson (2021) suggests that special-sciences entities might be (at least weakly) emergent. For instance, she writes:

> Special-science entities are characterized as having distinctive features, constitutive of the distinctive types under which they fall. A tree, for example, has roots, a trunk, branches, stems, leaves; it obtains nutrients from air, sun, soil, and water through leaves and roots; it reproduces via seeds and may bear fruit; it is deciduous or evergreen; it is hardy in certain climate zones, and so on. On the face of it, such features are not appropriately attributed to even complex configurations of fundamental physical entities; and the same is true for the characteristic features of other special-science entities (Wilson 2021: 4).

To conclude I want to discuss an application of the new mereology for emergence that I suggested to a particular example that combines different special-science entities. The example I have in mind is that of the particular *organism* mentioned in the passage above, a tree.[20] How does the new mereology handle the constitution of an organism such as a tree, where different parts of the tree are arguably themselves weakly emergent entities studied by different special sciences?[21] It is interesting to note that the passage to new special-science level with distinctive weakly emergent wholes is clearly mirrored in the mereological system I proposed. In particular it is mirrored in the passage from a reducible whole to an irreducible one of which the former is the matter. For instance, one starts with *atoms*, studied by *physics*.[22] Sums of atoms provide the matter of other weakly emergent wholes, *molecules*, studied by chemistry. Sums of molecules provide the matter for other weakly emergent wholes, *cells*, studied by *biology*. Finally, sums of cells provide the matter of other weakly emergent wholes, *organisms*, studied in the case of a tree, by *botany*. This is illustrated in Figure 4 below:[23]

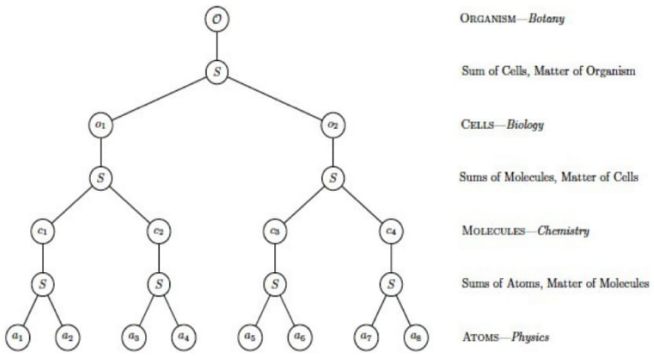


Figure 4: A Tree

[20] See also Calosi and Giordani 2023a.
[21] For a discussion of the relation between emergence, and a layered conception of reality with different levels studied by different special sciences see Wilson 2021: 12 and 24-30.
[22] For a discussion of atomism and emergence see Wilson 2021: 24.
[23] For the sake of clarity, I did not draw all the *Sum*-s.

To sum up. I argued that the possibility of emergence, as characterized in Wilson 2021, poses a threat to various reductive accounts of mereological structure. I then proposed a new account that seems to fit well with various intimations coming from the metaphysics of emergence, as applied to paradigmatic cases of emergent wholes. I admit this is just a first rung of a more thorough investigation of the mereological ladder of such emergent wholes. The hope is that this rung stands on solid ground.[24]

References

Baxter, D. and Cotnoir, A. (eds.) 2013, *Composition as Identity*, Oxford: Oxford University Press.

Bennett, K. 2013, "Having a Part Twice Over", *Australasian Journal of Philosophy*, 91, 1, 83-103.

Calosi, C. 2016, "Composition, Identity, Emergence", *Logic and Logical Philosophy*, 25, 3, 429-43.

Calosi, C. and Giordani, A. 2023a, "Atoms, Combs, Syllables and Organisms", *Philosophical Studies*, 180, 1995-2024.

Calosi, C. and Giordani, A. 2023b, "Universalism and Extensionalism Revisited", https://link.springer.com/article/10.1007/s11229-023-04091-9.

Canavotto, I. and Giordani, A. 2020, "An Extensional Mereology for Structured Entities", *Erkenntnis*, DOI: 10.1007/s10670-020-00305-5.

Cotnoir, A. 2013, "Strange Parts: The Metaphysics of Non-Classical Mereology", *Philosophy Compass*, 8, 9, 834-45.

Cotnoir, A. and Varzi, A. 2021, *Mereology*, Oxford: Oxford University Press.

Fine, K. 2010, "Towards a Theory of Part", *Journal of Philosophy*, 107, 11, 559-89.

Koslicki, K. 2008, *The Structure of Objects*, Oxford: Oxford University Press.

McDaniel, K. 2008, "Against Composition as Identity", *Analysis*, 68, 2, 128-33.

Sattig, T. 2015, *The Double Lives of Objects*, Oxford: Oxford University Press.

Sattig, T. 2019, "Part, Slot, Ground: Foundations for Neo-Aristotelian Mereology", *Synthese*, 198, 2735-49.

Schaffer, J. 2010, "Monism: The Priority of the Whole", *The Philosophical Review*, 119, 1, 31-76.

Sider, T. 2013, "Consequences of Collapse", in Baxter and Cotnoir 2013, 211-21.

Simons, P. 1987, *Parts*, Oxford: Oxford University Press.

Varzi, A. 2009, "Universalism Entails Extensionalism", *Analysis*, 69, 4, 599-604.

Wilson, J. 2021. *Metaphysical Emergence*, Oxford: Oxford University Press.

[24] For comments and suggestions on previous drafts of the paper I want to thank Alessandro Giordani. I also want to thank the editors of *Argumenta*. Needless to say, I owe more than gratitude to Jessica Wilson.

Metaphysical Emergence within Physics: Wilson's Degrees of Freedom Account

Nina Emery

Mt. Holyoke College

Abstract

Metaphysical emergence has often been used to help understand the relationship between the entities of physics and the entities of the special sciences. What are the prospects of using metaphysical emergence within physics, to help understand the relationship between three-dimensional physical entities, and the non-three-dimensional entities that have been recently posited in certain interpretations of quantum mechanics and quantum gravity? This paper explores Jessica Wilson's (2021) analysis of certain cases of metaphysical emergence in terms of degrees of freedom and raises several questions that need to be answered in order to better understand whether this analysis can be used to handle cases of metaphysical emergence within physics.

Keywords: Metaphysical Emergence, Quantum Mechanics, Quantum Gravity.

1. Introduction

In broad strokes, *metaphysically emergent* entities are characterized by being both in some sense *dependent* on some base entities, while also being in some sense *autonomous* from those base entities. Moreover, both the relevant notions of dependence and autonomy are supposed to be suitably *metaphysical*. It isn't enough for the emergent entities to either depend on or be autonomous from the base entities in some merely epistemic or pragmatic sense. Instead, the relevant kind of dependence and autonomy must be understood independently of the kinds of creatures we are, the kind of things we care about, and how we go about investigating the world.

Consider various kinds of special sciences entities—entities that play a role in our best geology and chemistry and biology and so on. On the one hand, the behavior of these entities seems to depend on our best physics; whether you're talking about tectonic plates or chemical solutions or alleles, they are ultimately composed of atoms and subatomic particles (and whatever else physicists turn up in their hunt for a final theory). At the same time, the behavior of entities like tectonic plates and chemical solutions and alleles seems in an important sense

Argumenta 10, 1 (2024): 257—266
ISSN 2465-2334

DOI 10.14275/2465-2334/202418.eme
First published: 30 November 2024

autonomous from the base entities that physics describes. At the very least, we can reliably predict the behavior of these special science entities without paying much attention at all to the details of our best physical theories—indeed that is what geologists and chemists and biologists spend quite a lot of their time doing. Is this type of autonomy suitably metaphysical? It's hard to say, but if it is, then these special science entities would be paradigm examples of metaphysically emergent entities.

So far, so good, but as the reader can surely tell, there's an enormous amount of philosophical work yet to be done both in spelling out precisely what is meant by dependence and autonomy as conditions of metaphysical emergence, and in clarifying when and where in our philosophical theories examples of metaphysical emergence arise. This is the work taken up in Jessica Wilson's important and timely new book, *Metaphysical Emergence* (Wilson 2021). In addition to putting forward a detailed account of metaphysical emergence, Wilson explores the wide range of philosophical arenas in which one might deploy this concept. There are, of course, the standard examples of special science entities mentioned above, as well as the familiar role that emergence has played in the literature on mental causation and causal overdetermination, but *Metaphysical Emergence* also shows how one might use this concept to help think through philosophical questions about the metaphysics of complex systems, ordinary objects, consciousness, and free will.

In this discussion, I'm going to focus on one particular area of application as a way of illustrating both the importance of Wilson's analysis of metaphysical emergence and raising a number of questions about that analysis. In particular, I will be focused on the ways in which the concept of emergence can be deployed *within* physics (as opposed to being deployed as a way of connecting special science entities with the entities of physics, as in the examples above). Wilson discusses this in her chapter on ordinary objects (Chapter 5). But the topic, as I see it, is much more expansive than she has space to take up there.

In recent years, philosophers of physics have gotten quite comfortable with appeals to emergence. Physicists are exceptionally good at generating mathematical formalisms that allow us to make accurate predictions, but the work of interpreting these formalisms—that is, the work of determining what these formalisms tell us about what the world is like—has become increasingly fraught. Often it is the case that the most straightforward or intuitive interpretation of the formalism tells us that the world is dramatically different than we expect it to be—even with respect to the kinds of entities that have traditionally been within the purview of physics. One example of this trend is found in foundations of quantum theory, where some philosophers of physics have begun to advocate for the view that the quantum formalism describes the evolution of a field in an extremely high-dimensional space—a space of 3×10^{80} dimensions.[1] The obvious question that this view raises is how we are supposed to think about the three-dimensional objects that have been the subject of all prior physics—are atoms and the like just an illusion? One way of resolving this question—or at least gesturing in the direction of a possible resolution—is to bring in the concept of metaphysical emergence, and claim that three-dimensional space and the three-dimensional entities occupying that space are metaphysically emergent entities.

[1] See Albert 1996 for an early version of this view and Ney 2021 for a recent comprehensive defense.

A similar line of thought has been highly influential in recent work on approaches to quantum gravity in which there is no spatiotemporal structure.[2] Obviously the world around us appears to have spacetime structure, so doesn't that make these approaches to quantum gravity non-starters? No, the standard line goes, not as long as one is willing to understand spacetime structure as in some sense metaphysically emergent.

These examples show that the concept of metaphysical emergence has the potential to play an important role in philosophy of physics. At the same time, the rules of the game in such debates are very unclear. There is little consensus on the definition or proper analysis of metaphysical emergence among philosophers of physics, or on the more general benefits and challenges of accepting this concept as a part of our overall metaphysical toolbox. Wilson's book therefore should be thought of as providing an important resource to help philosophers of physics think through these issues in a rigorous way that connects with the broader philosophical literature.

2. Wilson's DOF-based Account

As with any account of metaphysical emergence, Wilson's account has two parts: an analysis of the sense in which metaphysically emergent entities are dependent on some base entities, and an analysis of the sense in which metaphysically emergent entities are autonomous from those base entities. The latter is relatively simple (although see more on this in section 5). According to Wilson, the dependence aspect of metaphysical emergence is understood in terms of *cotemporal material dependence*. In paradigm cases (e.g. the special science cases) this involves the base entities *composing* the emergent entity.

The autonomy aspect of metaphysical emergence, on Wilson's view, is understood in two further, distinct ways. In some cases, autonomy is understood in terms of emergent entities having novel *powers* with respect to the base entities. In other cases, it is understood in terms of emergent entities having a proper subset of the powers had by the base entities. Thus we get two types of emergence:

> *Strong Emergence*. What it is for token feature S to be Strongly metaphysically emergent from token feature *P* on a given occasion is for it to be the case, on that occasion, (i) that S cotemporally materially depends on P, and (ii) that S has at least one token power not identical with any token power of P (Wilson 2021: 53).
>
> *Weak Emergence*. What it is for then feature S to be Weakly metaphysically emergent from token feature P on a given occasion is for it to be the case on that occasion, (i) that S cotemporally materially depends on P, and (ii) that S has a non-empty proper subset of the token powers had by P (ibid.: 72).

This classification is all well and good, but I fear that it doesn't help clarify when emergence occurs and when it does not unless we have a settled understanding of powers—when an entity has a power, when it does not, and what precisely powers are. And this, I strongly suspect, is a debate that many philosophers of physics will wish to avoid. With that in mind, it's also important to note that Wilson discusses various "implementations" of weak and strong emergence as defined above, and that one of these—the implementation of weak emergence in

[2] See, for instance, Wüthrich et al. 2021.

terms of degrees of freedom (DOF)—draws on a concept (degrees of freedom) that is already familiar in both physics and philosophy of physics.

Here's how the DOF-based implementation of weak emergence works. As always, the emergent entities need to cotemporally materially depend on the base entites. And then the autonomy condition is understood in the following way:

> […] at least one state of a Weakly emergent entity can be specified using strictly fewer degrees of freedom (independent parameters needed to specify states relevant to an entity's law-governed properties and behaviors) than are needed to specify the corresponding state of the system of entities upon which it cotemporally materially depends (ibid.: 18).[3]

The central example of DOF-based weak emergence, for Wilson is the relationship between the ordinary macrophysical objects that make up the world as we experience it, and the entities described by the quantum formalism. As Wilson writes, "Certain quantum DOF are…eliminated in the classical (macroscopic) limit. For example, entities of the sort treated by classical mechanics are ultimately composed of quantum entities, but the characteristics states of classical-mechanical entities do not functionally depend on the spins of their quantum components" (ibid.: 179).[4]

At least at first, this DOF-based implementation of weak emergence seems highly promising as a tool for understanding emergence within physics. But there are a number of questions that it inspires. In what follows, I'll discuss three of these questions, before returning to briefly discuss Wilson's notion of dependence.

3. The Limits of DOF-based Emergence

Perhaps the most obvious type of question that the introduction of the DOF-based implementation inspires, are questions about the limits of this way of understanding of emergence. First and foremost, we might wonder about the relationship between the DOF-based implementation and Weak Emergence as originally stated. Wilson's presentation of the concept suggests that DOF-based weak emergence only applies in particular cases, where as Weak Emergence is a more general concept. But why, exactly? What are the limits of DOF-based weak emergence? If we wanted to *exclusively* understand weak emergence in terms of the elimination of degrees of freedom, could we? If not, why not?

One way to try to figure out the answers to these questions is by looking at cases where Wilson posits weak emergence without any explicit discussion of degrees of freedom. One especially illuminating example is her application of weak emergence to free will. She writes,

> The prospects [for there actually being free will of the weakly emergent variety] are good. Though free choices are not taken to be part of a higher-level system of laws

[3] Note that Wilson says that the above description is rough. She gives a more thorough, technical definition in chapter 5.2.4. As far as I can tell, however, the details of the technical definition do not affect the discussion here.

[4] Note that although the discussion of ordinary objects being weakly emergent with respect to fundamental particles is the focus of just one subsection of the book (6.1.1), this example is repeatedly mentioned when DOF-based weak emergence is discussed. See, e.g., sections 3.2.3 and 5.2.4.

> on either compatibility or libertarian accounts, a compatibility account is one manifesting the usual Weak emergentist characterization of special science goings on as comparatively insensitive to lower-level physical details, in the sense that an agent's reasons for action in a given case float free of many such details (and in particular, are sensitive only to facts about 'relevant' causal antecedents) (ibid.: 274).

There's no explicit discussion of degrees of freedom here. Why not? One guess is that the mention of laws in the quote above is important. Perhaps on Wilson's view the DOF-based implementation is only possible when the emergent behavior is law-governed. Further support for this guess can be found in Wilson's definition of degrees of freedom. See the quote in section 1 from page 18 and also the following:

> Call states upon which the law-governed properties and behavior of an entity E (object, system, or other particular) functionally depends on the 'characteristic states' of E. A DOF is then, roughly, a parameter in a minimal set needed to describe an entity as being in a characteristic state (ibid.: 177).

From these quotes it looks as though it follows from Wilson's definition of degrees of freedom that if a certain kind of behavior isn't law governed then it won't have any associated degrees of freedom.

This restriction explains the thought that DOF-based weak emergence will only encompass a subset of the cases of weak emergence, but it is a somewhat surprising restriction to make. A fairly standard definition of degrees of freedom is that they are simply the number of independent parameters needed in order to specify a system's state. Of course we tend to only be interested in certain states of certain systems—and therefore we tend to only be interested in certain degrees of freedom. One such group is the states of systems that factor into the laws governing those systems behavior. But there are other salient groups—for instance the states of systems that factor into the explanation of those systems behavior, even if those explanations don't involve laws. And if we have this more expansive understanding of degrees of freedom—where degrees of freedom can be described for any behavior that has an explanation, even if it isn't law-governed—then we should be able to understand compatibilist-style free will as explicitly involving the elimination of degrees of freedom.

All of this by way of discussing how DOF-based weak emergence is related to weak emergence more generally. Another important question about the limits of the DOF-based implementation is whether it can be extended to help us understand strong emergence as well. In the book, Wilson presents this implementation exclusively as a variety of weak emergence. But it seems as though there ought to be a straightforward DOF-based implementation of Strong Emergence, along the following lines:

> *DOF-based Strong Emergence.* There is (i) cotemporal material dependence of the emergent entity on the base entity and (ii) least one state of the emergent entity must be specified using strictly more degrees of freedom than are needed to specify the corresponding state of the system of entities upon which it cotemporally materially depends.

Moreover, at least at first glance, there are some relatively straightforward examples of DOF-based strong emergence in philosophy of physics. For instance, on at least some interpretations of the quantum formalism, when two (or more)

particles become entangled one needs strictly speaking more degrees of freedom in order to specify the behavior of the system than one needs when specifying the behavior of the individual components of the system. For instance, if there are two particles whose spin states are entangled, it may be that all we can say about the behavior of the particles individually is that particle 1 has a .5 chance of having spin up in the z direction and a .5 chance of having spin down, and particle 2 has a .5 chance of having spin up in the z direction and a .5 chance of having spin down. But when it comes to the behavior of the system as a whole, there is an additional important pattern that comes to light, which is that when particle 1 has spin up, particle 2 has spin down. We capture this fact by saying that the wavefunction of the system as a whole takes a certain form, from which it can be derived (using Born's rule) that the probability of the particles having the same spin is 0. A natural way of thinking about this situation is that the entanglement of the particles' spin states results in there being emergent entity—the quantum system—whose state must be specified using strictly more degrees of freedom than are needed to specify the states of the individual particles.

4. Ordinary Objects as an Example of DOF-based Weak Emergence

Another way to try to better understand DOF-based weak emergence is to train a closer eye on some of the examples that Wilson provides. The central example, as mentioned above is ordinary, microphysical objects, which Wilson argues are weakly emergent (in the DOF-sense) from quantum parameters. Here's a bit more of what Wilson says about ordinary objects being weakly emergent.

> What I will call 'classical' objects are ordinary objects of the sort whose static and dynamic behaviors are appropriately treated by classical or Newtonian mechanics, understood as comprising, roughly, Newton's three laws of motion and the gravitational and electromagnetic force laws (ibid.: 192).
>
> The characteristic states of classical objects do not functionally depend on the spins of the quantum components of these entities. Hence notwithstanding that the values of quantum parameters may in some cases lead to macroscopic differences—for example, readings on a measurement apparatus, and the like, as in the case of Schrodinger's cat—it remains the case that DOF such as quantum spin are eliminated…from those needed to characterize entities of the sort appropriately treated by classical mechanics (ibid.: 194).

It is supposed to follow from all this that ordinary objects satisfy the DOF-based account of weak emergence.

The first thing to note about this example is that the details may be dependent on the interpretation that we give of the quantum formalism in fairly complicated ways. Just as one example, in Bohmian mechanics, you can talk about the spin properties of a particle, and use such talk to make predictions, but when you look more carefully, all of the behavior of a quantum particle is explained by its initial position, its initial wavefunction (in the position basis), and the two dynamical laws (the guidance equation and Schrödinger's equation). So it's not entirely clear how to think about the elimination of spin states as a degree of freedom on that interpretation. Was it ever really a degree of freedom to begin with? At the very least there seems to be room for some interesting additional work to be done in

sorting through how this example incorporates the details of various dynamical and ontological interpretations of the quantum formalism.

It's also interesting to note that it isn't immediately obvious why we need to discuss quantum parameters here at all. Consider the fact that ordinary objects like my coffee mug do not unexpectedly lift into the air and float around the room. This behavior is both predicted and explained by classical mechanics. One way of predicting and explaining it is by applying Newton's laws directly to the coffee mug. Another way is to use thermodynamics to predict and explain the behavior of the system involving the coffee cup, the table it is sitting on, and the air around it. Either way, note that you do not need to specify the position and momentum of each individual particle that is a part of the system.

It looks to me like this means that the coffee mug is a weakly emergent entity (on a DOF-based account). The mug cotemporally materially depends on the particles that compose it, but the state of the mug can be specified using strictly speaking fewer degrees of freedom than are needed to specify the states of the individual particles that compose the mug.

Call the argument just given the *classical argument for ordinary objects being weakly emergent* and Wilson's argument described above would be a *quantum argument for ordinary objects being weakly emergent*. At least at first glance it seems that the classical argument works just as well as the quantum argument for Wilson's purposes. And perhaps that's all to the good, since it means we don't have to sort through various interpretations of the quantum formalism in order to conclude that ordinary objects are in fact weakly emergent.

Of course, one thing that seems important about the classical argument is that our best physics says that classical particles with precise positions and momenta are not fundamental. But note first that it wasn't stated in the definition of DOF-based weak emergence that the base entities needed to be themselves fundamental. And second, as mentioned above, it is also controversial whether the quantum entities that instantiate properties like spin and which compose classical objects are themselves fundamental--those who think that the quantum formalism represents a field in a high-dimensional space, for instance, will disagree. So I don't think the non-fundamentality of classical particles is a good reason for treating the classical argument differently from the quantum argument unless you're willing to take a controversial stand with respect to quantum ontology.

4. When Is a Degree of Freedom Eliminated?

It's worth emphasizing the following complication in both the quantum and the classical arguments for the weak emergence of ordinary objects. In terms of the laws governing the base entities, it is *possible* for my coffee cup to lift up off the table and float around the room (or for it to, e.g. quantum tunnel through the table)—it's just very unlikely.

This is importantly different from the example that Wilson gives when discussing what it means for a degree of freedom to be eliminated. In Chapter 5, she writes:

> A case in point is that of a spherical conductor of the sort treated in electrostatics, which has DOF that are eliminated relative to the system of its composing entities; for while the E-field due to the free particles depends on all charged particles, the

> E-field due to a spherical conductor depends on the charges of particles on its surface. Certain quantum DOF are also eliminated in the classical (macroscopic) limit (ibid.: 179).

The case of the spherical conductor is one where degrees of freedom that are in other circumstances relevant to the behavior of the composing entities make *no difference at all* to the behavior of the electric field created by the conductor.

In the classical argument, the degrees of freedom that are in other circumstances relevant to the base entities (i.e. the exact position and momentum of each particle) are *very likely not to affect* the movement of the coffee mug. But there is some probability of them making quite a significant difference. The sense in which quantum degrees of freedom are eliminated in the coffee mug's behavior will also be merely probabilistic. (The exact details of the way in which they are probabilistic will depend on the interpretation one gives of the quantum formalism, but I will try to avoid going too far into the weeds here.)

So one of the key questions facing the DOF-based account is whether that is all that is necessary in order to say that a degree of freedom is eliminated—that it is *very likely not to* have an effect on the behavior of the emergent entity? Another way to put the same point: if a parameter is very likely not to have an effect on the behavior of some entity, is that sufficient to say that the behavior of that entity is *functionally independent* of that parameter?

In part this is just an interesting question to ask about this account. But it also gives rise to an interesting observation, namely that weak emergence might come in degrees, depending on the probability of the "eliminated" degree of freedom actually having an impact on the behavior of the emergent entity. For instance, in both the classical and the quantum case, the probability of a micro-parameter affecting the behavior of an ordinary object will typically decrease as the size of the ordinary object increases. So a larger ordinary object, like a school bus, might be thought of as weakly emergent *to a greater degree* than a smaller ordinary object, like a coffee mug, since the probability of a micro-parameter (e.g. the exact position and momenta of the individual particles) is less likely to affect the behavior of the school bus than the behavior of the coffee mug.

6. What Is Cotemporal Material Dependence?

All of the above discussion has focused on Wilson's understanding of autonomy. Let's turn now to think a bit more about her understanding of dependence. According to Wilson, the type of dependence involved in metaphysical emergence is *cotemporal material dependence*. As noted above, the central examples of emergence (e.g. the special science cases) are cases in which the base entities compose the emergent entities. One would be forgiven, then for thinking that cotemporal material dependence just is composition.

This is relatively straightforward, but it does raise some concerns, in particular about whether and to what extent Wilson's account of emergence can extend to contemporary debates in physics, where it isn't straightforward to understand the base entities as composing the emergent entities. Insofar as one thing helps compose another thing, both entities are standardly assumed to occupy the same physical space. But that assumption breaks down in the examples from philosophy of physics that I introduced at the beginning. If the based entity is a field in a high-dimensional space how can that field composed entities in

ordinary 3-dimensional space? And in interpretations of quantum gravity on which spacetime itself is the emergent entity, it similarly isn't obvious in what sense the base entities would compose the emergent entities.

Comments in the conclusion of the book show that Wilson is aware of this, and is leaving it to future work. That's fair enough, but it's worth pushing a little here, if only to try to get a sense of how this future work is likely to develop.

For instance, in some places in the book, Wilson says that cotemporal material dependence can be "understood as involving both (physical) substance monism and the minimal nomological supervenience of emergent feature types on base feature types" (ibid.: 73). One might take this as an indication that maybe physical substance monism in combination with minimal nomological supervenience is a sufficient condition for cotemporal material dependence.

This is likely to help with the extension of the account to at least some of the contemporary cases in physics. But it does raise some other questions. In particular, it seems like in some cases, composition as an indicator of cotemporal material dependence and minimal nomological supervenience as an indicator of cotemporal material dependence might be in tension. For instance, consider again the cases of quantum entanglement that I suggested in section 2 were potential cases of DOF-based strong emergence. Are these actually cases in which the emergent entity (the entangled system) in fact cotemporally materially depends on the base entities (the individual particles)? It isn't entirely clear.

On the one hand, the entangled system is plausibly composed by the individual particles. But also, the behavior of the entangled system does not nomologically supervene on the behavior of the individual particles—indeed it is the other way around. That's why the case seems like one that would give rise to DOF-based strong emergence.

In fact, if (substance monism plus) minimal nomological supervenience is a sufficient condition for cotemporal material dependence, then maybe cases of entanglement are better understood as cases where the individual particles are *weakly* emergent from the entangled system. After all, on this understanding, the individual particles cotemporally material dependent on the entangled system and you need *fewer* degrees of freedom in order to describe the behavior of those particles.

At any rate, all of this suggests that in order to understand the implications of Wilson's account—and in particular the DOF-based implementation of the account—in philosophy of physics, one will need to not only delve into the complexities of degrees of freedom as indicators of autonomy, but also into cotemporal material dependence as well.

7. Conclusion

The above discussion shows just how rich Wilson's account of metaphysical emergence is by exploring the ways in which just one implementation of her account (the degrees of freedom-based implementation) can be applied to debates within philosophy of physics. The questions raised above are, I think, quite difficult ones. But that just shows how interesting the concept of metaphysical

emergence is and the great potential for important further work on this topic within the philosophy of physics.[5]

References

Albert, D. 1996 "Elementary Quantum Metaphysics", in Cushing, J.T., Fine, A., and Goldstein, S. (eds.), *Bohmian Mechanics and Quantum Theory: An Appraisal*, Kluwer, 277-84.

Ney, A. 2021, *The World in the Wave Function: A Metaphysics for Quantum Physics*, New York: Oxford University Press.

Wilson, J. 2021, *Metaphysical Emergence*, Oxford: Oxford University Press.

Wüthrich, C., Le Bihan, B., and Huggett, N. (eds.) 2021, "Philosophy Beyond Spacetime: Implications from Quantum Gravity", Oxford: Oxford University Press.

[5] Thanks to Jessica Wilson for helpful discussion of the comments at the Eastern APA Author Meets Critics session on *Metaphysical Emergence*, and to Karen Bennett and Brian McLaughlin, who also participated in the session.

The Emerging Limits of Emergentism: Systematicity

Simone Gozzano

University of L'Aquila

Abstract

Taking steps from Wilson's distinction between strong and weak emergence, in this paper I cast doubts on the prospect of weak emergence. After discussing the relationship between properties set at different levels and supporting different counterfactuals and laws, I discuss one crucial condition for a property to be weakly emergent, one that is usually taken as the primary motivation for emergence, that of being "realization indifferent". I set an argument aimed at showing that this realization indifference does not accord with systematic relations holding between properties set at the mental level *vis-a-vis* their realizers. Since it is not possible to have mental properties which are not systematic, mental properties cannot be weakly emergent properties.

Keywords: Emergence, Systematicity, Multiple realization, Realization indifference, Subset.

1. The Making of Emergence

The issue of emergence still is the issue of whether special sciences are autonomous with respect to non-special, or fundamental, sciences. Such an issue was set by the debate, spanned over the years, between Jerry Fodor (1974, 1997) and Jaegwon Kim (1992, 1998 and 1999). The issue of emergence has both an epistemological side—the knowledge and methodology that we use to understand some properties in the world is absolutely specific to those properties?—and an ontological side—are there independent chunks of reality? How do they connect with other chunks?

Thus construed, emergence is seen as an articulated and robust phenomenon. *Articulated* inasmuch there are relations among properties (often called higher-level properties) which are taken to be *independent*, so distinct, of other properties (often called lower-level properties); *robust* inasmuch those relations support counterfactuals, thus allowing for predictions and explanations, that is, for a complex interrelation of epistemological procedures, tenets, and constraints. Or at least those who defend emergence seem to think.

Argumenta 10, 1 (2024): 267—277
ISSN 2465-2334

DOI 10.14275/2465-2334/202418.goz
First published: 30 November 2024

In her book *Metaphysical Emergence*, Jessica Wilson (2021, but see also 2015) argues that we have metaphysical emergence when macro-entities like humans, trees, rocks, and artifacts—as chairs and skyscrapers—are cotemporally materially dependent on but ontologically and causally autonomous from micro-entities, such as quarks and leptons, that ultimately form their base. On this general picture, two varieties of emergence are discussed: weak emergence, which occurs when a high-level feature (be it a property, state or behavior)[1] is both ontologically and causally autonomous and cotemporally materially dependent on a lower-level property or feature—where autonomy is guaranteed by having a subset of the powers had by its base features; and strong emergence, in which along with cotemporally material dependence there is a degree of autonomy to be found in the presence of a new causal power, not to be found in the base features. As such, strong emergence abandons the principle of the causal closure of the physical world, so a high-level feature occurrence cannot be traced back to the occurrence of lower-level physical features. The strong version of emergence proves to be very difficult to defend, while the weak version seems reasonable. But is this the case?

Emergence can be tackled via conceptual analysis and via metaphysics. On the conceptual side, Nicholas Humphreys (2016) has argued along two paths: one is positing that the presence of some properties cannot be derived from the presence of other properties. The other path says that taking certain configurations or patterns as evidence of emergence depends on our conceptualization of those configurations. The first path is conceptual because the notion of *derivation* is not the direct result of the adoption of the nomological-deductive method of science. So, it is a specially tailored notion. The second path, one that applies to phenomena such as flocks of birds or traffic jams, depends on our, presumably *Gestaltic*, capacity of recognizing groups of individual entities moving in a coordinated way as singular entities.

On the metaphysical side, Kim has argued that the nomological relations connecting higher-level *properties*, such as the movement of a flock, could be substituted by lower-level properties, the movements of each bird, thus favoring local reductions. Such local reductions have the burden to show that nothing is lost when the higher-level properties are split into lower-level properties, thus dissolving or reducing the seeming higher-level properties.

How did the attack on special sciences properties develop? One of the attacking points is to consider the predicates used by the special sciences to establish their own domains. For, any new science is characterized by a specific vocabulary, with its predicates and relations. Now, the predicates admissible in laws must be projectible and such that the laws mentioning them support counterfactuals. Being projectible means that the future applications of a predicate are warranted and supported by its past successes. Basically, it is a measure of inductive success, a measure of the force or strength of predicates.[2] Being counterfactual supporting means that the predicates that make a counterfactual true are those that can be included in science because they guarantee the truth of the covering

[1] Somehow betraying Wilson's wording, I will use "features" and "properties" interchangeably.

[2] As a side note in the philosophy of science, one may take it as a sign of resistance to change in science. E.g. "climate change" has not a deep entrenching in past scientific discussions, hence its projectability is modest. Consequently, it is very difficult to take it as a serious player in discussions on the future of climate.

law. Now, the wideness in the support of counterfactuals by a law is a measure of the scope of the application of the law itself. Such wideness can be evaluated both by the number and by the differences of these counterfactuals.

The number of counterfactuals is evidence of how much the law is applied, say, in the same field, thus providing more and more robustness to the projectability of its predicates. The difference in counterfactuals is to be considered in terms of type rather than of token. That is to say, a type different counterfactual establishes specific new relations and it is applied to type different entities and conditions. Clearly, there are cases where there can be type different counterfactuals, and a very high number of them, without this fact providing much insight, as when we say, e.g., that water freezes at 0°C or below and then we may formulate a counterfactual for each fraction of degree below 0°C, which is not very informative. But there are cases in which this number is of interest, as when we consider the angle at which an object bounces in a billiard table or a re-entry trajectory in the atmosphere is to be calculated. Also in this case, we may provide a counterfactual for each value, but the result could prove to be of great importance.[3]

Of greater importance is the number of *type* different counterfactuals supported by a law. Such a number depends on the adaptability of the predicates to new conditions, so by the inductive strength they have. Such strength is made evident exactly by the type-difference of the counterfactuals that the law supports, that is, as said, by the scope of the applicability of the law.

So, the number and types of counterfactuals that a law supports are determined by the strength of the projectability of its predicates, and how much a predicate is projectible depends on the inductive support given by it to successful applications of the law, success measured by the number of conditions in which the law holds. This may sound circular, but since the data and conditions are continuously changing, the circle is not vicious but rather virtuous. In a way, projectability and counterfactual support show us that conceptual analysis and metaphysics are the two sides of the same coin.

It seems, then, that what matters for the inclusion of a predicate into a law is what I would call predicate's *robustness*, namely its projectability and the counterfactual support of the law in which such predicate is included.

One of the most striking examples of this complexity is the way predicates used in psychology are now used in neurology and Artificial Intelligence. Let me contrast three different uses of "is perceiving". 1) A person is visually perceiving satisfactorily if she orientates and navigates herself properly into the world, namely if she finds her way, and does not bump into obstacles. 2) A person is visually perceiving if her eyes, lateral geniculate nuclei, and occipital areas V1-V5 are working and responding to the impinging stimuli determining the appropriate responses from the motor cortex. 3) A robot is perceiving if its cameras, processors, and CPU are such to activate its motor control engine to minimize the number of damaging interactions with the physical world while navigating it appropriately. So, the predicate "x is perceiving the environment" is used in several and type-different ways.[4]

This variety of applications, and this robustness, may come at a cost. On the one side, the wider the application, the wider the projectability and the support

[3] Thanks to Larry Shapiro for having pointed out this problem to me.

[4] I am not getting into the consciousness domain on purpose now, because I do not want to mix the issues.

for a variety of counterfactuals. On the other side, the counterexamples to the inductive base of such large-spectrum predicates can be quite different and revealing of their distinctness. This point was noted by Kim and discussed by Fodor, and the discussion was in terms of potentially disjunctive sets of confirmation.

In their original example, Fodor and Kim were considering "jade": a noun used to refer to two chemically different gemstones, jadeite and nephrite. Now, the sentence "jade is hard" is true both of jadeite and nephrite but this could be the case for different physical-chemical structures.

Fodor stressed that a high-level property could have an open or a closed set of realizers, where it being open is a crucial feature of special sciences. Now, I take the idea of an open set as quite idealized: a set should be closed for it to be defined, so let's say that what Fodor had in mind was an ideally very heterogeneous set. Let's consider pain: supposedly, in humans, it is realized by C-fiber firing, but it could be differently realized in other sentient beings and the realizers form an open set. So, we may take the property of being in pain as one that at a very high level can be shared by different entities, from human beings to other mammals, to other animals up to potentially extra-terrestrial individuals. At a finer level of detail, being in pain is multiply realized by structures that may have nothing in common.

So, is the latency, the wideness in the applicability of predicates and laws, tightly linked and supported by the projectability and number of counterfactuals or should we accept a loose relationship between the underlying (lower-level) structures supporting the higher-level phenomena?

2. Setting a Discussion

The above question bears directly to the issue of emergence, for emergence necessarily entails some form of autonomy between properties (and predicates) as referring to different levels of reality (whatever these levels are). In what follows I will consider weak emergence only, as the strong version seems to have little to no-prospects to be right. Indeed, strong emergentism entails abandoning the principle of causal closure which physicalists take to be non-negotiable. Vice versa, weak emergentism accepts the principle and tries to show that high-level and low-level features do not determine the pernicious overdetermination of so-called double-throw rock variety. Wilson's take on weak emergence is crucially set on the proper subset of power condition (PSPC) according to which a weak emergent feature S has on a given occasion powers that are a proper subset of the powers had by the Ps features on which, in that occasion, S cotemporally materially depend (CMD) (cf. Wilson 2021: 59).[5] In the terms of pain, we may say that John being in pain has both a special science feature (the phenomenal experience John is having) and a physical feature (his C-fibers firing) so that the S CMD on the Ps while being ontologically and causally autonomous from Ps. This PSPC is the way in which this autonomy is spelled out, and such condition is, in a way or another, endorsed or satisfied by all the weak emergentist parties, Wilson argues. This satisfaction, though, comes in different varieties. All these varieties are form of realization. These could be functional, constitutive-mechanistic, mereological, determinable-based or ontologically explanatory realization. Now, some of these varieties of realizations entail multiple realizability: surely functional realization does, but so mereological and determinable-based as well. To wit: one can multiply realize a wall out of the same bricks by

[5] From now on, references to Wilson's book will be just numbers in brackets.

having these parts rearranged (unless endorsing the very much debated constitution as identity thesis) or one may realize red by having either, say, crimson or scarlet and this goes hand in hand with the determinable type having fewer powers than its determinate types, thus satisfying the PSPC (65). Even if I prefer to leave it open whether all forms of realizations entail multiple realizability, we may stress that in most of the central cases of emergence, the way in which the weakly emerging property occurs is indifferent with respect to how it is realized, thus entailing some form of multiple realizability. I will say more on this later on, while defending the second premise of an argument that, I believe, could represent a problem for weak emergence. The argument goes as follows:

(i) Mental features are systematic;
(ii) (Many cases of) Emergence entails realization indifference;
(iii) Systematicity entails that realization indifference cannot hold;
(iv) Therefore, (in many cases) mental features can't be emergent.

3. Defending Premise (i)

We need to defend these premises. One crucial issue is whether mental properties Ss are systematic, as I will argue. That mental properties are systematic can be established via a sort of slippery slope: if some properties are in systematic relations, then you have a lot of systematicity.

Why accept systematicity? For the mental such acceptance is crucial: the more systematic the mental relations are, the less viable a complete reduction of them is. This was Davidson's point (1970) in stressing the anomaly of the mental (and hence its normative nature), or Fodor's (1975) point in stressing the holistic (*Quinean*: each belief is somehow confirmed by every other belief) and deeply inferential (*isotropic*: every belief is somehow pertinent to every other belief) nature of central systems.

The idea of such systematicity is that one can go from one mental state to another via logical or deductive relations. Now, this is surely true of intentional states: assuming rationality (Dennett 1971) or the principle of charity (Davidson 1974) amounts to assuming that crediting one subject with the belief that *p* entails also crediting the subject with those beliefs that follow from *p* at least directly and straightforwardly. Clearly, one has to refrain from assuming logical omniscience, but this can be limited, as I said, by taking only direct inferential links as acceptable. But is that true of qualitative or phenomenal states as well?

I think there are systematic relations also in the case of phenomenal states. Compare two phenomenal state tokens or properties Ss, say the property of feeling pain. We can consider many systematic relations. Let me make two cases for phenomenal states and one in which phenomenal and intentional states are mixed.

From stimulus to phenomenal state: if a subject is stimulated by stimulus R and enters into a phenomenal state S, it could be proved that if the subject receives stimulus 2R (double intensity) it will get into a state nS related to state S by some ratio (as per Weber-Fechner law). So, if these Ss are happening to the same subject along a short interval, we should imagine them being in a mathematical relation that somehow mirrors the values of the stimuli. This relation was supposed to be logarithmic, even if Johnson et al. (2002) have now demonstrated that the basic law of psychophysics vindicates linearity between a subjective experience (or magnitude, as they call it) and the neural activity on which it is based. According to them:

> [the] subjective magnitude, m, depends on a single, unidimensional measure, c, of the complex, multivariate neural response studied in the neurophysiological experiments: m = m(c). [Where] c = c(N), in which c(N) is the function (the operation) that yields the neural coding measure, c. If, for example, c is the mean firing rate of a population of neurons, then c(N) is the operation, summation, required to obtain c (Johnson et al. 2002: 113).

So, a set of phenomenal states, triggered by the same kind of stimuli, present internal relations that can be discovered empirically.

Let me now consider systematic relations among phenomenal states: if the subject gets a phenomenal state S such as to determine some sort of reaction (withdrawal, anxiety) it is natural to imagine that 2S will determine a modification in the speed or intensity of the reactions, even if the amount for such modifications can be hard to determine and may take a lot of empirical work, as happened in the case of the Weber-Fechner law. Again, we can imagine, and we can introspect ourselves to reveal the presence of internal relations between our phenomenal states. If both these cases were to hold, this would be in support of systematicity not only in the case of intentional features but also in the case of phenomenal features.

Finally, I take that there are systematic relations also if we consider a mix of intentional and phenomenal states in a practical argument. One may teach: if the fish stinks like that [experience this smell], throw it away. Then imagine the subject experiences the phenomenal odor of a rotten fish which prompts him to throw it away. However, if the odor is faint, the subject may take time to decide whether to throw the fish away, and this reaction time is systematically linked to the strength of the odor. So, there are systematic relations among phenomenal and intentional states as well as shown by the above *modus ponens*.

If there is systematicity at a high level, the mental, is there systematicity at a low level, the physical? This issue has to be faced by confronting the cotemporally material dependence (CMD) on which Wilson insists. Surely, if one aims at satisfying the PSPC and "realization indifference" as well, one is saying that for each single token S there could be wildly different Ps on which S supervenes. But if we consider the causal relations in which S is involved, and we should consider these because is on these that we assert that there are high-level laws of the sort discussed by special sciences, we may require a sort of systematic counterpart of supervenience: there cannot be systematic variations at a high level without systematic variations at a low level. And this should not be surprising: laws describe systematic relations. Laws in psychophysics, for instance, do exactly this: describe in mathematical terms the stable ratio between the felt sensation and the stimulus causing it.

This ratio determines a difference in the reactions, in the successive expectations, in the latency of the recovery from the stimulations, and so on. In the case of phenomenal features, the variations are embedded in systematic empirical relations.

Now, the more one considers the systematicity of the mental, the more constraints to be placed on the realizers even in case of singular realization. Systematic relations are constraints on realizability. Hence, not all realizers are fit to support all the systematic relations that you discover at the high phenomenal level.

The overall point, then, is that systematicity is a pervasive property of the relations among mental properties such that if you have some systematicity you have a lot of systematicity, and if you have systematicity all the way through, you can't have realization indifference.

4. Defending Premise (ii)

As I have discussed above, a feature being multiply realized is a primary motivation for the weak emergence of such a feature. However, Wilson denies that multiple realizability is a necessary condition for the proper subset condition to be met. Sometimes it looks like it could be a sufficient one:

> while multiple realizability is a good indicator of when a comparatively abstract ontological and causal joint is in place, that there is such a comparatively abstract joint does not hinge on multiple realizability (68).

However (see Ch. 5 on complex systems), Wilson argues that multiple realizability, if not coupled with the satisfaction of PSPC, is not even sufficient for emergence for in many (most) cases candidates for weak emergence are singularly realized. When this single realizability is the case, reductionist have an easy play and it is difficult to make a strong case for weak emergence in these terms. So, what really make the case for weak emergence are those cases in which a feature's powers are a subset of the powers of the realizers on which it cotemporally materially depends, and this may happen to be multiply realized.

As we have seen above, though, many analyses of realization crucially insist on having the weakly emerging features as multiply realized. This is the case with, at least, functional, mereological and determinable-based realization, but there are appeals to multiple realization also in ontologically explanatory realization. I think this appeal is due to the point I was mentioning in the first section: the more a feature or property can figure in type different counterfactuals the more its causal power is well established and robust. So, even in cases in which a feature is singularly realized, more than considering its actual subset of powers, one considers its *counterfactual* subset of powers, those that guarantees distinctive efficacy with respect to the superset powers on which it CMD. It is this the way in which the causal autonomy is robustly vindicated: we can establish the causal autonomy only if a feature makes some difference in a number of significant and causally different scenarios. And the best way to put it is to say that the Ss must determine a "realization indifferent regularity" (cf. Antony and Levine 1997), where the Ps are the differential realizers, no matter whether singularly or multiply. This means that a weakly emergent feature is a "tracker" (82) of difference makers (being causally efficacious) that could determine (potentially, i.e. counterfactually) a realization indifferent regularities (66-69). Such indifference can be as permissive as one can imagine it to be, as per Fodor's (1974) famous schema for special sciences. If S causes S* while CMD on P and P* respectively, this does not amount to S being a new power, for P* may well be caused by a disjunction of low-level properties P1, P2, …, Pn in each case S is instantiated. Suffice that all these Ps have a power in common, the one that satisfy the PSPC via S. I think this is enough to maintain premise (ii). In what follows I will refer to this premise in the shorthand terms: Emergence entails realization indifference.

5. Defending Premise (iii)

The third premise asserts that systematicity entails that realization indifference cannot hold. The following argument runs in support of this premise.

(1) If property S is systematic, the properties logically or empirically related to it are mentioned by or are causally covered by the same or similar laws and regularities.
(2) The Ps on which S cotemporally materially depends (CMD), should follow the same pattern of systematicity shown by S.
(3) If property S is realization indifferent, then it CMD on Ps that are not covered by the same law.
(4) If they are not covered by the same law, the Ps have different projectability patterns and support different counterfactuals.
(5) If they have different projectability patterns and support different counterfactuals, they do not establish the same systematic relations.
(6) If they do not establish the same systematic relation, property S cannot be realization indifferent.

Consider again the case of doubling the intensity of the stimulus. This case rests on using the same predicate, referring to the same property, as being in pain, so using the same projectability, and then embedding that predicate into the same law. But if we want serious realization indifference, this is not allowed, for the pattern of the projectability and counterfactual support of predicates and properties at the high-level disregard the patterns of predicates and properties at the low level. If these patterns are so distant, how can the patterns of projectability and counterfactual support at a high level be the same? These can be the same to a very limited range. For instance, you may realize a lever with iron or with wood to be included in the same machine or in two functionally identical machines: possibly the rigidity of the lever could be the same, but they may differ concerning resistance to fire or oxidation. One may say, this is not relevant. That really depends on the context. For, one may operate with the lever in certain contexts that make their resistance to fire or oxidation relevant, and this cannot be established a priori.

Similarly, if the S is a phenomenal property, it establishes systematic relations to other phenomenal or non-phenomenal properties. Consider seeing a ripe tomato. This produces a phenomenal property of appearance of full red. As such, this property is related to appearances of scarlet or crimson by a similarity relation, which could eventually be subsumed under being a determinate of the same determinable relations to those other shades. Now consider the Ps on which the S in question CMD. If the S is to be realization indifferent and respectful of PSPC, it could well be the case that the Ps on which it supervenes do not match the systematic relations established at the phenomenal level. The subset strategy would apply to them as well. But how far? Up to the point where just the P that happens to be CMD on the S in that token case? That would prevent the subset strategy of its generalization power.

One may wonder why the emergentist should accept premise (2) of this subargument. The emergentist can stress that each "level of reality", whatever that expression designates, is characterized by its laws and hence by its own projectability and counterfactual patterns, contra steps (3) and (4), and these laws could be such to determine different systematic relations or the same relations with a different degree of strength.[6] So, what consequences would bear having different systematic relations, if any at all?

[6] For this point I am indebted to Ivan Cotumaccio and Michele Paolini Paoletti, whom I thank.

According to the subset strategy a property is individuated by the set of its causal powers had by all its instances, hence these should be preserved by all its realizers (which are a subset of the causal power had by the single instances and their realizers). So, this set comprises all the properties that share the causal powers had by the realized property. But the causal powers defining the set do have causal relations to other powers. Say, a rubber band is elastic and green. Elasticity is shared among all elastic entities no matter their color. But elasticity determines fragility in cold conditions. Should we consider this as a condition on other elastic entities? I think we should but suppose we rather think not. Then we may ask a different question: should the elasticity also involve a specific ratio between, say, thickness and length of stretchability? If so, then it could be the case that only a specific realizer fits the bill. But if this is the case, then it seems Kim was right after all: each disjunct has its own merits and the high level is just a measure of our ignorance. Here, the slippery slope on systematicity I was mentioning is reflected in a similar slippery slope on causal powers: once the set is determined, several further causal relations are connected to the causal powers belonging to the set and is very difficult to imagine this being a matter of degree, because fixing the degree of resemblance sounds quite arbitrary.

I think this is a metaphysical point. The identity conditions of a property, what a property is, are determined by its causal relations, sometimes called its causal profile: what causes the property and what the property causes. If such relations are not preserved by its realizers, we can firmly question whether the realizing properties are just a superficial simulation of the property we are considering, mimickers of its behavioral performances in the *specific occasion* at hand, rather than the proper realizer of the high-level property we are considering.

6. Previous Attacks

A different and much more articulated attack on realization indifference comes from Tom Polger and Lawrence Shapiro (2016). Consider, they say, two types of entities A and B which are taken to be of the same kind by taxonomic system S1 and of a different kind by taxonomic system S2. If the factors that lead A and B to be differently classified by S2 are among those that lead them to be commonly classified by S1 and the relevant S2-variation between A and B is distinct from the S1 intra-kind variation between A and B, then we have a real case of multiple realization. However, they continue, no real-life examples come to the rescue. This may seem like an a posteriori argument, open to empirical challenges, though. They confront this argument with possible realizers as well, stressing multiple realiza*bility* rather than multiple realizations, but one may wonder how much their point generalizes.

Also, Paul Thagard (2022) has argued that realization indifference (which he calls "substrate independence") is false. Here is his argument, resting on the assumption that any mental process is an information process:

(1) Real-world information processing depends on energy.
(2) Energy depends on material substrates.
(3) Therefore, information processing depends on material substrates.
(4) Therefore, substrate independence is false.

However, one may defend realization indifference by noting that the kind of difference that energy consumption may make is not relevant to the realization of content.

Another attack comes from Matthew Rellihan (2023). He has argued that realization indifference, which is a basic tenet of functionalism, is a much weaker identity criterion than the one defended by the subset theory of realization. This point is much more relevant than the previous one, being devoted to the strategy at issue in Wilson's book. Realization indifference allows for substituting a causal element for another, provided that it satisfies the same functional role. But these elements may have very different causal powers, and having the same causal powers is required by the subset model. So, such realization indifference is not a guarantee of the sameness of causal power. Consider again the lever of iron and that of wood: they may play the same functional role in their respective machine, but the lever in iron may have a different breaking point from that of wood. So same functional role but different causal power: realization indifference is then to be relativized.

Even if I think this is an effective argument, the reply could be that functional identity has to be all the way down: the two levers must respect the same functional definition in all the relevant aspects. Even if this were the case, it is obscure why we should place such a restrictive constraint. With phenomenal properties, this contextual problem is much deeper.

As I have argued, it is not the external condition that constrain the viability of realization indifference, but systematic relations in which the high-level properties are embedded. After all, these are the properties that determine how a subject feels or what it associates that condition with something else. It is now very difficult to see how this can be guaranteed by realization indifference. Such systematic relations by themselves constrain the realizations allowed.

7. A Different Look at the Whole Argument

An alternative way to put the argument I have been defending so far is the following, which I provide in probabilistic terms:

(1) The higher the similarity in the systematicity of the relations, the lower and less probable that the realizers are wildly realization indifferent;
(2) The lower the probability of realization indifference the higher the probability of having the same realizers;
(3) The higher the probability of having the same realizers, the higher the probability of having the same laws involved;
(4) The higher the probability of the same laws involved the less distinct or causally relevant the Ss involved;
(5) The less distinct and causally relevant the Ss involved, the less their projectability and use in appropriate counterfactuals;
(6) The less their projectability and use in counterfactuals the less the autonomy of the special sciences, *pace* Fodor.

What the argument is saying is that if there is a stable relation between an isolated (not systematic) S property and a P property (these Ss and Ps are kinds) then the S is not realization indifferent, and reduction is viable. If, on the contrary, S is embedded in a pattern of regular and rational relations, hence systematic, then the viability of realization indifference is threatened if not completely undermined. I have used "threatened" and "undermined" because the argument has a

probabilistic nature. So, I admit, it is not a knockdown argument, but one that makes the relation between empirical and logical features evident, and the empirical features, as per scientific practice, point to probability rather than certainty.

On the other hand, if to defend the distinctness and causal relevance of the mental one defends their being nonsystematic, possibly one gains the realization indifference but gets closer to local reductions of the sort advocated by Kim. Now, I agree with Wilson that emergence comes in only two varieties and that the strong one comes with a very high cost that would run against physicalism. If I am right that systematicity puts a serious constraint on the viability of weak emergence, at least the one in which multiple realizability plays a crucial role, it seems that emergence in general has very few hopes to be a viable option in metaphysics.[7]

References

Antony, L. and Levine, J. 1997, "Reduction with Autonomy", *Philosophical Perspective*, 11, 83-105.

Davidson, D. 1970, "Mental Events", in Davidson, D., *Essays on Actions and Events*, Oxford: Clarendon Press 1980.

Davidson D. 1974, "Belief and the Basis of Meaning", in Davidson D., *Inquiry into Truth and Interpretation*, Oxford: Clarendon Press.

Dennett, D. 1971, "Intentional Systems", *Journal of Philosophy*, 68, 87-106.

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

Fodor, J. 1975, *The Language of Thought*, New York: Crowell.

Fodor, J. 1997, "Special Sciences: Still Autonomous After All These Years", *Noûs*, 31, 149-63.

Humpreys, P. 2016, *Emergence: A Philosophical Account*, Oxford: Oxford University Press.

Johnson, K., Hsiao, S., and Yoshioka, T. 2002, "Neural Coding and the Basic Law of Psychophysics", *Neuroscientist*, 8, 2, 111-21.

Kim, J. 1992, "Multiple Realization and the Metaphysics of Reduction", *Philosophy and Phenomenological Research*, 52, 1, 1-26.

Kim, J. 1998, *Mind in a Physical World*, Cambridge, MA: MIT Press.

Kim, J., 1999, "Making Sense of Emergence", *Philosophical Studies*, 95, 3-36.

Polger, T. and Shapiro, L. 2016, *The Multiple Realization Book*, Oxford: Oxford University Press.

Rellihan, M. 2023, "A Familiar Dilemma for the Subset Theory of Realization", *Analytic Philosophy* 64, 68-90.

Thagard, P. 2022, "Energy Requirements Undermine Substrate Independence and Mind-Body Functionalism", *Philosophy of Science*, 89, 70-88.

Wilson, J. 2015, "Metaphysical Emergence: Weak and Strong", in Bigaj, T. and Wüthrich, C. (eds.), *Metaphysics in Contemporary Physics*, Brill: Leiden, 251-306.

Wilson, J. 2021, *Metaphysical Emergence*, Oxford: Oxford University Press.

[7] For comments on a previous draft, I express my gratitude to Ivan Cotumaccio, Michele Paolini Paoletti, Larry Shapiro, and Jessica Wilson.

Wilson on Metaphysical Emergence

Brian P. McLaughlin

Rutgers University

Abstract

I critically examine Jessica Wilson's views concerning the relationship between Weak emergence and Physicalism and between Strong emergence and Physicalism, and also her defense of libertarian free will in *Metaphysical Emergence* (2021).

Keywords: Metaphysical emergence, Weak emergence, Strong emergence, Physicalism, Fundamental interactions, Free will.

Jessica Wilson's *Metaphysical Emergence* (2021) is a wonderful book. It addresses a wide range of central metaphysical issues from an overarching theoretical perspective. Not only is it must-reading for anyone who works on metaphysical emergence, it contains a wealth of material that should be of interest to anyone who works on physicalism, realization, the metaphysics of complex systems, the metaphysics of ordinary objects, consciousness, mental causation, or free will.

As the title of her book makes evident, Wilson is concerned with metaphysical emergence—metaphysical, rather than merely epistemic emergence. More specifically, she is concerned with whether special science and (scientific and folk) mental kinds, properties, and their instances metaphysically emerge, respectively, from physical kinds, properties, and their instances. A central aim the book is to examine the relationship between that issue and physicalism (15).[1] I'll focus on that aim.

What, then, is physicalism? Wilson takes the core idea of physicalism to be that our world is fundamentally physical.[2] What counts as physical? Wilson appeals to a physics-based conception of the physical, with a caveat in response to Hempel's (1969) famous dilemma (23). The first horn of that dilemma is that if by the physical we mean what is posited by current physics, then, since current physics is incomplete and at least to some extent inaccurate, the claim that our world is fundamentally physical is false. The second horn is that if instead we mean what would be posited by an ideally completed physics that is in fact true of our world, then, since we don't know what such a physics would posit, the

[1] Numerals in parentheses are references to page numbers in the book.

[2] She takes the notion of fundamentality as a primitive (31).

Argumenta 10, 1 (2024): 279—290
ISSN 2465-2334

DOI 10.14275/2465-2334/202418.lau
First published: 30 November 2024

claim that our world is fundamentally physical is largely vacuous. Current physics, for instance, has no need of the hypothesis that there are mental phenomena, but mightn't it turn out to be the case that the physics in fact true of our world does? As Wilson conceives of the physical, it is whatever would be posited by the completed physics in fact true of our world, with the following caveat: A mental feature is not to be counted as a physical feature even if that physics would posit it. She calls this constraint on her physics-based conception of the physical "the no fundamental mentality constraint" (23). She uses it to impose a constraint on physicalism: any doctrine deserving of the name 'physicalism' should be incompatible with the physics in fact true of our world having to posit mental phenomena. She doesn't state a "no fundamental chemical" or a "no fundamental biological" constraint. When discussing physicalism, her attention is typically focused on the place of the mental in nature. I think she would, though, accept such additional constraints. It is clear, for instance, that if the physics in fact true of our world would have to posit entelechies or a fundamental vital force, she would take physicalism to be false (8).

Unlike a term like 'causation', the term 'emergence' is a term of art. Its uses are many and varied both in the philosophical and in the scientific literature.[3] Indeed, they are so diverse that one wonders whether there is even any common core idea. Focusing on metaphysical emergence narrows things down. It is fairly common ground in the philosophical literature at least that whenever there is metaphysical emergence, there is something that emerges and *something else* that it emerges from; that metaphysical emergence is incompatible with reduction; that it always involves emergent properties; and, moreover, that the bearers of emergent properties are complex entities: macro-entities constituted by micro-entities.

Wilson maintains that the core idea of metaphysical emergence is that of dependence with autonomy (1). Emergents are dependent on what they emerge from, yet autonomous from them. She is concerned with emergence from the physical. She calls the kind of dependence that she maintains is required for it, "co-temporal material dependence" (1); and she distinguishes two kinds of autonomy: ontological and causal. She states: "The coupling of co-temporal material dependence with ontological and causal autonomy [...] is most basically definitive of the notion of (metaphysical) emergence" (1). Let's consider, in turn, her notions of ontological autonomy, co-temporal material dependence, and causal autonomy.

What ontological autonomy from the physical comes to is just failure of emergents to be identical with anything physical. Following Wilson in using 'feature' as a blanket term for kinds and properties (including relational properties), if a feature S metaphysically emerges from a physical feature P, then S is not identical with P or any other physical feature. Following her in using 'token feature' as a term for a particular entity's having a feature at a time or throughout an interval of time, if a token feature S emerges from a token physical feature P, then S is not identical with P or any other physical feature token. Further, if a feature S emerges from a physical feature, then any entity that has S is not identical with any physical entity. She takes reduction to require identity claims, and so maintains that metaphysical emergence is incompatible with reduction.

[3] See, for example, the essays in Bedau and Humphreys 2008.

Wilson doesn't explicitly state a definition of 'co-temporal material dependence'. But from her discussion (Ch.1), I take it that she holds that an entity's having a feature S at a time t (what she calls "a token feature S") co-temporally materially depends at t on a configuration of fundamental physical particles having a physical feature P at t (what she calls "a token feature P") just in case at t, the configuration of fundamental particles is coincident with the entity and has a physical feature P that minimally nomologically necessitates S. (Wilson suggests how this could be modified should our world turn out to be gunky (24), but the modification needn't concern us here.) I take it that although a physical feature P must minimally nomologically suffice for S if S emerges from P, P needn't be nomologically necessary for S. Co-temporal material dependence on the physical is compatible with an emergent feature's having multiple physical emergent bases. A token of feature S might emerge from a token of feature P, while a different token of feature S emerges from a token of feature P*, where P and P* are distinct physical features.

Turn to causal autonomy. Wilson holds that emergent features have causal powers: powers to produce certain kinds of effects when an entity has them in certain circumstances. She takes token features, an entity's having a feature at a time or throughout an interval of time, to be the primary *relata* of the causal relation (40). She takes token features to have causal powers too, "token powers" (72). By that I take it she just means that they have causal effects in virtue of being tokens of the features in question and the circumstances in which they are instantiated. She distinguishes two kinds of causal autonomy, and uses the distinction to distinguish two kinds of metaphysical emergence. Her distinction between the two kinds of metaphysical emergence plays a major role throughout the book, so let's turn to it.

Wilson characterizes the two kinds of metaphysical emergence as follows:

> Weak Emergence. What it is for a token feature S to be Weakly metaphysically emergent from token feature P on a given occasion is for it to be the case, on that occasion, (i) that S co-temporally materially depends on P, and (ii) that S has a non-empty proper subset of the token powers had by P (72).
>
> Strong Emergence. What it is for token feature S to be Strongly metaphysically emergent from token feature P on a given occasion is for it to be the case, on that occasion, (i) that S co-temporally materially depends on P, and (ii) that S has at least one token power not identical with any token power of P (120).

The definitions include the same first condition, co-temporal material dependence (explained earlier), but their respective second conditions express different kinds of causal autonomy. In cases of Weak emergence, the token feature S is causally autonomous from the token feature P in that it has a different complete causal profile from the complete causal profile of the token feature P: The token powers of the token feature S (i.e., its effects) are a proper subset of the token powers (the effects) of the token feature P. Thus, every effect of the token feature S is an effect of the token feature P, but the token feature P has effects that the token feature S doesn't have. In cases of Strong emergence, a token feature S has at least one token power (one effect) that is not identical with any token power (any effect) of the token feature P; it does so in virtue of feature S's having a causal power not possessed by P.

I regard Wilson's characterizations of Weak and Strong emergence as entirely stipulative, and so to be judged solely in terms of their theoretical fruits. Of each, we should ask whether there are any instances of the kind of emergence in question, and, if so, what theoretical consequences that has. I'll be concerned with whether there are any instances of the kinds in question, and, if so, the theoretical consequences of that for physicalism, where physicalism is understood to be the thesis that our world is fundamentally physical.

Before turning to those issues, however, I want to first briefly consider other notions of emergence in the literature. Some theorists would deny that causal autonomy, in either of Wilson's two senses, is among the conditions "most basically definitive of the notion of (metaphysical) emergence" (1). They maintain that emergent features can be epiphenomena, and so devoid of causal effects.[4] Let's call that kind of emergence "epiphenomenal metaphysical emergence". One might try to characterize it along Wilson's lines in terms of co-temporal material dependence with the null set of causal powers. Wilson discusses epiphenomenalism (97–101, 140–141). She points out that in the literature, the leading candidates for epiphenomena are the phenomenal or qualitative characters of subjective experiences—their what it is like for the subject aspects—, and argues that they are in fact causally efficacious. I agree with her view that they are causally efficacious. Still, the notion of epiphenomenal metaphysical emergence is coherent; it is an *a posteriori* issue whether there is any. Let it suffice to note, then, that although Wilson sometimes seems to suggest that Weak and Strong emergence are the only two basic kinds of metaphysical emergence, I take it that her considered position is that they are the only basic kinds of metaphysical emergence that we have reason to believe may be found in our world. Of course, epiphenomenal emergentists will disagree even with that weaker claim, but I'll say no more about epiphenomenalism.

As concerns a number of other at least apparently different notions of emergence in the literature, Wilson argues either that they fail to be notions of *metaphysical* emergence or else they in fact involve either Weak or Strong emergence. I recommend in this connection reading her chapter "Complex Systems". It is informative, but it would have benefited from a discussion of the notion of emergence used in solid state physics. That notion is certainly not the notion of Strong emergence in her sense. It would have been instructive to know whether she thinks it involves Weak emergence or instead that it isn't a kind of metaphysical emergence, and why. Be that as it may, I'll now focus just on her notions of Weak and Strong emergence.

Weak and Strong emergence are not so-called because Strong implies Weak but Weak doesn't imply strong. Neither implies the other. They are incompatible: It is impossible for a token feature S to be both Strongly and Weakly emergent from a token feature P, for the simple reason that it can't be the case that the token causal powers of S are a proper subset of the token causal powers P and also the case that S has a token causal power not had by P. Given that they are incompatible, one might wonder why she labels them "Weak emergence" and "Strong

[4] See, for example, Chalmers 1996.

emergence".[5] She doesn't explicitly say, but I take it that she so labels them because she holds that Weak emergence from the physical is weaker than Strong emergence from the physical in the following way: Weak is compatible with physicalism, while Strong is not.

Wilson defends the twofold claim that (a) there is Weak emergence and there may well be Strong emergence, and that (b) while Weak emergence is compatible with physicalism, Strong emergence is incompatible with physicalism. This twofold claim will be my central focus.

Wilson tells us that physicalism is committed to Physical Causal Closure: the thesis that "every lower-level physical effect has a sufficient purely lower-level physical cause" (41). (I take it that the thesis isn't supposed to entail causal determinism. A sufficient cause of an effect must determine the objective probability of the effect, but that can be less than 1 if causal determinism is false.) Weak emergence is compatible with Physical Causal Closure, since the causal powers of the emergent will be a proper subset of the causal powers of its physical base. In contrast, Strong emergence, she tells us, is incompatible with Physical Causal Closure: If there is Strong emergence, then there are at least some lower-level physical effects that do not have any purely physical lower-level sufficient cause (41).

Wilson's formulation of Physical Causal Closure invokes a notion of level, and so presupposes a notion of levels in nature. To be sure, proponents of metaphysical emergence standardly maintain that nature is layered, with higher levels metaphysically emerging from lower levels. Wilson could of course appeal to Weak and Strong metaphysical emergence to characterize two different notions of levels in nature. But the Physical Causal Closure thesis is not supposed to entail that there is metaphysical emergence of even the Weak kind. If, then, the notion of levels invoked in Physical Causal Closure is not to be understood in terms of metaphysical emergence, how should it be understood? What is a level? It is uncontroversial that there are macro-micro levels, but they are just a matter of scale. A proper micro-constituent of a macro-entity will be at a lower level, lower scale, than the macro-entity. But any micro-configuration of physical particles that makes up an entity (at a time) will be at the same scale as that entity (at that time). Systems of particles arranged mountain-wise are at the same scale as mountains, and so not at a different level in the micro-macro sense. So what, then, is a level? Wilson discusses that question (24–30), but doesn't commit to a definitive answer to it since she seems to want to remain neutral on certain issues.

I won't pursue the question of how 'level' should be understood in the Physical Causal Closure thesis. The reason is that I think that Wilson needn't appeal to a notion of levels in order to formulate a physical causal closure thesis that is suitable for her purposes. Given her no fundamental mentality constraint, she could reformulate Physical Causal Closure just as the thesis that every physical effect has a sufficient purely physical cause (one that determines its objective probability). She could then claim that if any mental features are Strongly emergent, that thesis is false, and so physicalism is false since there are *fundamenta* that are not physical. (To address the issue of whether there is chemical or biological

[5] The terms 'weak emergence' and 'strong emergence' get used in the literature, though not in a uniform way. I'm here just concerned with her terms 'Weak emergence' and 'Strong emergence' as she defines them.

Strong emergence, issues she doesn't pursue, one could appeal to a no chemical or no biological constraint on the physics-based conception of the physical.)

Mainly for readability, rather than using 'features' and 'token features', I'll now, for the most part, frame the issues in terms of properties (monadic properties, dyadic ones, etc.), and in terms of states and events as the *relata* of the causal relation. Nothing, I believe, will turn on this shift in terminology. Unless I explicitly indicated otherwise, I'll take states and events to be an entity's having a property at a time or throughout an interval of time, and so what she calls a token feature.

Wilson maintains that Weak emergence is widespread among the special sciences yet compatible with our world being fundamentally physical. Reductive physicalism, she holds, requires that every contingent entity, event, or property be identical, respectively, with some physical entity, event, or property, but that isn't required for our world to be fundamentally physical, and so isn't required for physicalism. A kind of non-reductive physicalism could be true (55–58). She doesn't herself embrace non-reductive physicalism, however, at least not across the board. As I mentioned, she takes there to be reason to believe that there may very well be certain cases of Strong emergence, and so reason to believe that even non-reductive physicalism, as a general doctrine, may very well be false; but of that, more shortly. Let's first look more closely at the relationship between Weak emergence and non-reductive physicalism.

Wilson's notion of Weak emergence requires a modification if Weak emergence across the board is supposed to guarantee non-reductive physicalism. The nomological requirement on Weak emergence is that if a feature S Weakly emerges from a physical feature P, then P is minimally nomologically sufficient for S. That condition is compatible with the law linking S and P being a fundamental law of nature, a law that doesn't hold in virtue of other laws and conditions. The notion of Weak emergence is thus silent about whether the laws linking Weak emergents with their physical bases hold in virtue of physical laws and physical conditions. If S is, for instance, a mental property, the law will be a psychophysical law. The existence of fundamental psychophysical laws is incompatible with physicalism, reductive or non-reductive. If mental properties are distinct from physical properties, and there are fundamental laws in which they figure, then it's not true that our world is fundamentally physical, even if the instances of mental properties don't make a non-redundant causal contribution to the course of physical events (or indeed even if they are epiphenomenal). Mental properties and their instances would be, respectively, fundamental properties and property instances. Since Weak emergence is compatible with fundamental psychophysical laws, it is possible for Weak emergence to hold across the board and yet non-reductive physicalism be false. To avoid this result, the condition of co-temporal material dependence must be amended. It must be amended to include the requirement that the law linking S and P not be a fundamental law of nature; it must be a law that holds in virtue of physical laws and physical conditions.

It should be noted that while this amendment is needed if Weak emergence is to serve the purpose in question, the condition of co-temporal material dependence should not be so amended in the characterization of Strong emergence if Strong emergence is to do the work Wilson intends it to do. A Strong emergentist should hold that laws linking emergents with their physical bases are fundamental laws; and so, not ones that hold in virtue of physical laws and physical conditions. Thus, if Weak and Strong emergence are to do the work that Wilson intends, the

two kinds of emergence require different kinds of co-temporal material dependence, not just different kinds of causal autonomy.

It is fairly common for self-billed non-reductive physicalists to claim that although there are contingent objects, events, and properties that are not physical, they are *realized*, respectively, by physical objects, events, and properties. Realizers are supposed to be more ontologically fundamental than what they realize, thus allowing a kind of non-reductive physicalism. This agreement among non-reductive physicalists is thin, however. 'Realization', like 'emergence', is a term of art. We must be told what's meant by the term. Non-reductive physicalists oblige, but there are a number of non-equivalent relations that get called 'realization' in the literature. As Wilson makes clear, she takes Weak emergence to be realization (vii).[6] She readily acknowledges that there are various notions of realization in the literature, but she seems to hold that they all involve the notion of Weak emergence. She seems to view them as invoked to try to help explain how the kind of causal autonomy required for Weak emergence is implemented. Her view seems to be that if there is realization of any of the kinds in question, then there is Weak emergence.

If, as I've argued, in cases of Weak emergence, the laws linking an emergent with its physical bases must be non-fundamental, it cries out for explanation how it is that such laws hold in virtue of physical laws and physical conditions. Non-reductive physicalists typically want an account of realization that yields such explanations. The role-functionalist notion of realization as causal role occupancy, for instance, yields an explanation of why laws that invoke functional properties hold in virtue of physical laws and physical conditions, and so are not fundamental laws, even though functional properties are not identical with the physical properties that occupy the roles in question. The notion of Weak emergence itself won't yield an explanation of how laws citing Weakly emergent properties hold in virtue of physical laws and conditions.

It is important to note, moreover, that while a role functionalist may hold a view of causation according to which functional states and their physical realizers meet the causal autonomy condition for Weak emergence, a role functional needn't hold such a view. Role functionalists hold that a functional state is a second-order state of being in some state or other that has certain causal effects, and that the first-order states that have those effects realize the functional state. It is open to a role functionalist to maintain that a functional state, a state of being in some state or other that has certain effects, does not itself cause those effects. Its realizers do. That's compatible with functional states figuring in causal explanations of the effects in question.[7] But it is incompatible with Weak emergence.

Weak emergence requires that there be a certain kind of causal overdetermination. As Wilson points out, the kind in question will be different from the familiar kind of causal overdetermination that occurs when, for instance, the shattering of a window is overdetermined by two rock throws (40–46). If one of the rocks throws had not occurred, the window would still have shattered, but not in precisely the manner and at precisely the time in which it in fact shattered. Weakly emergent events, if there are such, don't overdetermine the effects of their

[6] See also Shoemaker's (2009) subset view of realization. Wilson tells us that the subset view of realization was first proposed by Michael Watkins (vii).

[7] For details, see McLaughlin 2006, 2015.

physical bases in that way. The effects of a Weakly emergent event will be precisely the same in manner and time of occurrence as those of a proper subset of the causal effects of its physical event base. Wilson regards this kind of overdetermination as unproblematic, since it is compatible with Physical Causal Closure. It is indeed compatible with Physical Causal Closure. But it cries out for explanation how such overdetermination could occur in our world. We need an explanation of how emergent events can have certain causal effects that their physical base events have, even though those effects would have occurred in precisely the same manner and time even if the emergent event had not occurred.

Whether there is overdetermination of the kind Weak emergence requires, and so whether there is Weak emergence, depends on the answers to questions about the *relata* of the causal relation and about the nature of causation. As Wilson points out (40–44), Jaegwon Kim (1998, 2005) wonders what causal work an emergent state or event could possibly be doing were there such overdetermination, given the causal work done by its physical base. A leading non-reductive physicalist response to Kim's no-work objection is that he is assuming a productive notion of causation, and causation is, rather, a kind of counterfactual dependency (Loewer 2007). Whether this response is available to Wilson depends on some issues about which she is silent. If the entity, feature, or time of a token feature are essential to the token feature, then token features are too fragile to serve as the *relata* of the causal relation on a counterfactual theory of causation.[8] It thus matters whether they are essential to the token feature. Wilson is silent about that.

It is, moreover, uncertain why a non-reductive physicalist would have to appeal to the kind of overdetermination required for Weak emergence. That isn't required if role functionalism counts as a kind of non-reductive physicalism, since, as I've noted, it is at least open to a role functionalist not to countenance the kind of overdetermination Weak emergence requires. It also remains open to a non-reductive physicalist to eschew Wilson's view of the *relata* of causal relations as feature tokens in favor of a coarse grained view of events, and to maintain that every event is identical with some physical event, but deny that special science and mental event types reduce to physical event types.[9] Further, it remains open to a non-reductive physicalists who embraces Wilson's view of the *relata* of causal relations as feature tokens to argue that special science and mental tokens have novel causal powers in a way that is compatible with Physical Causal Closure: They could have novel effects without having novel physical effects. It's been argued, for instance, that special science and mental events will screen off their underlying physical bases from having certain non-physical effects that those special science and mental events have.[10]

Notice that if the kind of view of causation last mentioned is viable, then Strong emergence, as Wilson defines it, isn't incompatible with Physical Causal Closure. A Strongly emergent state or event can have an effect that its physical base doesn't have, yet not have any physical effect that its physical base doesn't have. That's compatible with Physical Causal Closure. Wilson's intent, though, is clearly that Strongly emergent features have novel physical effects, physical effects that lack sufficient purely physical causes (54), so that if there are Strongly

[8] See Lewis 1986.
[9] See, for example, Davidson 1970.
[10] See, for example, Yablo 1992.

emergent features, then Physical Causal Closure is false, and hence physicalism is false. She may be taking it as given that an emergent couldn't have a novel effect (one its physical base doesn't have) without having some or other novel physical effect (one its physical base doesn't have). That may be so, but the issue has certainly not been settled. There is no such consensus about causation. I suggest that rather than getting into the weeds about whether a special science state or event could have novel effects without having novel physical effects, Wilson should modify the definition of Strong emergence so that it explicitly requires that Strongly emergent token features have at least one physical effect that their physical token feature base lacks.

To return to Weak emergence, although Wilson has much of interest to say about non-reductive physicalism and causation, she doesn't say enough to establish that any doctrine deserving of the label "non-reductive physicalism" requires appeal to the kind of overdetermination Weak emergence requires. Moreover, if a non-reductive physicalist maintains there is overdetermination of the kind in question, she owes us an explanation of how it is that there is such overdetermination. The notion of Weak emergence won't help to answer that question. As concerns Weak emergence and non-reductive physicalism, then, my main take away points are that it remains unresolved whether there is overdetermination of the sort Weak emergence requires, and so whether there is Weak emergence, and also whether any doctrine that counts as non-reductive physicalism must appeal to Weak emergence.

Let's turn, finally, to Strong emergence. Wilson claims that libertarian free will requires the Strong emergence of decisions and acts of will, and so is incompatible with Physical Causal Closure, and thus incompatible with physicalism (281). Of course, if there is in fact no such libertarian free will, physicalism faces no such threat. The book's jaw dropper is that Wilson maintains that there is "good reason to think that we have free will of libertarian, Strong emergent variety" (281). She makes a case that we have *prima facie* reason to believe that we have libertarian free will, and that that *prima facie* reason has thus far not been defeated. Her considered position seems to be that we are entitled to believe it until it has been defeated. At one point, though, she says something stronger: "I conclude that there is actual free will of both Weak and Strong varieties" (281). That, however, can't be the best way to state the conclusion she intends. Weak and Strong emergence, you'll recall, are incompatible. If decisions or acts of will are Weakly emergent, then they are not Strongly emergent; and if they're Strongly emergent, then they are not Weakly emergence.

In what remains, I'll focus just on Wilson's claim that decisions and acts of will are Strongly emergent. I'll simply assume, for the sake of argument, that a libertarian notion of free will requires that.

Wilson tells us a novel causal power of a Strongly emergent feature will be a novel fundamental power (54), a power to influence the course of physical events that no physical feature has. Indeed, Strong emergentism, she tells us, "is committed to there being at least one other fundamental force beyond those fundamental forces currently posited" (50) by physics. The force would be a configurational force, a fundamental force, yet one that can be exerted only by complex configurations of particles. As she notes (46-49), in McLaughlin 1992, I claimed that one finds this idea in some of the literature in the British Emergentist tradi-

tion, and that such configurational forces are compatible with Schrödinger's equation, and also that it is an empirical question whether there are such forces. I stand by those claims.

I also claimed in McLaughlin 1992 that I am deeply skeptical about whether there are any fundamental configurational forces, that there seems to be no evidence for their existence, and compelling empirical reason to think there are no such forces. I stand by those claims too. Such forces would involve complex configurations of physical particles participating in fundamental interactions in the physicist's sense of "fundamental interactions". As concerns fundamental interactions in that sense, Wilson says whether there are fundamental configurational interactions is an "open empirical question contingent on as yet unconducted experiments establishing that […] one or more fundamental interactions come into play only under certain comparatively complex circumstances" (283). If, however, that were such fundamental configurational interactions, then current physics would be wrong in a deep way that there is no evidence to believe it is. I'll now elaborate on this point, drawing heavily from a pair of superb articles by the physicist Sean Carroll (2021, 2022). I'll briefly sketch things in broad strokes; for technical details presented in an accessible way, see the Carroll articles.

Quantum field theory includes the Standard Model of particle physics and also gravitation in the weak-field limit of general relativity. It doesn't cover gravitation near black holes; it is silent about the very early universe, about dark matter and dark energy, and also about interactions energies below certain thresholds. Conditions required for its applicability are that gravity is weak and interactions involve energy transfers below a certain threshold. But as Carroll (2021, 2022) points out, human brains and our earthly environment fall well within its scope of applicability.

The key point for present purposes is this: In the field dynamics of quantum field theory, interactions are *local*.[11] They are local in that fields directly interact with other fields only at spacetime points. That is to say, the dynamics of each field at any spacetime point are directly influenced only by the values and derivatives of the other fields at that same point, and not by anything happening elsewhere. That fundamental interactions are local is inextricably baked into the theory. Quantum field theory could, for instance, accommodate new kinds of particles and new kinds of fundamental forces. But the discovery of fundamental configurational interactions would refute the theory. It thus isn't just that quantum field theory doesn't now posit fundamental configurational interactions, it cannot countenance them. Such direct fundamental interactions would involve whole regions of spacetime. That is incompatible with relativity theory.

Quantum field theory has been enormously successful in its regime of applicability, and, as noted, human brains fall well within that regime. The truly enormous empirical support quantum field enjoys soundly defeats any intuitions we might have about there being a fundamental force of will.

Still, to be sure, fundamental configurational interactions can't be ruled out *a priori*. Suppose, then, that current physics has gone very badly wrong indeed, since there are fundamental configurational interactions (relativity theory be damned). Suppose further that acts of will are co-temporally materially dependent on complex neural events, which are in turn co-temporally material dependent on

[11] Entanglement is not local, but it isn't an interaction in the physicist's sense.

events involving astronomically complex micro-configurations of physical particles that participate in fundamental interactions, and so locality fails. Physical particles don't obey the same basic equations when they are in a human brain that they obey when inside a block of ice, even though at some scale human brains fully decompose into physical particles.

Suppose all that is so. Why would it follow that there is libertarian free will? Why would the imagined yet undiscovered fundamental force be a force of will, rather than a fundamental configurational physical force? If acts of will are not identical with the events involving the astronomically complex configurations of particles that (by hypothesis) participate as wholes in such fundamental interactions, but only materially dependent on them, then the question remains whether the acts of will themselves participate in fundamental interactions. Any physical event from which an act of will Strongly emerges will (by definition) nomologically necessitate the act of will, as will any other physical event that nomologically necessitates the physical event in question if nomological necessitation is transitive. Mightn't the acts of will only Weakly emerge from their complex physical base events? Mightn't the acts of will even be epiphenomena, devoid of any effects, and so only be epiphenomenally emergent from those complex physical events? I take it that Wilson's answer to both questions would be "No," but I myself don't see why the answers would be "No". I find it deeply obscure how fundamental configurational interactions, even if there were such, could yield libertarian free will.

Since I've focused mainly on what I take to be some remaining issues for Wilson's view, let me once again express my admiration for *Metaphysical Emergence*. There is much of interest in the book that I haven't even touched on. The book will, I believe, contribute to setting the research agenda on a wide swath of metaphysical issues for years to come.

References

Bedau, M.A. and Humphreys, P., eds., 2008. *Emergence: contemporary readings in Philosophy and Science*. MIT Press.

Carroll, S., 2021. Consciousness and the laws of physics. *Journal of Consciousness Studies*, 28 (9), 16–31.

Carroll, S.M., 2021. The Quantum field theory on which the everyday world supervenes. *In*: O. Shenker, M. Hemmo, S. Iannids and G. Vishine, eds. *Levels of Reality: a scientific and metaphysical investigation*, Jerusalem studies in Philosophy and History of Science. Copenhagen: Springer, 27–46.

Chalmers, D.J., 1996. *The Conscious Mind: in search of a fundamental theory*. New York: Oxford University Press.

Davidson, D., 1970. Mental events. *In*: L. Foster and J.W. Swanson, eds. *Experience and Theory*. Oxford: Clarendon Press, 207–224. Reprinted in Davidson 1980.

Davidson, D., 1980. *Actions and Events*. Oxford: Clarendon Press.

Hempel, C.G., 1969. Reduction: ontological and linguistic facets. *In*: S. Morgenbesser, P. Suppes and M. White, eds. *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel*. New York: St Martin's Press, 179–199.

Kim, J., 1998. *Mind in a Physical World: an essay on the mind-body problem and mental causation*. Cambridge, Mass: MIT Press.

Kim, J., 2005. *Physicalism, or Something Near Enough*. Princeton, N.J: Princeton University Press.

Lewis, D., 1986. Events. *In*: *Philosophical Papers Vol. II*. Oxford: Oxford University Press, 241–269.

Loewer, B., 2007. Mental causation, or something near enough. *In*: B.P. McLaughlin and J. Cohen, eds. *Contemporary Debates in Philosophy of Mind*. Hoboken, NJ: Blackwell Publishing, 243–264.

McLaughlin, B.P., 1992. The rise and fall of British emergentism. *In*: A. Beckermann, H. Flohr and J. Kim, eds. *Emergence or Reduction? Prospects for a Nonreductive Physicalism*. Berlin: De Gruyter, 49–93. Reprinted in Bedau and Humphreys 2008.

McLaughlin, B.P., 2006. Is role functionalism committed to epiphenomenalism? *Journal of Consciousness Studies*, 13 (1–2), 39–66.

McLaughlin, B.P., 2015. Does mental causation require psychophysical identities? *In*: T. Horgan, M. Sabates and D. Sosa, eds. *Qualia and mental causation in a physical world: themes from the philosophy of Jaegwon Kim*. Cambridge: Cambridge University Press, 64–104.

Shoemaker, S., 2009. *Physical Realization*. Oxford: Oxford University Press.

Wilson, J., 2021. *Metaphysical Emergence*. Oxford: Oxford University Press.

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

Questioning, Rather Than Solving, the Problem of Higher-Level Causation

Erica Onnis

University of Turin

Abstract

In *Metaphysical Emergence*, Jessica Wilson recognises the problem of higher-level causation as "the most pressing challenge to taking the appearances of emergent structure as genuine" (2021: 39). Then, Wilson states that there are "two and only two strategies of response to this problem" (2021: 40) that lead to Strong and Weak emergence. In this paper, I suggest that there might be an alternative strategy—not opposite, but different in kind—to approach this difficulty. As noticed by Wilson, the problem of higher-level causation was formulated and made central by Jaegwon Kim. However, Kim's arguments were grounded on distinct metaphysical principles—including Alexander's Dictum and its analysis in terms of causal powers. Rather than following Kim's formulation and responding to the problem he raised in his own terms, a different approach may be to question the pertinence of the metaphysical framework in which these arguments were originally grounded. The problem of higher-level causation, in other words, might be less "pressing" if ontological emergence came with a less strict and univocal view of causal novelty and ontological relevance.

Keywords: Emergence, Alexander's Dictum, Causation, Causal powers.

1. The Troubles of the Nonreductionist

Jessica Wilson's *Metaphysical Emergence* (2021) is devoted, as the title suggests, to the analysis of metaphysical forms of emergence. Wilson's focus is on special science macro-entities, whose ontological and causal autonomy are issues close to her heart. She ascribes two features to these entities. First, they depend upon certain complex configurations of fundamental entities, being cotemporally materially composed by them. Second, despite this dependence, special science entities exhibit some ontological and causal autonomy, being "[…] distinct from, and distinctively efficacious with respect to, the micro-configurations upon which they depend" (2021: 2). Special science entities, in short, present both (i) cotemporal material dependence on micro-configurations, and (ii) ontological and causal

Argumenta 10, 1 (2024): 291—302
ISSN 2465-2334

DOI 10.14275/2465-2334/20230.onn
First published: 3 July 2023

autonomy. The coupling of these features provisionally defines metaphysical emergence because (i) and (ii) are real features of the entities at issue.

The compatibility between dependence and autonomy in special science entities, however, is a debated issue. This compatibility problem, indeed, corresponds to a generalisation of the more specific problem of nonreductive materialism highlighted by Jaegwon Kim. This issue arises from embracing both ontological physicalism (the claim that all is physical) and property dualism (the claim that psychological properties belong to a domain which is autonomous and irreducible to the physical one (1989: 32)). The topic that Wilson is addressing is a generalisation of Kim's problem because she is not just interested in *mental* properties and powers, but in a wider range of higher-level entities, such as cells, organs, trees, birds, humans, and so on (2021: 1). The autonomy of these phenomena, however, is under the same threat as the mental properties discussed by Kim, because recognising their autonomy requires solving the so-called "problem of higher-level causation".

The problem was first presented by Kim in 1989, when he argued that no physicalist worthy of the name can be a nonreductionist about psychological phenomena. Kim's analysis proceeds as follows. Nonreductionists accept physicalism. Hence, they accept the so-called "causal closure of the physical", i.e., the assumption that every physical event has a sufficient physical cause. This means that "if we trace the causal ancestry of a physical event, we need never go outside the physical domain" (1989: 43). Consequently, nonreductionists admit that physical events can have only physical causes. However, they reject eliminativism, and are therefore realists about mental properties. This entails that to grant a legitimate existence to mental properties, nonreductionists must find a causal work that is done by mental properties *qua* mental properties (we will soon see why, in Kim's view, it must be so).

Yet nonreductionists already subscribed to the causal closure of the physical, so they seem to come to a dead end: if mental phenomena exert a genuine causal efficacy, then the causal closure of the physical is violated (in addition to the problem of overdetermination, because the effect of a mental cause must have a physical cause as well). If the causal closure is respected, on the contrary, mental phenomena have no genuine causal efficacy and, consequently, no genuine existence. In light of this, Kim concludes that a physicalist has to be either a reductionist or an eliminativist, for she has to reject the distinct autonomy of the mental or the mental *tout court*.

Before turning to Jessica Wilson's presentation of the problem, a relevant remark is in order. Among the premises that lead to the nonreductionists' dead end, Kim briefly mentioned the idea that "to be a mental realist, […] mental properties must be *causal properties*" (1989: 43). Kim fully formulated this principle in a later paper focused again on nonreductionists' troubles with mental causation (2006). Here, Kim asks: "[…] what does the commitment to the reality of mental properties amount to? What is the significance of saying of anything that it is real?" (2006: 436). In Kim's opinion, the answer to these questions is provided by the British Emergentist Samuel Alexander, for whom "To be real is to have causal powers" (ibid.). Kim named this principle "Alexander's Dictum" and its importance within the problem of higher-level causation is evident. If the principle is rejected, entities can have a legitimate existence even without exerting causal efficacy. If the nonreductive physicalist has to give up her nonreductionism,

therefore, it is because of Alexander's Dictum. Let's now turn to Jessica Wilson's formulation and treatment of Kim's problem.

2. The Problem of Higher-Level Causation

As already mentioned, Wilson considers the problem of higher-level causation as "the most pressing challenge to taking the appearances of emergent structures as genuine" (2021: 39). The problem, also known as the overdetermination or the exclusion problem,[1] lies in the apparent impossibility, for a higher-level entity, to be distinctively efficacious in a world where every physical effect is supposed to have an equally physical cause. In this framework, if a non-physical cause is admitted, it follows that the same effect has two sufficient causes, leading to a case of causal overdetermination.

For Wilson, the problem presented by Kim can be exhaustively rephrased starting from six premises. Four of them—*Dependence*, *Reality*, *Efficacy*, and *Distinction* (1-4)—are claims about the nature of higher-level entities; the remaining two—*Physical Causal Closure* and *Non-overdetermination* (5-6)—concern the nature of causation. The premises are the following:

> (1) *Dependence*. Special-science features cotemporally materially depend on lower-level physical features […] in such a way that, at a minimum, the occurrence of a given special-science feature on a given occasion minimally nomologically supervenes on base features on that occasion.
> (2) *Reality*. Both special-science features and their base features are real.
> (3) *Efficacy*. Special-science features are causally efficacious.
> (4) *Distinctness*. Special-science features are distinct from their base features. […]
> (5) *Physical Causal Closure*. Every lower-level physical effect has a sufficient purely lower-level physical cause. […]
> (6) *Non-overdetermination*. Except for cases of the double-rock-throw variety, effects are not causally overdetermined by distinct individually sufficient cotemporal causes (Wilson 2021: 41).

Wilson notices that accepting the dependence, reality, efficacy, and distinctness of special science entities implies the failure of one of the two other premises, and the same can be said about the commitment to the last two premises: if both *Physical Causal Closure* and *Non-overdetermination* are accepted, at least one of the features of special science entities listed above must go.

To solve the problem of higher-level causation there are different strategies, each coinciding with the rejection of one or more premises of the list. In Wilson's opinion, substance dualism rejects *Dependence*, eliminativism *Reality*, epiphenomenalism *Efficacy*, and reductive physicalism *Distinctness*. All these strategies succeed in preserving *Physical Causal Closure* and the *Non-overdetermination* requirement, but they do so by weakening the ontological and causal autonomy of special science entities. Wilson's strategy, conversely, consists in accepting the first four premises about higher-level phenomena, alternatively denying the legitimacy of the other two premises. By doing so, she defines her two schemas for emergence. The rejection of *Physical Causal Closure* leads to Strong Emergence, while that of *Non-overdetermination* leads to Weak Emergence. As we will see in the next paragraph, the first produces a metaphysical position that is not compatible with physicalism, while the

[1] Wilson refers to Kim's (1993) and Merricks' (2003) formulations of the argument.

second allows for a position that is compatible with it. In short, Wilson accepts the structure of Kim's argument, but chooses to reject a different premise than the one chosen by Kim and builds her models of emergence starting from this move.

3. Wilson's Two Schemas for Strong and Weak Emergence

In her book, Wilson poses two key questions. The first is what is emergence, while the second is whether there are real cases of emergence in nature. To answer these questions, while curbing the detrimental effects of the problem of higher-level causality, she designs her two schemas for metaphysical emergence.

The forms of emergence she recognises depend upon the satisfaction of two conditions, the *New Power Condition*, and the *Proper Subset of Powers Condition*. The fulfilment of the first one leads to Strong emergence, while the fulfilment of the second one leads to Weak emergence.

3.1 Strong Emergence

The *New Power Condition* states the following:

> *New Power Condition*: Token feature *S* has, on a given occasion, at least one token power not identical with any token power of the token feature *P* upon which *S* cotemporally materially depends, on that occasion (Wilson 2021: 51).

In this case, to fulfil the condition, it is necessary that the higher-level feature *S* has at least one power that its lower-level base feature *P*, on which *S* materially depends, does not have. If this feature *S* has this new power, then that feature can be considered Strongly metaphysically emergent.

The point to clarify, here, is how the fulfilment of the *New Power Condition* leads to Strong emergence. The answer is that a feature having a new fundamental power cannot (by Leibniz's law) be identical to a feature that does not exert that power. The argument leads, therefore, to the ontological autonomy of the feature at issue. As for its causal autonomy, the argument is much the same. The higher-level feature having a novel power can produce an effect that its base feature cannot because the latter has different powers. Being therefore both ontologically and causally distinct because of the presence of a new power, the feature fulfilling the *New Power Condition* is Strongly metaphysically emergent. In Wilson's words:

> *Strong emergence*: What it is for token feature *S* to be Strongly metaphysically emergent from token feature *P* on a given occasion is for it to be the case, on that occasion, (i) that *S* cotemporally materially depends on *P*, and (ii) that *S* has at least one token power not identical with any token power of *P* (Wilson 2021: 53).

3.2 Weak Emergence

Let's turn to the second schema. The *Proper Subset of Powers Condition* states the following:

> *Proper Subset of Powers Condition*: Token feature *S* has, on a given occasion, a non-empty proper subset of the token powers of the token feature *P* on which *S* cotemporally materially depends, on that occasion (Wilson 2021: 59).

To fulfil the condition, it is necessary that the higher-level feature *S* has a proper subset of the powers possessed by the lower-level base feature *P* on which *S* one materially depends. If the feature at issue has this proper subset of powers, then the feature can be considered Weakly metaphysically emergent.

Similarly to the case of the *New Power Condition*, the fulfilment of the *Proper Subset Condition* entails both ontological and causal distinctness of the higher-level feature. Having different sets of powers, the higher-level and the lower-level features will be ontologically distinct by Leibniz's law and will produce different effects, having causal distinctness due to their different causal profiles (2021: 79). In Wilson's words:

> *Weak emergence*: What it is for token feature *S* to be Weakly metaphysically emergent from token feature *P* on a given occasion is for it to be the case, on that occasion, (i) that *S* cotemporally materially depends on *P*, and (ii) that *S* has a non-empty proper subset of the token powers had by *P* (Wilson 2021: 72).

3.3 How to Be Causally Effective?

As the schemas show, for Wilson it is possible to save the distinctness and causal efficacy of special science entities having (at least) one novel causal power—as in the fulfilment of the *New Power Condition*—or having "a distinctive set (collection, plurality) of powers" (2021: 79)—as in the fulfilment of the *Proper Subset of Powers Condition*. There are therefore two ways in which a higher-level feature—and a special-science entity—can be causally autonomous: it "may have more powers than its base feature", or, alternatively, "fewer powers than its base feature" (2021: 74). If the emergent entity has more powers, some genuine causal novelty appears and violates the Causal Closure. If it has fewer powers, no real causal novelty is involved, but the difference in features and powers had by the entity ensures its ontological and causal autonomy.

In Wilson's opinion, therefore, these are the only two ways in which a higher-level entity can be genuinely efficacious, and for this reason she thinks that every viable account of emergence offered by the literature can be rephrased in her two schemas, which represent the only two appropriate responses to the problem of higher-level causation.

4. Questioning, Rather Than Responding To, the Problem of Higher-Level Causation

In the first paragraph, I described the premises recognised by Kim as underlying the problem of higher-level causation. These are (i) ontological physicalism, (ii) mental realism, and (iii) Alexander's Dictum. These three premises give rise to five of the six premises listed by Wilson. Roughly, *Dependence* and *Physical Causal Closure* originate from ontological physicalism; *Reality* and *Distinctness* descend from mental realism; finally, *Efficacy* derives from the coupling of mental realism with Alexander's Dictum. The sixth premise, *Non-overdetermination*, is independent from the others and is the (unacceptable) consequence, in Kim's opinion, of nonreductionist assumptions. As already suggested, Wilson's and Kim's views about the problem of higher-level causation are structurally similar, even if they solve the problem differently, with Kim rejecting *Distinctness* and Wilson rejecting, alternatively, *Physical Causal Closure* or *Non-overdetermination*.

However, some details of these arguments can be questioned, and in this paper, I would like to focus on those involved with the acceptance of Alexander's Dictum. Specifically, there are three issues that need to be addressed. The first one concerns the Dictum itself: one may want to reject it and assume other criteria about existence. The second one is about the power-based interpretation of the Dictum: one may want to accept the latter, while considering its power-based interpretation as too strict. The third one is about the metaphysical underdetermination of the powers involved in the power-based interpretation: one may want to accept the Dictum and its power-based interpretation, while requiring a differentiation between microscopic physical powers and macroscopic emergent powers. In the next paragraphs, I will examine each of these issues, suggesting that a less strict and univocal view of existence and causal efficacy might render the problem of higher-level causation less "pressing".

4.1 Alexander's Dictum

The first issue is presented here for the sake of the argument, because I think that Alexander's Dictum is reasonable and convincing. I will start with a quick overview about it.

The Dictum is a reformulation of what is known as the Eleatic principle, which owes its name to the visitor coming from Elea who discusses with Theaetetus in Plato's *Sophist* (Oddie 1982). Towards the end of the dialogue, the Eleatic Visitor describes the so-called "battle of gods and giants" (*Soph*. 246e-249d), namely a dispute over the nature of being in which two contrasting views can be recognised. The first one is that assumed by the Gods, i.e., the friends of the forms, who are committed to their immaterial existence; the second, the Giants, are the "earth people", who only grant existence to material and tangible bodies (Assaturian 2021). The Giants' criterion for reality, which can be roughly formulated as "being is being tangible", poses a serious problem: if only tangible bodies exist, how can virtues or souls be accommodated in the resulting ontology? How can something like justice influence the behaviour of the individual, if justice has no tangible body? In this frame, the Eleatic Visitor tries to make the Giants' views more coherent, suggesting that their criterion for reality might be improved. In doing so, he enunciates the Eleatic principle, according to which everything that really is must possess some power or capacity ("τὸ καὶ ὁποιανοῦν τινα κεκτημένον δύναμιν", 246a). The Eleatic principle, therefore, suggests that being, rather than being equivalent to tangibility, is equivalent to having some sort of causal capacity.

Now, the principle (or the Dictum) seems reasonable and convincing because an existing entity unable to produce any sort of causal effects would be hardly conceivable. Still, one might reject it and assume other criteria for existence. Without going too far, while examining free will, Wilson writes that a good reason to take free will at realistic face value is our direct introspective access to it. The fact that we "experience ourselves as seeming to freely choose, in ways transcending any nomological (deterministic or indeterministic) goings-on" (2021: 278) is therefore enough for accepting the genuine existence of free will. Wilson states that "in the absence of good reasons to think that our experience of nomologically transcendent free will cannot be taken at face value, we are entitled to take this experience at realistic face value" (2021: 278). Direct introspective access, therefore, seems a valid criterion for the existence of free will and is different

from Alexander's Dictum, as different as other criteria that have been formulated during the history of philosophy—e.g., being tangible or admitting direct epistemic access, as we already saw, but also being indispensable to our scientific theories (Putnam 1979; Quine 1980), being robust (Levins 1966; Wimsatt 1981 and 1994), and so on. Alexander's Dictum, in short, is not the only reasonable criterion for existence, and admitting other criteria seems to make the problem of higher-level causation less challenging.

4.2 The Power-Based Interpretation of Alexander's Dictum

As mentioned, it is possible and legitimate to assume Alexander's Dictum, namely the principle whereby existence corresponds to the capacity of being causally efficacious. Kim's formulation of the Dictum, however, does not merely equate existence and causal efficacy in general, but rather being with the exertion of causal powers.

This stricter equation might nonetheless be problematic for at least two reasons. The first is historical. As already noticed, Kim states that in Samuel Alexander's opinion being is having some causal powers (2006),[2] but this attribution originated from a misunderstanding. In *Space, Time and Deity* (1920), Alexander expresses an anti-epiphenomenalist position on consciousness, stating that epiphenomenalism is to be rejected (among other reasons) because "it supposes something to exist in nature which has nothing to do, no purpose to serve, a species of noblesse which depends on the work of its inferiors, but is kept for show and might as well, and undoubtedly would in time be abolished" (1920: Vol. II, 8). Kim translates this passage into a power-based vocabulary, but this approach does not reflect Alexander's intentions, as his view of causation was closer to that of Hume than to that of Aristotle. For Alexander, in other terms, causation does not correspond to the exertion of causal powers, but to the relationship of continuity and succession that exist between different regions of Space-Time—the fundamental element of Alexander's metaphysical monism. In *Space, Time and Deity*, Alexander clearly expresses his aversion to the concept of causal power, which, in his view (as also in Hume's), cannot be admitted in our ontologies:

> If all we observe in external events is uniform succession, to impute to one of them a power to produce the other is a fiction, the fiction which Hume set himself to discredit. It may be serviceable anthropomorphism, but it is not science nor philosophy. If there is no power traceable in things, then there is none (1920: 188).[3]

However, Kim is not the only one attributing to British Emergentists some sort of theory of causal powers; Robert McLaughlin did the same in his well-known and

[2] See also Kim: "Prominent […] is the claim that the emergents bring into the world new causal powers of their own, and, in particular, that they have powers to influence and control the direction of the lower-level processes from which they emerge. This is a fundamental tenet of emergentism, not only in the classic emergentism of Samuel Alexander, Lloyd Morgan, and others but also in its various modern versions" (Kim 1999: 5-6).

[3] A little further, Alexander adds: "causality is not the work of power" (1920: 290) and then he goes on to say "The mischief of the conception that a cause has power to produce its effect is that it introduces some mysterious element of connection other than that of simple continuity" (Alexander 1920: 291).

influential paper about the rise and fall of British Emergentism (1992).[4] The problem with these misreadings is that the power-based interpretation, even if only sketched, is not metaphysically neutral (besides being historically inaccurate) and can be misleading.

On the one hand, therefore, the British Emergentists were not committed to a power-based view of emergent causal efficacy. On the other hand, this account of causation might not be the most appropriate for conceptualizing emergence, given its central role in reductionist—i.e., anti-emergentist—strategies. This brings us to the second problem with the power-based interpretation of Alexander's Dictum.

Starting from Kim's causal inheritance principle (1993) and arriving at Elanor Taylor's collapse objection (2015), the notion of causal power has played a pivotal role in strategies aimed at excluding the possibility of higher-level causal efficacy. Kim's causal inheritance principle suggests that higher-level causal efficacy is not genuine, but is derivative from the lower-level by means of the inheritance of lower-level causal powers:

> **Causal Inheritance Principle** (CIP): If mental property M is realized in a system at time t in virtue of physical realization base P, the causal powers of this instance of M are identical with the causal powers of P (Kim 1993: 326).

Taylor's argument (2015), instead, focuses on latent dispositional properties. In her view, higher-level causal efficacy is not genuine because the alleged causal powers of emergent, higher-level phenomena correspond to the dispositional properties belonging to the low-level components on which the emergent phenomena depend. These dispositional properties are latent when the components are in isolation, and their effects become manifest only when they are organised in complex manners: hence the illusion that these properties belong to a higher-level.

What I am suggesting here is that the concept of causal power is central to classic reductionist strategies and seems to already carry anti-emergentist implications. Its introduction into the emergentist debate, moreover, is recent and appears to be related to the recovery of the notion of emergence as an alternative view to contemporary reductionism and physicalism. However, this emergence *vs*. reduction battle is played out within the framework of the latter and draws upon its conceptual repertoire, referring to issues such as realisation, dispositionalism, causal inheritance, and so on. Reading—or re-reading—the emergentist debate in this contemporary key is not necessarily a bad thing, but it is important to recognise that doing so is not metaphysically neutral, nor is it the only approach available.

[4] See McLaughlin (1992: 20): "British emergentism maintains that some special science kinds from each special science can be wholly composed of types of structures of material particles that endow the kinds in question with fundamental causal powers. Subtleties aside, the powers in question emerge' from the types of structures in question". McLaughlin cites C.D. Broad, who indeed uses the term 'power' more than Alexander does. A careful reading of Broad's passages in which the term power is used, however, shows that the term is employed in a non-technical way. Broad, who is referenced by Alexander, similarly believes that causation is a matter of regularity, uniformity, and continuity between spatiotemporal regions (see Broad 1925: 454-56).

There are different interpretations of the Eleatic principle—Samuel Alexander and the British Emergentists provided at least one—and these alternatives seem to make the problem of higher-level causation less challenging.

4.3 The Metaphysical Underdetermination of the Power-Based Interpretation of Alexander's Dictum

While it is perfectly possible to accept both Alexander's Dictum and its power-based interpretation, describing emergent causal efficacy in power-based terms might lead to new problems, rather than solving old ones.

Admitting emergent causal powers seems to naturally raise questions about their nature, namely about what kind of powers they are and whether these emergent powers are different from non-emergent ones.

In the first chapters of *Metaphysical Emergence*, Wilson provides some characterisations of these powers by stating that they are fundamentally novel—this is the reason why Strong emergence is incompatible with physicalism. As for fundamentality, Wilson defines it in primitivist terms: the fundamental is simply what God had to create (2014 and 2021). Wilson adds, however, that a nonfundamental power is a summation or aggregation of already existing lower-level powers (2021: 48), so fundamentality is also defined in terms of compositional basicness: a fundamentally novel power is a non-aggregative power.

Fundamentality, however, does not exhaustively define higher-level causal powers, because microphysical causal powers (those possessed by the emergence base) are fundamental as well. At a first glance, therefore, higher-level causal powers seem to differ from lower-level ones simply by being at a different level.

Further information about these novel powers can be gathered in another passage from *Metaphysical Emergence*. Emergent powers may be intended as grounded in fundamental interactions that are different from physical fundamental interactions (i.e., interactions other than strong and weak interactions, electromagnetism, and gravity) (2021: 133).

These suggestions, however, do not really clarify the nature of these emergent powers, how they act, and how they are exerted by their bearers. Wilson simply states that Strong emergence corresponds to the fulfilment of the condition of having (at least) one novel causal power, but what this power is, is left programmatically undiscussed. For Wilson, that of power is an "operative notion [that is] metaphysically highly neutral" (2021: 32) and "no 'heavyweight' notion of powers or causation need be presupposed" (2021: 33).

Now, the absence of a precise description of emergent powers seems to indicate that there is no relevant difference, in Wilson's view, between lower-level and higher-level causal powers. In other words, it may be reasonable to assume that if there had been a relevant difference, Wilson would have highlighted it.

However, by leaving the power-based interpretation of causal efficacy metaphysically underdetermined and disregarding the hypothesis that emergent causal powers might be relevantly different from low-level ones, two suggestions emerge. First, powers are conceived as a sort of universal and undifferentiated currency for causal processes, regardless of the ontological domain in which they appear. Second, this currency is not "bearer sensitive". Even if emergent properties and entities are different from the properties and entities from which they emerge, the powers of the former are not relevantly different from those of the latter. Here, I use the word "relevant"—or "relevantly"—repeatedly because low-level and

high-level causal powers are obviously different in some way, but the crucial difference I am pointing out is not just any difference, but a difference in kind that might be able to weaken the problem of high-level causation.

By examining the nature of causal powers, for instance, it might be discovered that higher-level powers cannot really collapse, while lower-level ones cannot really emerge. Emergent and non-emergent causal powers, in other words, might simply be non-interchangeable powers of a different kind. Let's try to develop this hypothesis.

Traditional (non-emergent) causal powers are often intended as fundamental, (micro)physical powers. A classic example of these powers is the electron's charge, which is mentioned by several authors involved in the debate (Psillos 2006; Marmodoro 2010 and 2013; Engelhard 2010; Williams 2019) and has peculiar properties that are commonly—though not universally—attributed to powers: being fundamental, essential, intrinsic, intrinsically active, and productive. These features accurately describe many microphysical powers, but macroscopic powers seem more difficult to describe in these terms. Defining the electron's charge as a causal power, in short, seems simpler and more accurate than defining my ability to roller-skate as one.

Emergent causal powers, despite being sometimes intended as ontologically fundamental (Wilson 2021; Barnes 2012), are often conceived as nonfundamental, extrinsic, context-sensitive, and constraining (Thorpe 1974; Mitchell 2012; Gillett 2016; Onnis 2021). These properties appear to be not intrinsically causal but rather determinative in a different (perhaps weaker) sense. Carl Gillett (2016), for instance, defines the causal efficacy of emergent phenomena as a role-shaping, non-productive determination which he dubs "machresis". In his framework, machresis is a "non-powerful" relationship that does not involve the exercise of active and productive causal properties but constrains the already existing contributions of the latter, and in so doing determines reality in "making a difference" to the world. The most striking difference between micropowers and emergent powers would therefore be the intrinsic activity and productivity of the former and the extrinsic non-productive constraining capacities of the latter.

It should be noted that the previous analysis is a preliminary and brief examination of the possible differences between non-emergent and emergent powers. However, it might be useful to engage in a more thorough investigation because powers can easily collapse if they are understood as properties that can be indifferently instantiated at both higher and lower levels. Conversely, differentiating between micropowers and macropowers might make this collapse more difficult. For instance, let's suppose that the macroscopic causal powers exerted by a biological complex system require a biological complex bearer. In that case, a non-biological system or a biological isolated component could not instantiate those macropowers, which would therefore become non-collapsible.

Ultimately, overcoming the metaphysical underdetermination of the power-based view by recognising relevant ontological differences between micropowers and macropowers appears to be another promising approach to making the problem of higher-level causation less challenging.

5. Conclusions

In *Metaphysical Emergence*, Jessica Wilson recognises the problem of higher-level causation as "the most pressing challenge to taking the appearances of emergent

structure as genuine" (2021: 39). As I have attempted to show in this paper, the problem might be less "pressing" if emergence were related to a less strict and univocal view of existence and causal efficacy, and to a more detailed examination of the nature of causal powers.

References

Alexander, S. 1920, *Space, Time, and Deity*, New York: Macmillan.

Assaturian, S. 2021, "What's Eleatic About the Eleatic Principle?", *Archai*, 31, e-03122.

Barnes, E. 2012, "Emergence and Fundamentality", *Mind*, 121, 873-901.

Broad, C.D. 1925, *The Mind and Its Place in Nature*, Brace: Harcourt.

Engelhard, K. 2010, "Categories and the Ontology of Powers", in Marmodoro 2010: 49-65.

Gillett, C. 2016, *Reduction and Emergence in Science and Philosophy*, Cambridge: Cambridge University Press.

Kim, J. 1989, "The Myth of Nonreductive Materialism", *Proceedings and Addresses of the American Philosophical Association*, 63, 3, 31-47.

Kim, J. 1993, "The Non-Reductivist's Troubles with Mental Causation", in Heil, J. and Mele, A.R. (eds.), *Mental Causation*, Oxford: Clarendon Press, 189-210.

Kim, J. 1999, "Making Sense of Emergence", *Philosophical Studies*, 95, 1-2, 3-36.

Kim, J. 2006, "Emergence: Core Ideas and Issues", *Synthese*, 151, 3, 547-59.

Levins, R. 1966, "The Strategy of Model Building in Population Biology", *American Scientist*, 54, 421-31.

Marmodoro, A. (ed.) 2010, "The Metaphysics of Powers: Their Grounding and Their Manifestations", Abingdon: Routledge.

Marmodoro, A. 2013, "Aristotelian Powers at Work: Reciprocity Without Symmetry in Causation", in Jacobs, J.D. (ed.), *Causal Powers*, Oxford: Oxford University Press, 57-76.

McLaughlin, B. 1992, "The Rise and Fall of British Emergentism", in Bedau, M. and Humphreys, P. (eds.) 2008, *Emergence. Contemporary Readings in Philosophy and Science*, Cambridge, MA: MIT Press, 19-59.

Merricks, T. 2003, *Objects and Persons*, Oxford: Clarendon Press.

Mitchell, S. 2012, "Emergence: Logical, Functional and Dynamical", *Synthese*, 185, 171-86.

Oddie, G. 1982, "Armstrong on the Eleatic Principle and Abstract Entities", *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 41, 2, 285-95.

Onnis, E. 2021, *Metafisica dell'emergenza*, Turin: Rosenberg & Sellier.

Psillos, S. 2006, "What Do Powers Do When They Are Not Manifested?", *Philosophy and Phenomenological Research*, 72, 1, 137-56.

Putnam, H. 1979, *Mathematics, Matter and Method*, *Philosophical Papers*, Vol. 1, Cambridge: Cambridge University Press.

Quine, W.V.O. 1980, *From a Logical Point of View: Nine Logico-Philosophical Essays*, Cambridge, MA: Harvard University Press.

Taylor, E. 2015, "Collapsing Emergence", *The Philosophical Quarterly*, 65, 261, 732-53.

Thorpe, W.H. 1974, "Reductionism in Biology", in Ayala, F., Ayala, F.J. and Dobzhansky, T. (eds.), *Studies in the Philosophy of Biology: Reduction and Related Problems*, Berkeley: University of California Press, 109-38.

Williams, N.E. 2019, *The Powers Metaphysic*, Oxford: Oxford University Press.

Wilson, J. 2014, "No Work for a Theory of Grounding", *Inquiry*, 57, 5-6, 535-79.

Wilson, J. 2015, "Metaphysical Emergence: Weak and Strong", in Bigaj, T. and Wüthrich, C. (eds.), *Metaphysics in Contemporary Physics*, Leiden: Brill, 251-306.

Wilson, J. 2021, *Metaphysical Emergence*, Oxford: Oxford University Press.

Wimsatt, W.C. 1981, "Robustness, Reliability, and Overdetermination", in Brewer, M. and Collins, B. (eds.), *Scientific Inquiry and the Social Sciences*, New York: Jossey-Bass, 124-63; repr. in Wimsatt 2007: 43-74.

Wimsatt, W.C. 1994, "The Ontology of Complex Systems: Levels of Organization, Perspectives, and Causal Thickets", *Canadian Journal of Philosophy*, 20, 207-74; repr. in Wimsatt 2007: 193-240.

Wimsatt, W.C. 2007, *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*, Cambridge, MA: Harvard University Press.

Not So Weak Emergence

Michele Paolini Paoletti

University of Macerata

Abstract

In this article, I shall examine Jessica Wilson's schema for weak emergence in connection with two questions: why are only certain proper subsets of the powers borne by lower-level features associated with higher-level, weakly emergent features? Why is a certain proper subset of the powers borne by a given lower-level feature associated with a certain higher-level, weakly emergent feature, and vice versa? I shall consider and criticize four possible answers to these questions, including Wilson's own view. Finally, I shall suggest my own solution, which is based on something akin to grounding categoricalism. I shall also explore some consequences of accepting my view.

Keywords: Emergence, Physicalism, Grounding categoricalism, Powers, Subset account.

1. Introduction

I shall discuss in this contribution Jessica Wilson's schema for weak emergence. I shall show that this schema comes together with two crucial questions. First question: why are only certain proper subsets of the powers borne by lower-level features associated with higher-level, weakly emergent features? Second question: why is a certain proper subset of the powers borne by a given lower-level feature associated with a certain higher-level, weakly emergent feature, and vice versa?

I shall show that answering such questions implies that one rediscusses, *inter alia*, the compatibility between weak emergence and physicalism. In Section 2 I shall briefly introduce Wilson's schema for weak emergence and the two questions I anticipated above. In Section 3 I shall consider three ways of answering (or dissolving) such questions: the suggestion that they ask for explanations of modal facts; primitivism; deflationism about powers. I shall criticize each way. In Section 4 I shall examine and discuss Wilson's own view. Finally, in Section 5, I shall suggest that one should embrace—with respect to higher-level, weakly emergent features and the powers they confer—something akin to grounding categoricalism. I shall also explore some consequences of accepting this view.

Argumenta 10, 1 (2024): 303—312
ISSN 2465-2334

DOI 10.14275/2465-2334/202418.pao
First published: 30 November 2024

2. Weak Emergence and the Two Questions

Jessica Wilson (2021: 72) presents the following schema for weak emergence:

> (WE) a token feature S weakly emerges (on a given occasion) from a token feature P if and only if, on that occasion, (i) S cotemporally materially depends on P and (ii) S has a non-empty proper subset of the token powers had by P.

Token features are particular property-instances. The properties involved in S and P are properties that belong to different levels of the universe. Cotemporal material dependence may be interpreted in different ways, depending on one's favorite theory of ontological dependence. Finally, token powers need *not* be taken as *sui generis* entities, to be distinguished from P, S and their particular instances. For example, on a deflationary view of token powers, the latter may be taken as descriptions of what token features S and P are able to cause in specific circumstances.

In this contribution, I shall dwell on condition (ii). I shall extend the discussion a bit beyond Wilson's original project of providing a schema for weak emergence. And I shall introduce further issues concerning weak emergence and its compatibility with physicalism.

On condition (ii), token feature P has a certain set of token powers associated with it. Assume that this set includes four token powers: p_1, p_2, p_3 and p_4. Following (ii), token feature S has another set of token powers associated with it. Crucially enough, the latter set includes some, but not all of the token powers associated with token feature P. Namely, the set of token powers associated with token feature S is only a proper subset of the set of token powers associated with token feature P. Assume that the set of token powers associated with token feature S includes three token powers: p_1, p_2 and p_3.

This guarantees that, on the one hand, token feature S is *not* endowed with any novel power with respect to the token feature P on which it depends. If token feature P and all of its powers are physical, the weak emergence of S from P is fully compatible with the acceptance of physicalism. Yet, on the other hand, token feature S has a distinctive causal profile with respect to token feature P. Indeed, the distinctive causal profile of S is associated with distinctive laws of nature and distinctive difference-making considerations.

So far, so good. Let me recall the set of powers associated with P, i.e., p_1, p_2, p_3 and p_4. Call this set the "causal role of P". And the proper subset of powers associated with S, i.e., p_1, p_2 and p_3. Call this proper subset the "causal role of S". Three questions arise.

First question: are *all* of the proper subsets of the causal role of P associated with higher-level token features such as S? For example, is there a token feature S_1 associated with p_1 and p_2, another token feature S_2 associated with p_2 and p_3, and so on?

It seems that the answer to this question must be negative. *Not all* of the proper subsets of the causal role of P are associated with higher-level token features. In most cases, *only some* proper subsets are. In our example, only the proper subset including p_1, p_2 and p_3 is associated with a higher-level token feature such as S. Otherwise, we may turn to postulate the existence of higher-level token features that are scientifically irrelevant. Indeed, their distinctive causal profiles/causal roles may be associated with no distinctive law of nature and no distinctive difference-making

consideration. Thus, such higher-level token features would find no place in the best theories of special sciences.

We grant that *only some* proper subsets of the causal role of P are associated with higher-level token features such as S. In our example, only the proper subset including p_1, p_2 and p_3 (i.e., the causal role of S) is associated with a higher-level token feature, i.e., S itself. The next question is: why is the proper subset made of p_1, p_2 and p_3 the only one (in our case) that is associated with *a* higher-level token feature? Namely, why is it the only one that is relevant for the weak emergence of *a* higher-level token feature?

Another question is in order. Even if we concede—*contra hypothesin*—that every proper subset is associated with a higher-level token feature such as S, it seems that the proper subset made of p_1, p_2 and p_3 is the *only one* that is associated *with S*. And it is associated *only with S*. This seems to happen in the actual world not by sheer coincidence, but at least as a matter of nomological necessity. Thus, why is the very proper subset made of p_1, p_2 and p_3 (i.e., the causal role of S) the only one that is associated with S—and only with S? Why is it not associated with any other higher-level token feature? More strongly: why *can't* it be associated—at least as a matter of nomological necessity—with any other higher-level token feature? And why *can't* S have—at least a matter of nomological necessity—any other proper subset of powers associated with it, i.e., any other causal role? In sum, why must S and its causal role be associated with each other (*and only with each other*) at least as a matter of nomological necessity?

We have two questions to face:

1. Why is the proper subset made of p_1, p_2 and p_3 the only one that is associated with a higher-level token feature?
2. Why must S and the proper subset made of p_1, p_2 and p_3 (i.e., its causal role) be associated with each other (and only with each other) at least as a matter of nomological necessity?[1]

3. Three Attempts

These questions cannot be dismissed by claiming that they look for explanations of *modal* facts. First of all, question (1) is not explicitly put in modal terms. Moreover, many questions in the business of metaphysics and philosophy of science are actually put in modal terms, insofar as they ask for explanations of what *can* and *cannot* happen.

Suppose now that, in order to answer both questions, we embrace some sort of *primitivism*. Namely, suppose that we claim that it is a primitive and inexplicable fact of the matter that the proper subset made of p_1, p_2 and p_3 (i.e., the causal role of S) is the only one that is associated with a higher-level token feature. And, more crucially, that that proper subset is only associated with token feature S and S is only associated with that proper subset.

[1] Elder (2004 and 2011) considers similar questions with respect to the restricted composition of everyday objects and with respect to micro-physical causation. In a similar vein, Inman (2018) raises the following problem with respect to the essences of natural substantial kinds: if such essences were nothing but sets of specific properties, why would such properties be unified/clustered together? He criticizes several attempts to solve this problem, e.g., by appealing to homeostatic mechanisms or to specific laws of nature. And, as we shall see, he embraces a non-reductionist solution similar to the one I suggest here.

To make sense of this situation from an ontological standpoint, we may hold that there is some irreducible relation R that links S (and only S) with its causal role (and only with it). Consider now P, i.e., the physical, lower-level token feature. As far as P and its token powers are concerned, R does *not* link any other proper subset of those powers with any other higher-level token feature. Moreover, that R holds between S and its causal role has no further metaphysical explanation. Finally, R may be taken as a nomologically necessitating relation, i.e., as a relation that implies certain nomologically necessary goings-on. This seems to answer both questions.

There are three problems with primitivism. The first problem is that it seems to overpopulate our ontology with many irreducible facts of the matter such as: the fact that R holds between S and the very causal role associated with it.

Secondly, such facts are *not* enough in order to answer question (2). It is *not* enough that R holds between S and its causal role in order to guarantee that S is *only* associated with that role and that role is *only* associated with S. In a given possible world, R may hold between S and its (actual) causal role. But it may *also* hold between S and another causal role. In another possible world, R may *not* hold between S and its (actual) causal role, but between S and another causal role. In sum, there should be something else (a negative fact? A totality fact?) that guarantees that S is *only* associated with its causal role and its causal role is *only* associated with S—both in a given possible world and across possible worlds.

Thirdly and finally, that R holds between S and its causal role is an irreducible fact of the matter. Thus, it is a *fundamental* fact. Moreover, this fact constitutively includes a non-physical token feature such as S. Thus, there are fundamental facts with non-physical token features such as S. The constituents of fundamental facts are fundamental.[2] Therefore, non-physical token features such as S are fundamental.

This conclusion may be hard to swallow for physicalists. True: on one plausible interpretation of physicalism (the one embraced by Wilson 2021), physicalism is only taken to hold that the only powers existing in the (actual) universe are physical powers primarily and non-derivatively borne and exercised by physical entities. Therefore, according to this interpretation, every causal going-on turns out to be exhaustively produced and explained by physical powers. This version of physicalism is fully compatible with there being fundamental facts such as: the fact that R holds between S and its causal role. It is also compatible with S's being a fundamental entity, insofar as S is not endowed with novel powers.

However, that R holds between S and its causal role is *not* a purely physical fact. The former also includes S, which is non-physical. Moreover, that R holds between S and its causal role cannot be fully explained in fully physical terms, since it is a fundamental fact. Thus, that R holds between S and its causal role is at odds with a stronger version of physicalism, according to which everything (at least in the actual universe) is physical or can be fully explained in fully physical terms (i.e., in the end, it entirely depends on the physical and only on the physical).

Invoking deflationism about token powers, causal roles and/or properties does not help either. Assume that "S" is nothing but a scientifically relevant but non-physical predicate and the causal role of S is nothing but a complex description of the nomological regularities connected with "S". In this context, it still

[2] See Sider 2011: 126-32.

makes sense to ask why "S" is associated with *a* description of nomological regularities, why it is associated with *that* description and not with other descriptions, why that description is *only* associated with "S", and so on. From the standpoint of physicalists, the answers to such questions should not (irreducibly) invoke non-physical terms and predicates.

Alternatively, one may hold that causal roles are nothing but complex descriptions of possibly regular behaviors, without the need to invoke non-physical predicates such as "S". Fine. Still, some sets of such descriptions may turn out to *correctly* describe the universe and/or be *useful* when describing the universe. And other sets may turn out to be incorrect and/or useless for such purposes. What accounts for the relevant distinction between correct/useful sets of descriptions and incorrect/useless ones? In order to answer this question, one should find some feature or another in the universe. The alternative would be to adopt a radically anti-realist stance on the bearings of such descriptions. But this would be a non-starter for a project on the metaphysics of emergence. And, more importantly, it would leave something unexplained i.e., the fact that only certain sets of descriptions are correct/useful.

4. Wilson's Physicalist Solution

Wilson (2010; 2021: 177-85) puts forward an account of weak emergence based on degrees of freedom. I cannot enter into detail here. Roughly, the idea is that a weakly emergent entity emerges from its base if, *inter alia*, at least one of the degrees of freedom required to characterize its base is eliminated by imposing certain constraints on the base. Such constraints should be entirely placed at the level of the base. In the end, these constraints must be entirely physical or entirely dependent on the physical.

By eliminating specific degrees of freedom, the powers associated with such degrees are eliminated. Thus, weakly emergent entities turn out to have only a proper subset of the powers associated with their bases.

This mechanism is compatible with the acceptance of physicalism, even in its stronger version. Nevertheless, it is necessary to clarify what one means by "physical constraints". Indeed, by "physical constraints", one may first mean "naturalistically acceptable constraints", i.e., constraints that do *not* involve the existence and/or the action of supernatural entities. This understanding is too weak. For it is compatible with the possibility that some of such constraints are irreducibly non-physical and/or result from the exercise of non-physical powers—even if they still belong to the 'natural world'. For example, some of such constraints may irreducibly belong to the biological level of the universe, so that they still belong to the 'natural world', even if they are not physical.

Secondly, by "physical constraints", one may mean "constraints that necessarily operate through and come together with specific physical processes and changes". This understanding is still too weak. Indeed, if one were to believe in irreducible downward causation, some of such constraints could still be non-physical and/or be caused by irreducibly non-physical entities and/or result from the exercise of non-physical powers—insofar as, in all such cases, the relevant constraints operate through and/or are caused through specific physical processes and changes (by downward causation). For example, an irreducibly biological constraint may still operate through and/or be caused through specific physical processes and changes (by downward causation).

The relevant understanding of "physical constraints" at work here is a stronger one. A physical constraint is one that only involves (in itself and in its own causes) entities and processes that are entirely physical[3] and/or entities and processes that entirely depend on further entities and processes that are entirely physical. This understanding of "physical constraints" makes Wilson's mechanism fully compatible with all versions of physicalism. But it may run into the risk of narrowing down the range of weakly emergent phenomena. Some of such phenomena may result from constraints that—for what we know—do *not* clearly satisfy the third characterization of physical constraints. In other terms, we cannot now assume—and we cannot be now sure—that all of the constraints that contribute to weak emergence are such that they only involve entirely physical entities and processes and/or entities and processes that entirely depend on further entities and processes that are entirely physical.

At any rate, with respect to questions (1) and (2), Wilson's mechanism does *not* provide satisfactory answers. First of all, the characterization of weak emergence in terms of degrees of freedom only provides a *sufficient* condition for weak emergence. Thus, it is *not* guaranteed that every weakly emergent entity will arise through this sort of mechanism. Secondly and more importantly, it seems that *not* every possible elimination of the degrees of freedom required to characterize a base is also able to bring about the causal role of a weakly emergent entity (in our case, of a weakly emergent token instance). On the contrary, it seems that only the elimination of *specific* degrees of freedom—and not others—guarantees this result. Why so? Question (1) is left unanswered.

Thirdly and finally, one must still explain why a certain weakly emergent token feature is only associated with a certain causal role and why the latter is only associated with the former. Question (2) is left unanswered.

In reply to this last worry, one may well embrace a view of token features according to which they are nothing but bundles of token powers. Yet, first, one would then be committed to token powers instead of token features. And, secondly, one would still need to explain why *only certain* bundles of token powers (and not others) seem to 'give rise to' or 'be legitimately describable as' token features.

5. Grounding Categoricalism, or Something Near Enough

In my opinion, the best way to answer questions (1) and (2) consists in embracing something akin to 'grounding categoricalism', i.e., the doctrine according to which the causal roles of categorical properties are somehow grounded on those very properties (see, among others, Tugby 2012, 2021, 2022a, 2022b, Yates 2018, Kimpton-Nye 2021 and Paolini Paoletti 2022).

In Paolini Paoletti 2022, I have defended the following form of grounding categoricalism: by virtue of its own essence, the causal role C of a categorical property P (i) is the causal role *of* P, so that it essentially depends (also) on P, (ii) it depends for its origins on P (i.e., it starts to exist as a causal role thanks to P or thanks to the instantiation of P) and (iii) it depends for its continuing to exist (also) on P (i.e., it continues to exist also or only thanks to P or to the instantiation of P). This entails

[3] An entirely physical entity/process is one that, in principle, can be only characterized (with respect to its essence and with respect to all of its features) in physical terms.

that, as a matter of necessity, the existence of C implies the existence of P: necessarily, C cannot exist without P. And it also entails that, as a matter of necessity, C is the causal role of P and of no other property distinct from P.[4]

By the "essence" of something (be it a property or something else), I mean what that entity non-derivatively is (or could be) in all possible circumstances. Namely, the features to be included in the essence of an entity should *not* derive from other features of that entity and they should necessarily come together with that entity whenever it exists. This view of essences is compatible with the possibility that the essence of an entity is identical with that entity or it is only a description of that entity.

My view is compatible with different conceptions of causal roles. Indeed, causal roles may be nothing but descriptions of regular behaviors.

Please also note that, if one believes that all the (nomologically) possible causal roles exist even if they are not associated with any property, one could modify my view as follows: by virtue of its own essence, the causal role C of a categorical property P (i) is the causal role *of* P, so that it essentially depends (also) on P, and (iv) it (also or only) depends on P for its being a causal role that correctly describes the universe and/or that is 'useful' for the purpose of describing the universe. Indeed, not all the (nomologically) possible causal roles that exist correctly describe the universe and/or are 'useful' for this purpose.

At any rate, if, by virtue of its own essence, the causal role C of a categorical property depends in such-and-such a way on P itself, it seems that C obviously depends on the essence of P, i.e., on what P non-derivatively is (or could be) in all possible circumstances.

We can now apply this view to weakly emergent features and their causal roles.

Roughly, there are three facts to be accounted for: that the proper subset that only includes powers p_1, p_2, p_3 is the causal role of *a* token feature; that it is the causal role *of* token feature S and *only* of token feature S (at least as a matter of nomological necessity); that S *cannot* have any other causal role (at least as a matter of nomological necessity).

The first two facts are easily accounted for by my doctrine. The causal role of a token feature S depends on the property involved in that token feature, i.e., the weakly emergent property in S. It is (also or only) by virtue of the property involved in S that causal powers p_1, p_2 and p_3 are put together so as to constitute the causal role of *a* token feature, so that the relevant causal role starts and continues to exist.

Secondly, it is by virtue of that property that such powers constitute the causal role *of* token feature S, and only of it (or only of token features of that property). And this seems to be part of the essence of the causal role of S[5]. Yet,

[4] I offer a proof of this latter thesis in Paolini Paoletti 2022.

[5] The connection between the weakly emergent property involved in S and the causal role C does not merely hold as a matter of nomological necessity. For there is no possible world with other laws of nature in which C is associated with a property distinct from the one involved in S. C, by virtue of its own essence, is only associated with the property involved in S. This seems reasonable in light of the physicalist commitments of weakly emergentists. Indeed, if C were associated with the property involved in S in one possible circumstance and with some other property in another possible circumstance, then there would be nothing at the level of C (nor at the level of the causal powers included in C) to account for this difference.

my view does *not* entail that powers p_1, p_2 and p_3 turn out to be non-physical. Indeed, such powers may well be physical powers, so that they do *not* depend for what they are on token feature S, nor on the weakly emergent property involved in S. It is only the relevant causal role made of powers p_1, p_2 and p_3 that depends on the weakly emergent property involved in S.

In Paolini Paoletti 2022, I have also defended the following thesis: the categorical property P can have other causal roles different from C in other possible worlds and/or at other times. When applied to weakly emergent properties/token features and the causal roles associated with them, this is at odds with the third fact to be accounted for: that the token feature S (and, presumably, the weakly emergent property involved in it) cannot have any other causal role (at least as a matter of nomological necessity).

If we wish to stick to this fact, we can argue that, as a matter of metaphysical necessity, the weakly emergent property involved in S is realized by causal role C and only by C, so that it cannot have any other causal role. Namely, the weakly emergent property involved in S necessarily depends for its being causally effective on (i.e., is realized by) causal role C and only on it. I assume that dependence for causal effectiveness (i.e., realization) and the other relations of dependence mentioned above are distinct and non-equivalent. I shall expand on this point in a few lines.

Something similar to the solution I suggest here is explored by Wilson (2021: 96-97) in reply to Melnyk (2006). Wilson objects to this solution that scientific truths about scientific features do *not* depend on the presence or on the absence of quiddities (i.e., of qualitative aspects of properties). Moreover, she claims that quiddities are mostly required for transworld individuation, whereas the individuation of properties in worlds that share our laws of nature only proceeds by reference to powers.

What I suggest here is that we *do* need quiddities for metaphysical reasons, i.e., in order to answer questions (1) and (2). Or, at least, we need to appeal to (the essence of) higher-level properties, not fully exhausted by their causal powers. Additionally, not all the facts mentioned in such questions as *explananda* are 'other-wordly' facts. For example, that the proper subset with p_1, p_2 and p_3 is associated with a higher-level token feature is not an 'other-wordly' fact.

In a similar vein and in the footsteps of other authors[6] Inman (2018) suggests that the irreducible essences of higher-level substantial kinds play two roles. First, they structure the modal profiles associated with such kinds, i.e., they connect all the possible ways the relevant substances can be characterized and modified. Secondly, the irreducible essences of higher-level substantial kinds fix the causal profiles associated with such kinds, i.e., all the causal powers the relevant substances possess by necessity whenever they exist.

By embracing my solution, we avoid introducing primitive and *sui generis* connections between token features and proper subsets of powers. However, two problems are left open.

The first problem is that this solution is incompatible with some versions of physicalism. If the causal role of token feature S depends on the higher-level and weakly emergent property involved in S, then it is *not* the case that everything depends on the physical. Secondly, assume that token feature P is physical. P does

[6] Inman (2018: 49) cites Scaltsas (1994: 78-80), Des Chene (1996: 71-75), the Early Modern metaphysician Francisco Suárez (2000), Lowe (2006: 135) and Oderberg (2011).

not depend on the property involved in S. Nor do its physical causal powers depend on that property. However, on the one hand, it seems that the causal role of S depends on the property involved in S. Yet, on the other hand, it seems that the property in S depends—for its being causally effective—on that very causal role. There seems to be a circle of dependence here.

To solve these problems, I suggest that we should first swallow the fact that weak emergence is not so weak. Weak emergence is incompatible with the idea that everything whatsoever is physical or fully depends on the physical.

Moreover, I also suggest that different dependence relations may actually be at stake with the property involved in S and the causal role of S. Indeed, the causal role of S may depend *in a certain respect* (e.g., for its being the causal role of S and for its starting and continuing to exist) on the property involved in S. Yet, the property involved in S may depend *in another respect* (e.g., for its being causally effective, or 'realized') on the causal role of S. Such respects are associated with distinct and non-equivalent dependence relations that may run in opposite directions and still remain by themselves asymmetrical.[7]

By invoking distinct dependence relations, we can then construct distinct and non-equivalent versions of physicalism. We can also generalize in order to make sense of the idea that the physical is more fundamental than the non-physical. Intuitively, we can take into account all the dependence relations that involve physical entities and all those that involve non-physical entities. We can then determine the overall degree of dependence of the former and the overall degree of dependence of the latter. Finally, we can find out that the overall degree of dependence of physical entities is lower than that of non-physical entities, so that the former are more fundamental than the latter.

In sum, there are two lessons to be learnt here. The first lesson is that weak emergence should be accepted in conjunction with metaontological pluralism, i.e., the view that distinct and non-equivalent dependence relations are at stake in the universe. The second lesson is that weak emergence is *not always* compatible with physicalism, i.e., it is not compatible with all forms of physicalism.

It may be objected that my approach is no better than primitivism. Indeed, even primitivism is somehow incompatible with physicalism. And even primitivism turns out to take higher-level, weakly emergent properties as fundamental. However, unlike primitivism, my approach does *not* take the *explanandum* (i.e., the connection between S and its causal role) as a primitive fact of the matter. On the contrary, it explains this connection by appealing to the weakly emergent property involved in S. And my approach postulates no special entity such as the relation R. On the contrary, only the weakly emergent property involved in S and the relevant causal role are taken into account.[8] In turn, the weakly emergent property involved in S is something we are already committed to if we believe that S is a token *feature*. And it need *not* be a universal property. Therefore, *ceteris paribus*, my approach is also ontologically more parsimonious than primitivism.[9]

[7] More on this in Paolini Paoletti 2019 and 2021.

[8] The dependence relations at stake in my approach turn out to be internal relations, i.e., relations whose presence is determined just by the essence and/or the existence of their own relata. On the contrary, the relation R postulated by primitivism is *not* internal. For the weakly emergent property involved in S and its causal role are *not* enough (through their essence and/or existence) to make it the case that R holds between them.

[9] I wish to thank Jessica Wilson and the audience at the Sixth Italian Conference on Analytic Metaphysics and Ontology (L'Aquila 2022).

References

Des Chene, D. 1996, *Physiologia: Natural Philosophy in Late Aristotelian and Cartesian Thought,* Ithaca: Cornell University Press.

Elder, C.L. 2004, *Real Natures and Familiar Objects,* Cambridge, MA: MIT Press.

Elder, C.L. 2011, *Familiar Objects and their Shadows,* Cambridge: Cambridge University Press.

Kimpton-Nye, S. 2021, "Reconsidering the Dispositional Essentialist Canon", *Philosophical Studies*, 178, 3421-41.

Lowe, E.J. 2006, *The Four-Category Ontology: A Metaphysical Foundation for Natural Science,* Oxford: Oxford University Press.

Melnyk, A. 2006, "Realization and the Formulation of Physicalism", *Philosophical Studies*, 131, 127-55.

Oderberg, D. 2011, "Essence and Properties", *Erkenntnis*, 75, 85-111.

Paolini Paoletti, M. 2019, "Respects of Dependence", *Studia Neoaristotelica*, 16, 49-82.

Paolini Paoletti, M. 2021, "Respects of Dependence and Symmetry", *Studia Neoaristotelica*, 18, 31-68.

Paolini Paoletti, M. 2022, "A Brighter Shade of Categoricalism", *Axiomathes*, 32, 1213-42.

Scaltsas, T. 1994, *Substances and Universals in Aristotle's Metaphysics,* Ithaca: Cornell University Press.

Sider, T. 2011, *Writing the Book of the World,* Oxford: Oxford University Press.

Suárez, F. 2000, *On the Formal Cause of Substance: Metaphysical Disputation XV,* translated by J. Kronen and J. Reedy, Milwaukee: Marquette University Press.

Tugby, M. 2012, "Rescuing Dispositionalism from the Ultimate Problem: Reply to Barker and Smart", *Analysis*, 72, 723-31.

Tugby, M. 2020, "Grounding Theories of Powers", *Synthese*, 198, 11187-216.

Tugby, M. 2022a, "Dispositional Realism without Dispositional Essences", *Synthese*, 200, article 222.

Tugby, M. 2022b, *Putting Properties First: A Platonic Metaphysics for Natural Modality*, Oxford: Oxford University Press.

Wilson, J. 2010, "Non-reductive Physicalism and Degrees of Freedom", *British Journal for the Philosophy of Science*, 61, 279-311.

Wilson, J. 2021, *Metaphysical Emergence,* Oxford: Oxford University Press.

Yates, D. 2018, "Inverse Functionalism and the Individuation of Powers", *Synthese*, 195, 4525-50.

Author Meets Critics Session on *Metaphysical Emergence*: Replies

Jessica Wilson

University of Toronto

Introduction

I'd like to start by thanking Simone Gozzano, the patient editor and shepherd of this volume, Massimo Dell'Utri, the benevolent editor-in-chief of Argumenta, and Michele Paolini Paoletti, who initially suggested that an issue of Argumenta be devoted to *Metaphysical Emergence*. Simone, special thanks for your encouragement and your efforts; this is a great honour for me, and you have been a fantastic (and patient) collaborator, in print and in song. I'd also like to sincerely thank my commentators for their illuminating, fruitful, and provocative discussions of my book. The diversity of topics they have addressed, highlighting connections between metaphysical emergence and areas ranging from ontology to property theory to counterfactuals to mereology to quantum field theory to biochemistry and beyond, is truly striking, and a real testament to the wide-ranging import and applications of the notion of metaphysical emergence. Every contribution has given me substantive food for thought. For reasons of space I have focused my replies to each commentator on what I see as the most pressing of their remarks, but of course there is more to say, and I hope and anticipate that these conversations will continue on beyond this volume.

1. Replies to Bellazzi

Bellazzi offers a novel application of Weak emergence as the operative relation between the (broadly biological) function and (broadly chemical) structure of biochemical molecules, such as vitamin B12. As Bellazzi notes, biochemistry stands as a kind of 'hybrid domain' between chemistry and biology, with biochemical kinds understood as having micro-structural features of the sort characteristic of chemical kinds, and certain functions of the sort operative in biological systems. Given that the characterization of a biochemical kind incorporates both structural and functional features, the question arises of how these features stand to one another, as per what Bellazzi calls 'the relation problem'—a problem, and not just a question, reflecting a certain trickiness in identifying a relation capable of

Argumenta 10, 1 (2024): 313—364
ISSN 2465-2334

DOI 10.14275/2465-2334/202419.wis
First published: 30 November 2024

accommodating certain constraints on the connection at issue. These constraints reflect that biochemical kinds are typically both multiply realizable (MR)—such that the same biochemical function can be realised by multiple microstructures—and multiply determinable (MD)—such that the same biochemical structure can realise multiple biochemical functions (see Slater 2009; Bartol 2016; Tahko 2020). These joint features of, or constraints on, the relation at issue are in place for Bellazzi's case study of vitamin B12, whose biochemical functions can be realised by any of four distinct vitamers, and whose biochemical structure(s) can play different roles in human physiology, including in DNA and RNA production, and in hematopoiesis/erythropoiesis.

Bellazzi convincingly argues, to my mind, that taking the relation between biochemical structures and functions to be one of Weak emergence provides an illuminating basis for accommodating MR and MD in the case of vitamin B12, and more generally in other cases of biochemical kinds. I will not repeat the details of her application here, but will rather highlight and discuss what I think are three important ramifications of Bellazzi's discussion for investigations in intra-level metaphysics. I close with some related questions about the specific application at issue.

The first moral of Bellazzi's application is that cases of emergence need not be associated with different 'levels.' Discussions of emergence tend to take for granted that this relation holds between goings-on (in the usual case: features) in different sciences. Hence in my book I focus on cases, e.g., where certain features of ordinary objects of the sort treated by Newtonian mechanics might emerge from features of quantum mechanical aggregates; or where certain thermodynamic properties of complex systems might emerge from properties of statistical mechanical aggregates; or where certain conscious mental states might emerge from neurological and ultimately lower-level physical states; and so on. In the case of biochemical kinds, however, and notwithstanding the connection to chemical and biological kinds and features, what appears to be at issue is the relation between seemingly distinct features of a kind treated by a single special science. The possibility of such intra-level emergence complexifies the structure of special scientific goings-on, both expanding the range of cases which might potentially involve metaphysical emergence, and also suggesting that we should be cautious about assuming that any case of metaphysical emergence is one generating a new 'level' of natural reality.[1] That said, the case of biochemical kinds and features also raises the questions of what relations (most saliently: identity or emergence?) hold between, first, the individual structural and functional components of biochemical kinds, and second, the features in the proximal sciences—i.e., between the structure of a biochemical kind and chemical structure, and the function of a biochemical kind and biological function. I'll return to this issue down the line.

A second moral of Bellazzi's application is that MD is an underappreciated resource so far as theorizing about inter-level metaphysics, and emergence in particular, is concerned. Discussions of emergence often advert to cases of multiple realizability (MR) of a given feature as providing some reason to think that the

[1] A similar moral might be seen as read off of diachronic or 'transformational' conceptions of emergence (see, e.g., Humphreys 1997 and Guay and Sartenaer 2016) as involving fusion or some other interaction at a single level. Bellazzi's moral rather applies to cotemporal emergence of the sort traditionally associated with leveled structure.

feature cannot be treated in reductive (identity-based) terms, and is rather better treated as metaphysically emergent, one way or another. Hence in my book the potential bearing of multiple realizability on a given claim of metaphysical emergence (typically, of the Weak variety) comes up several times. As it happens, a theme of my discussions on this topic is that a feature's being multiply realizable isn't in itself sufficient to establish that the feature is Weakly emergent, at least antecedent to engaging with certain reductionist strategies for accommodating multiple realizability in identity-based terms—most commonly, by taking the lower-level feature to which the higher-level feature is supposed to be identical to be a disjunction of *S*'s realizers; and I also argue that a feature's being multiply realizable isn't necessary for its being Weakly emergent. That said, it remains that the multiple realizability of a higher-level feature is the feature most commonly offered as indicative of a feature's being Weakly emergent. Now, as above biochemical kinds are MR, in that the same biochemical function can be realised by multiple microstructures; but they are also MD, in that a single biochemical structure may realize, or determine, multiple biochemical functions.

To see that MD is an underappreciated resource in theorizing about interlevel metaphysics, note that, notwithstanding that MR poses a prima facie difficulty for reductionism, there is in such cases at least an available candidate lower-level feature (namely, the feature consisting in the disjunction of the multiple lower-level realizers) for the reductionist to appeal to in conformity with their claim that every higher-level feature is in fact identical to some or other lower-level feature. But in cases of MD, it is less clear how an identity-based strategy is supposed to be implemented. Suppose that a single lower-level feature F is capable of determining multiple higher-level features (functional or otherwise) S_1, S_2, and S_3. Each determined feature is, according to the reductionist, identical to some or other lower-level feature, but which one? F can't be identical to just S_1, since in that case F's determination of S_2 and S_3 is unaccounted for. An alternative strategy would be to identify S_1 with some part or aspect of F, and similarly for S_2 and S_3; but even granting that such parts or aspects are available for the identification, as it stands it is unclear that these parts or aspects are properly seen as themselves being lower-level features, as the reductionist requires. Indeed, on some accounts of realization (per, e.g., Shoemaker 2000/2001 and Clapp 2001), token realized features are taken to be proper parts of their realizers. From this perspective, multiple determination poses even more of a challenge to reductionism than multiple realization.

A third moral of Bellazzi's application is that it encodes a distinctive response to the question of which subsets of powers of a given dependence base feature are, or can be, associated with a Weakly emergent feature. In my book, I largely leave it to the scientists to discover which entities and features, and associated powers, are plausibly hypothesized as making sense of natural reality, taking my goal to be that of saying how, given that such-and-such entities and features are supposed to have the key features of metaphysical emergence (as coupling dependence with ontological and causal autonomy), we can make sense of this supposition. I do offer one more specific answer to this question, in the context of discussing an implementation of Weak emergence involving an elimination in degrees of freedom; here the idea is that which degrees of freedom (and associated powers) are eliminated from the characterization of the higher-level feature will often reflect the holding of certain lower-level constraints. But attention to Weak emergence in biochemical kinds provides the basis for a new specific answer to the question

of which subsets of powers are associated with genuine features—namely, that this may be, as Bellazzi puts it, "a product of evolution", and more specifically (as per her forthcoming) that biochemical functions are "associated with a set of chemical powers to bring out a specific effect within biological processes" where these processes are a product of evolution, such that "the relevant chemical powers are indirectly evolutionary selected" (see also Santos et al. 2020). This 'evolutionary' route to identifying which subsets of powers of a given feature are associated with genuine, and moreover Weakly emergent features, is an important part of the background story about why natural reality has the structure it has, which promises to illuminate and apply to kinds and features in biological, ecological, and many other sciences. It also serves to show that there are apparently at least two quite distinct sources capable of generating Weakly emergent features: one broadly synchronic (as in the cotemporal imposition of constraints), and one broadly diachronic. As such, it is unclear whether we should expect a unified metaphysical explanation of which higher-level features come to exist, and why—an important result in its own right.

I want to turn now to raising some questions about Bellazzi's application, falling under the rubric of a single question—namely, how many (potentially instantiated) relations of Weak emergence might be associated with a given biochemical kind?

Let's assume that Bellazzi is right that biochemical functions Weakly emerge from biochemical structures. As above, in being MD, a given biochemical kind may have multiple biochemical functions, each of which would presumably be Weakly emergent from whatever biochemical structure is associated with the kind on a given occasion. So a biochemical kind is plausibly associated with as many Weak emergence relations as the kind has biochemical functions. But now recall that, in being MR, a given biochemical kind may have multiple biochemical structures.[2] And for each such biochemical structure, the question arises of whether it is identical to, or rather (presumably, Weakly) emergent from, a chemical structure. Perhaps each biochemical structure is just identical to some chemical structure, as is suggested by the characterization of biochemistry as "the science that considers the behaviour and effects of chemical processes in biological systems" (Bellazzi, this volume, per Santos et al. 2020). But perhaps there are cases to be made that some or all biochemical structures have only a proper subset of the token powers of associated chemical structures. In that case, a biochemical kind would be associated with as many Weak emergence relations as the kind has distinct realizers. Finally, just as there is a question of what relation holds between chemical and biological structures, there is a question of what relation holds between biochemical and biological functions. Might the latter relation(s) also be ones of Weak emergence? If so, a biochemical kind would be associated with as many Weak emergence relations as the kind has biochemical functions—now running not (as in Bellazzi's case) from biochemical structure to biochemical function, but rather running from biochemical function to biological function.

I offer these questions as further food for theorizing for Bellazzi and others working on the metaphysics of biochemistry. In any case, I'm well convinced that attention to the distinctive characteristics of biochemical kinds points the way towards several new avenues of investigation in the metaphysics of emergence.

[2] I assume that each such structure can serve as a dependence base for any (i.e., all) of the biochemical kind's biochemical functions.

2. Replies to Bennett

In my book, I motivate the powers-based schemas for Weak and Strong metaphysical emergence by attention to the problem of mental/higher-level causation, pressed by Kim (1989 and elsewhere); my basic line is that the two schemas encode the strategies operative in the only responses to Kim which accommodate metaphysical emergence, understood as coupling cotemporal material dependence and (ontological and causal) autonomy. I motivate the schema for Weak emergence, more specifically, by attention to non-reductive physicalist (NRPist) responses to Kim's problematic, which posit diverse relations (functional realization, compositional mechanism, the determinable-determinate relation, and so on) advanced as making sense of how cotemporally dependent higher-level features may be distinct and distinctively efficacious as compared to their physical base features, in a way not involving causal overdetermination of the 'double-rock-throw' variety that makes little sense for the cases at issue. I argue that "a deeper unity of strategy" underlies the seemingly diverse NRPist accounts—namely, that the posited relations[3] each guarantee that, on any given occasion, the higher-level feature has only a proper subset of the token powers of the physical feature upon which it cotemporally materially depends; and I argue that the holding of the Proper Subset of Powers condition, along with the cotemporal material dependence condition, captures what is core and crucial to metaphysical emergence of a physically acceptable variety.

In her contribution, Bennett offers three challenges to this motivation for my account of Weak emergence. The first is that there is an alternative NRPist response to Kim's problematic—Bennett's 'Counterfactual Strategy'—which also encodes "a deeper unity of strategy", but which does not involve any reference to the Proper Subset of Powers Condition. The second is that the Proper Subset Strategy itself does not establish the efficacy of the mental (or Weak emergents more generally). And the third is that the means by which Weak emergent efficacy avoids overdetermination is not as ontologically neutral as I have made it out to be. These challenges are well worth considering; in what follows, I present and respond to each in turn.

2.1 Challenge 1: The Counterfactual Strategy

As noted, I see the deeper unity of strategy underlying diverse NRPist accounts posits as reflecting that their chosen relations guarantee satisfaction of the Proper Subset of Powers condition at the heart of my schema for Weak emergence; but drawing on her 2003 and 2008, Bennett suggests that the underlying unity reflects that the relations posited by NRPists allow implementation of what she calls the 'Counterfactual Strategy' in response to Kim's concerns about overdetermination:

> Talk of overlapping sets of causal powers is not the only way to explain how various intimate relations between the causes defuse the threat of overdetermination. In a (2003) paper, I offered a different explanation. I provided a necessary condition on overdetermination (genuine, 'double-rock' overdetermination), and argued that it is

[3] Not including supervenience or other mere modal correlations, which for various reasons are too weak for physicalist purposes; see Wilson 2005 and McLaughlin and Bennett 2018.

> not met by pairs of causes related in any of the ways [Weak emergentists/NRPists] think that mental and physical phenomena are.
>
> The necessary condition is simply that two causes overdetermine an effect only if had either happened without the other, the effect would still have occurred. That is, causes c_1 and c_2 overdetermine e only if both of the following counterfactuals are nonvacuously true:
>
> $(c_1 \wedge \neg c_2) \rightarrow e$
>
> $(c_2 \wedge \neg c_1) \rightarrow e$
>
> This is a very intuitive test for overdetermination. […] if the test is legitimate, the [Weak emergentist/NRPist] is again in good shape. At least one of these counterfactuals will be vacuous or false when (2003) and only when (2008) the mental and physical causes stand in one of the […] favored relations. […] the basic idea is that on any such relation, the physical base necessitates the weakly emergent mental phenomena, rendering one of the counterfactuals vacuous. (241)

As Bennett's past work makes clear, the necessitation at issue here is metaphysical, such that in every possible world where the physical base feature is instanced, so will be the higher-level mental feature. As such, if c_1 is a mental feature M, and c_2 the mental feature's physical base P, then the counterfactual '$(P \wedge \neg M) \rightarrow e$' will be vacuously true, and the necessary condition for overdetermination will fail to be met.

Bennett offers the Counterfactual Strategy as a kind of 'minimalist' response to Kim's problematic, in the sense that it provides a basis for denying one of the premises in Kim's argument—namely, on Bennett's reconstruction, the premise ('Exclusion') according to which all events that have multiple sufficient causes (that are not themselves causally related) are overdetermined. The Counterfactual Strategy is minimalist in being silent on further details about how, exactly, a higher-level feature might be efficacious in such a way as to avoid overdetermination. That said, as above Bennett does suppose that the Counterfactual Strategy unifies NRPist approaches, and relatedly (as is developed in her 2008) is not available to dualists, including Strong emergentists. In what follows I'll offer three reasons for thinking that the Counterfactual Strategy is subject to problems rendering it unsuccessful even with respect to these minimalist aims. As I'll also observe, the Proper Subset Strategy does not incur these problems, and so is correspondingly advantageous.

2.1.1 Response 1: The Illegitimacy of the Test

Is Bennett's test 'legitimate,' in being a necessary condition on overdetermination, such that failure of one or other counterfactual to be non-vacuously true will get one off the overdetermination hook? No, for it is easy to construct cases of clear overdetermination, where the overdetermining phenomena are nonetheless sensitive to whether the other occurs. Indeed, the whole point of firing squads is to ensure that everyone pulls the trigger, so that no individual is to blame. We can similarly set things up so that Billy and Suzy make a pact that they will each throw the ball at the window only if the other does, so that in the closest worlds where either doesn't throw, neither does the other.

It is an advantage of the Proper Subset Strategy that, unlike the Counterfactual Strategy, it doesn't rely on a condition on overdetermination that is subject to clear counterexample.

2.1.2 Response 2: The Controversy and Context-sensitivity of Counterfactual Assessment

Counterfactual deliberation and assessment are subject to controversy and context-sensitivity. The controversy at issue pertains not so much to the general account of counterfactual truth—most accept some kind of similarity-based account, where a counterfactual is true just in case in the closest world(s) where the antecedent is true, the consequent is true—but rather to the question of how worlds are to be ordered with respect to similarity, given that (as Fine, 1975, nicely established) overall similarity won't do. At present there is no agreement either on more specific criteria of similarity or their ranking. Relatedly, similarity judgements are highly context-sensitive. Bennett briefly registers this in discussing a move according to which (relative to some contexts) events are highly fragile—so fragile that in cases of overdetermination, it turns out to be false that had one but not the other event occurred, then the (same type of) effect would still have been produced.[4] But the more general point is that, given the context-dependence of similarity, whether the counterfactual conditions on overdetermination are or are not met is going to depend on context. Relative to one context, perhaps, there's no overdetermination; relative to another, there is. In that case, Bennett's condition does not provide a clear basis for a response to Kim, but rather pushes the bump in the rug to the question of which contexts are most crucial so far as questions of overdetermination are concerned.

It is an advantage of the Proper Subset Strategy that, unlike the Counterfactual Strategy, it (and the associated response to Kim) isn't subject to the controversy and context-dependence of counterfactuals.

2.1.3 Response 3: failing to distinguish Weak and Strong emergentist responses to Kim

As above, Bennett intends that the Counterfactual Strategy unify Weak emergentist/NRPist responses to Kim's problematic, and distinguish these from anti-physicalist dualist, including Strong emergentist, responses. But as I'll now argue, the Weak and Strong emergentist can implement the Counterfactual Strategy in exactly the same way. Bennett can distinguish these responses, but at the price of taking on board certain controversial metaphysical commitments—commitments not needed to implement the Proper Subset Strategy.

To start, consider the overdetermination counterfactuals for a mental feature M that is supposed to be Weakly emergent. The counterfactual '$(M \wedge \neg P) \rightarrow e$' will likely be non-vacuously true, given the usual assumption that mental states may have diverse physical bases (in a physicalist context: are 'multiply realizable'); for then the nearest antecedent worlds will likely be ones where M has a slightly different physical base (realizer), and M causes e. However (per Bennett's characterization of the NRPist's response to Kim), '$(P \wedge \neg M) \rightarrow e$' will be only vacuously true, given that P metaphysically necessitates M.

[4] Note that this amounts to another 'Counterfactual Strategy' that the NRPist could avail themselves of in response to Kim. Bennett suggests that those endorsing fragile events take the effect to be jointly caused by higher-level and base features, but that diagnosis of the effect's fragility is optional—the fragile event NRPist can just adopt Bennett's minimalist stance and resist calls to provide details about how, exactly, higher-level features enter into causing effects.

Now consider the overdetermination conditionals for a mental feature M that is supposed to be Strongly emergent. The counterfactual '$(M \wedge \neg P) \rightarrow e$' will likely be non-vacuously true, given the usual assumption that mental states may have diverse physical bases (in anti-physicalist context: are 'multiply determined'); for then the nearest antecedent worlds will likely be ones where M has a slightly different physical base, and M causes e. What about '$(P \wedge \neg M) \rightarrow e$?' In her 2008, Bennett argues that the NRPist treatment of this counterfactual "is not available to the dualist": "the dualist cannot say that [this counterfactual] is either false or vacuous […] For the dualist, cases of mental causation do meet the necessary condition on overdetermination". Most relevant here is Bennett's reason for thinking that the dualist (Strong emergentist) cannot claim that the relevant counterfactual is vacuous:

> It is clear that only the physicalist can say that ['$(M \wedge \neg P) \rightarrow e$'] ever comes out vacuous. The dualist cannot, because she does not think that there are any physical events or properties that metaphysically necessitate mental ones. She precisely thinks that there are—at best!—contingent psychophysical laws that link the two. So the dualist denies that there is any legitimate substitute for [P] that would make the antecedent metaphysically impossible. She at most thinks that there are choices of [P] that would make the antecedent nomologically impossible. So the dualist cannot claim that any instance of [the counterfactual] is vacuous. (2008: 290)

This line of thought builds in a controversial metaphysical commitment, however—namely, that Strong emergents are nomologically but not metaphysically necessitated by their physical bases. As I discuss in my (2005), however, there are several views on which Strong emergents are metaphysically necessitated by their physical bases, including a modally consistent Malebranchean occasionalism, a view of properties as essentially constituted by all of the laws into which they enter, and a view of fundamental interactions as holistically unified. Moreover, I argue, the latter two views enjoy considerable empirical support, by contrast with Humean 'anything goes' versions of contingentism which greatly depart from scientific theorizing and practice. Whether or not one accepts any of these views, the fact remains that Bennett's Counterfactual strategy does not itself distinguish between the Weak and Strong emergentist strategies, independent of further controversial assumptions about the modal strength of the connections at issue.

Indeed, upon closer examination even the supposition that the NRPist's favoured relations are such that a physical base metaphysically necessitates a Weak emergent can be denied. Consider functional realization, according to which, e.g., mental feature M is associated with a distinctive causal or functional role, which on a given occasion is played by some lower-level physical feature P. Need P metaphysically necessitate M? Not on causal contingentist views, on which properties and powers may come apart; for on such views there is no guarantee that P, instanced in worlds with different laws of nature, will have the powers requisite unto playing M's causal role. For such a contingentist functionalist NRPist, it might well be that both of the counterfactuals in the Counterfactual Strategy turn out to be non-vacuously true. Correspondingly, the success of the Counterfactual Strategy requires a further metaphysical commitment—namely, the rejection of causal contingentism.

By way of contrast, the Proper Subset Strategy clearly distinguishes between the Weak and Strong emergentist responses to Kim, in a way that is moreover

neutral both on whether either relation holds with metaphysical necessity (requiring only, as per the cotemporal material dependence condition which NRPists and Strong emergentists agree is in place, that emergents supervene with at least nomological necessity on physical goings-on) and on whether causal contingentism is true.

2.2 Challenge 2: No Explanatory Advantage

Putting aside the previous concerns and granting that Bennett's Counterfactual Strategy suffices to undercut the Exclusion premise of Kim's argument, one might wonder whether the Proper Subset Strategy is more explanatory than the Counterfactual Strategy, in going beyond a minimalist response to establish that mental goings-on, in particular, are efficacious. Bennett registers, however, that she is skeptical of this:

> [Wilson's strategy] could solve the exclusion problem and secure the causal efficacy of the mental. But I am still skeptical; I do not think the strategy actually does secure that. All the work is done by Wilson's claim that weakly emergent entities have a nonempty proper subset of the causal powers of their bases. This is the only reason we are guaranteed that weakly emergent entities have causal powers. But Wilson never argues that any particular thing or kind of thing has a non-empty set of causal powers; that is just part of her definition of weak emergence. So those who are inclined to be worried about the causal efficacy of the kinds of phenomena she takes to be weakly emergent—like the mental—will simply deny that they are weakly emergent in her sense. (244)

I agree with Bennett that the Proper Subset Strategy qua response to Kim doesn't itself establish that the mental or any other phenomena is efficacious. The Strategy qua response is at that point in-principle, specifying what it would take for some phenomenon to be Weakly emergent in a way in line with NRPist intentions and accounts. Similarly for the New Power Strategy at the heart of the schema for Strong emergence.

Arguments that mental or other phenomena actually have "a non-empty set of causal powers" come later. Hence after arguing for the in-principle viability of (my conception of) Weak emergence (Ch. 3), I argue that there are good cases to be made that complex systems (Ch. 5), ordinary objects (Ch. 6), qualitative mental states (Ch. 7), and (compatibilist) events of free choosing (Ch. 8) satisfy the conditions in the schema for Weak emergence. I motivate the satisfaction of the Proper Subset of Powers condition by attention to a variety of (empirical, philosophical, introspective, etc.) considerations. In brief (see the chapters for details): for complex systems, satisfaction of the condition mainly hinges on the applicability of the renormalization group method and associated elimination of microphysical degrees of freedom (DOF), coupled with my DOF-based account of Weak emergence; for ordinary objects, satisfaction hinges, alternatively, on the elimination of quantum DOF, on sortal practices of individuation, and on ordinary objects' having metaphysically indeterminate boundaries, understood as per my determinable-based account of metaphysical indeterminacy and coupled with a determinable-based account of Weak emergence; for conscious (qualitative) mental states, satisfaction mainly hinges on perceptions' being determinable, coupled with a determinable-based account of Weak emergence; and for (non-libertarian) free will, satisfaction hinges on an understanding of 'relevant antecedent'

approaches to compatibilist free will according to which the powers of the complex event comprising the relevant antecedents are a proper subset of those associated with the complex event comprising the complete antecedents.

Does the fact that qua response to Kim, the Proper Subset Strategy doesn't itself establish that the mental is actually efficacious mean that the Strategy doesn't have any explanatory advantage over the Counterfactual Strategy? I'm inclined to deny this, for two reasons. First, unlike the Counterfactual Strategy, the Proper Subset Strategy provides an explanatory basis for not just the efficacy, but moreover the distinctive efficacy, of Weak emergents—a distinctive efficacy which tracks difference-making considerations (if my thirst had been differently realized, I would still have reached for the Fresca) associated with comparatively abstract systems of laws or levels of causal grain. Independently of further investigations into which phenomena are actually Weakly emergent, this conception of distinctive efficacy provides the basis for a more compelling NRPist response to Kim than does the Counterfactual Strategy; for it undercuts Kim's incorrect supposition that the distinctive efficacy of a higher-level feature can only lie in the having of a novel power, contra Physical Causal Closure, hence contra Physicalism. Second, unlike the Counterfactual Strategy, the Proper Subset Strategy provides a blueprint for establishing that a given phenomenon is Weakly emergent, and so is not just efficacious but distinctively so—a blueprint that is, as I argue, often realized.

2.3 Challenge 3: Undue Ontological Commitment

Bennett's third challenge is that on the face of it, implementing the Proper Subset Strategy for avoiding overdetermination requires "ontological commitment to trackable, countable causal powers".

> [T]he success of the Proper Subset Strategy entirely depends on the idea that the causal powers of the emergent phenomena are numerically identical to the causal powers of the base. And this in turn requires that token causal powers are the sort of thing that can not only be counted but also individuated. Indeed, it is very, very hard not to imagine them as pebbles in a bucket—and Wilson's diagrams on page 70 suggest that she cannot resist this picture either. But this is a serious and rather discombobulating ontological commitment. I will not argue here that causal powers are not like that, but I suspect others will share my reticence. Even Wilson takes pains to insist that her causal powers are nothing dubious or creepy:
>
> > Talk of powers is simply shorthand for talk of what causal contributions possession of a given feature makes […] to an entity's bringing about an effect, when in certain circumstances… no controversial theses pertaining to the nature of powers, causation, properties, or laws are here presupposed. (32–33; also 45)
>
> But the question is, can she really make good on this neutrality? More precisely, can she assuage my ontological qualms while retaining the nice claim that strictly speaking, there is really only one cause of an effect caused both by a weakly emergent phenomenon and its base? That is the challenge I lay before her. […] My real point here is that one cannot have the Proper Subset Strategy on the cheap; the cost-benefit analysis must be made. We can shoulder the ontological commitment to trackable, countable causal powers and accept the benefits, or we can be squeamish and reject the whole picture. (245–44)

I think this is a fair question, but by way of convincing the skeptic I'm not sure what to say beyond what I've already said. As above, and notwithstanding the convenient schematic representation of powers as 'pebbles in a bucket,' I am explicit about the operative notion of 'power' as simply tracking what (actual or potential) causal contributions the having of a given feature makes when instanced in certain circumstances. As I note by way of proof of metaphysical neutrality, even a contingentist categoricalist Humean can make sense of powers in this sense:

> [E]ven a contingentist categoricalist Humean—someone who thinks that causation is a matter of regularities, such that features have their powers contingently, and that all features are ultimately categorical—can accept powers and the associated notion of causation in the neutral sense(s) here: for such a Humean, to say that an (ultimately categorical) feature has a certain power would be to say that, were a token of the feature to occur in certain circumstances, a certain (contingent) regularity would be instanced. Contemporary Humeans implement more sophisticated variations on this theme; but the point remains that no 'heavyweight' notion of powers or causation need be presupposed in what follows. (33)

So far, so metaphysically neutral. But is it really the case that, as I claim in discussing the schemas, "effectively all participants to the debate can make sense of such identity (non-identity) claims as applied to token (actual or potential) causal contributions (token powers)" (45)? If one has a conception of dispositions or powers as ontological existents, then presumably there is no in-principle difficulty with making sense of these being token identical, in any given case. But as I note by way of proof of metaphysical neutrality, even a contingentist categoricalist Humean can make sense of such identification:

> For example, suppose a contingentist categoricalist Humean wants to take a physicalist approach to the problem of higher-level causation, and so aims (as I will expand on §2.3) to identify every token power of a token higher-level feature with a token power of its lower-level base feature. As previously discussed, such a Humean understands powers in terms of actual or potential instances of a (contingent) regularity. Where the aim is to avoid overdetermination, the Humean may suppose, to start, that the (relevant instances of the) regularities overlap, both with respect to the (single) effect, and with respect to the (single) circumstances in which the two token features occur. If the Humean aims to be a reductive physicalist, they may suppose that such overlap motivates identifying the token features at issue, and hence the associated powers. If the Humean aims to be a nonreductive physicalist, they can reject this identification of features, on difference-making or other grounds of the sort to be discussed §2.3. Such a Humean will suppose that attention to broader patterns of regularities can provide a basis for identifying token powers of token features, even when the token features are not themselves identical. Whether reductive or nonreductive, the contingentist categoricalist Humean can make sense of the claim that some, all, or none of the token powers of token features are identical. As I observed in my (2015: 35), this case is like the case of New York: if we can make it (out) here, we can make it (out) anywhere. (45–6, note 15).

That said, it is worth clarifying that it isn't any part of my view that "there is really only one cause of an effect caused both by a weakly emergent phenomenon and its base"—i.e., the causal power that the mental feature shares with its physical

base. If that were part of my view, I can see why one might be skeptical about the supposed metaphysical neutrality of powers: plausibly, a cause must be some kind of real existent! But on my view it is features (properties, events, etc.) or associated objects which are causes; and talk of powers is (again) just talk of what contribution the having of a given feature can make to the production of certain effects when the feature is instanced in certain circumstances. As such, in any given case of Weak emergence there are indeed two causes on the scene: the two features which share the token power—that is, which are such that their contributions to producing the effect in the circumstances overlap. Relatedly, in her note 6, Bennett says that "given the Proper Subset of Powers strategy, [Wilson] should not think that the effects of mental causes are overdetermined *at all*. For an effect to be overdetermined, it must have at least two distinct causes. But the only sense in which Wilson's [Weak emergentist/NRPist] thinks there are two distinct causes is that there are two distinct phenomena that literally share the efficacious part". Some (e.g., Shoemaker) might want to think of powers or associated potential contributions to the production of effects in mereological terms (as "efficacious parts") but even for such a person, it is the features having the power, not the power itself, that cause the effect. In any case, it's no part of my view that the real 'cause' of a Weak emergent effect is a shared power—so perhaps this clarification will assuage at least some of Bennett's skepticism.

3. Replies to Calosi

In his contribution, Calosi advances a novel mereology—a broadly formal theory of parts and wholes—which aims to (a) accommodate the possibility of metaphysical emergence, without (b) introducing non-mereological structure (as on variations on the theme of hylomorphism; see Koslicki 2008, Fine 2010, and Sattig 2015) or multiplying notions of parthood (as per Cameron 2007 and Canavotto and Giordani 2020). On Calosi's view, a single notion of sum provides the means of accommodating both reducible and irreducible—i.e., emergent—wholes.

The basis of Calosi's mereological framework (following Calosi and Giordani in progress*a* and in progress*b*) is a new conception of sum:

$$Sum(xx, y) \equiv xx < y \wedge$$
$$\forall x\,(\neg x \circ xx \rightarrow \neg x \circ y) \wedge$$
$$\forall x\,(xx < x \rightarrow y < x)$$

Sum is distinct from, and moreover stronger than, existing notions in the literature (see Cotnoir and Varzi 2021), in entailing each other notion while not being entailed by any. The associated mereology assumes an unrestricted composition principle (whereby any plurality of objects composes a Sum), and various axioms governing parthood, including antisymmetry, transitivity, and quasi-supplementation. System in hand, Calosi defines the notions of a 'simple' (having no object as a proper part) and a 'composite,' as the negation of 'simple'; and by appeal to the unrestricted composition principle defines a total function assigning to each object the 'matter' of the object, where the matter of a simple object is the object itself, and the matter of a composite object is the Sum of its components. Calosi is thereby able to distinguish between what he calls a 'Reducible Whole'—a whole that is identical to its matter—and an 'Irreducible Whole'—a whole that is not so identical, which distinction he takes to intuitively correspond to the distinction beween a whole's being 'nothing over and above' its parts (his illustrative cases

being a heap of sand and a lump of clay) vs. 'something over and above' its parts (his illustrative cases being tables, trees, organisms, and statues). And Calosi observes that, given all this, it follows that any simple object is Reducible, and any Irreducible object is composite; but the converse entailments do not hold (some Reducible objects may not be simple; some composite objects may not be Irreducible). Now for the connection to emergence:

> It should be clear why the present proposal has a chance to provide a mereology for emergent wholes: it allows for irreducible wholes that are something over and above their proper parts, i.e., their matter. Indeed, I suggest that, faced with cases of emergent wholes (E) we should endorse the following conditional:
>
> if emergent(x) then Irreducible(x)
>
> That is, Irreducibility as defined above is a necessary condition for emergence. (250–51)

Given that (as above) any Irreducible object is composite, it moreover follows on Calosi's system that

if emergent(x) then composite(x)

That is, being composite is a necessary condition for emergence. Calosi is officially neutral on whether being Irreducible (hence being composite) is sufficient for emergence, since he allows that there might be other 'grounds' for irreducibility. So as I understand Calosi's suggestion, if we have reason to think that some goings-on are emergent, then Calosi's mereology can accommodate them, at least to the extent of satisfying certain key necessary conditions. In this latter respect, Calosi takes his mereology to do better than certain alternative mereologies—most saliently, reductivist conceptions on which composition is identity, and eliminativist conceptions on which there are no composed entities (as per mereological nihilism), which (for reasons that I'll return to below) have been taken to be incompatible with the possibility of emergence, at least of a Strong variety.

By way of further motivating his proposed connection between mereology and emergence, Calosi argues that his account provides a basis for accommodating certain features of emergence as highlighted in my book. First, that emergents depend on yet are distinct from their bases is accommodated in that an Irreducible whole depends on its parts (its 'matter') in that "were we to annihilate its matter, it is unclear that anything would remain of the whole"; yet an Irreducible whole is by definition distinct from its matter. Second, that emergents are typically compositionally flexible is accommodated, at least potentially, in that Irreducible wholes are not identical to compositionally inflexible Sums (Reducible wholes). Third, that emergent entities typically fall under sortals (e.g., 'being a table' or 'being a statue') is accommodated by taking a given sortal to refer to an Irreducible whole as opposed to its Sum (matter). Correspondingly, one need not resort to a non-extensional notion of sum (as on Simons 1987) in order to make sense of, e.g., the applicability of the sortal 'statue' to a lump of clay. Finally, Calosi suggests, his mereological system provides a basis for the leveled structure associated with the special sciences, with special science entities at a level being Irreducible wholes that at each level emerge from sums of Reducible or Irreducible wholes, characteristic of the next level down.

Calosi and Giordani's distinctive mereological framework strikes me as in many ways intuitively plausible and theoretically powerful; in particular, it is a

significant accomplishment to identify and systematize a conception of Sum that unifies and asymmetrically entails existing conceptions. Moreover, I am inclined to agree with Calosi that his application of this framework can be seen as providing a basis for a common characterization of emergent entities as wholes that in some sense exist 'over and above' the mere sum of their parts, which in turn might be seen as confirming an also-common supposition that the notions of emergence and of mereology are deeply connected, such that (at a minimum) emergent entities are necessarily composite, and emergent features are necessarily features of composites.

Even so, in what follows I want to cast a somewhat skeptical eye on the extent to which Calosi's mereology can provide a basis for emergence, and on the more general supposition that emergence and mereology are necessarily connected. I'll start by arguing that while Calosi's application of his mereological framework plausibly provides a basis for *a* conception of emergence, this conception is different both from that which he seemed to have in mind in offering his illustrative cases of Reducible and Irreducible wholes, and from that which I aim to characterize in my book; and I'll draw out certain implications of this result for his project. I'll then highlight some considerations which indicate that the connection between emergence and mereology is not as deep (or necessary) as has sometimes been assumed.

To begin: recall that Calosi characterizes Reducible wholes as those which are (as he puts it) intuitively 'nothing over and above' their parts, with his examples being of unstructured entities or aggregates such as heaps and lumps of clay, and Irreducible wholes as those which are intuitively 'something over and above' their parts, with his examples being those of structured entities such as tables, trees, organisms, and statues; and he wants to make use of the distinction between Irreducible and Reducible wholes to at least make room for entities to be emergent, or not.[5] Now, an initial problem here, which poses a problem for identifying a purportedly Reducible heap or lump of clay with its 'matter,' is that heaps and lumps aren't identical to the sum of their scattered parts, in which case Calosi's mereology deems heaps and lumps Irreducible as opposed to Reducible wholes, and so doesn't distinguish his illustrative paradigm cases (which in turn were supposed to be candidates for non-emergent vs. emergent wholes). In any case, at best the Reducible/Irreducible distinction operative here is apt for distinguishing completely unstructured objects—mere collections, as fusions—from any at-all-structured objects.

Now, the distinction between structured and (completely) unstructured entities is no doubt important. It has played an important role, in particular, in discussions of the metaphysics of ordinary objects, as entities which are structured as opposed to unstructured collections of parts, as in Koslicki's (2008) motivating case of a (structured) motorcycle and an (unstructured) heap of motorcycle parts. But this distinction has not played an important role in debates about whether seemingly higher-level (ordinary, special scientific) goings-on are reducible or rather emergent.

[5] Calosi does not specify whether the emergence at issue is to be understood in Weak or Strong terms. In discussing the application of his framework to accommodating leveled structure of the sciences he seems to have Weak emergence in mind; but on the other hand concerns about whether emergence is compatible with reductive or eliminativist conceptions of composition typically suppose that the emergence at issue is Strong. In any case, which form of emergence is at issue won't matter for my present point.

To see why this is so, note that reductive physicalists, who think that any given special science entity or feature is type identical to some or other (perhaps logically or otherwise complex) lower-level physical feature, *take for granted* that the entities to which special science entities are identical are structurally complex (which is not to say that they are committed to composites as distinct from pluralities, about which more anon). It's no part of the reductive physicalist's view to maintain that tables, trees, organisms, or statues are identical to unstructured entities or aggregates. Rather, to take a toy example, a reductive physicalist might identify a certain table with a relational aggregate of atoms standing in atomic relations (or a disjunction of such aggregates, to allow for the table to persist through some change), and so on.[6] So the distinction between something that is in some sense just an unstructured sum of parts and something that is rather in some sense a structured aggregate isn't, at least in the usual cases, what is at issue in the physicalism debates, or in the related debates over whether or not there are multiple 'levels' of natural reality. And nor is it what is at issue in my attempts (in my book and elsewhere) to characterize metaphysical emergence in a way making sense of the appearances of higher-level reality. Rather, what is at issue in these contexts is the question of whether, *in addition* to whatever massively complex, typically highly structured, lower-level physical goings-on there might be, there are moreover any goings-on which are properly seen as (cotemporally materially) dependent on and (ontologically and causally) autonomous from the (massively complex, typically highly structured) lower-level physical goings-on that emergentists and non-emergentists alike agree exist.

Again, this is not to deny that there might be a different, weaker conception of metaphysical emergence that the broad distinction between unstructured sums and structured wholes might latch onto. It would serve, for example, to characterize an extreme form of reductive physicalist—call them 'the reductive pluralist'—who maintains that every apparently structured entity is really identical to some unstructured lower-level physical entity (or logical construction thereof). My point here is just that Calosi's conception of emergence as 'mirrored in' the distinction between (unstructured) Reducible and (structured) Irreducible entities is not obviously suited to accommodating metaphysical emergence of the sort at issue in debates over leveled structure, and which I aim to characterize.

The previous result has certain implications for Calosi's advertised characterization of his mereology as able to accommodate emergence without requiring additional (e.g., hylomorphic) non-mereological resources or multiplying notions of parthood. For insofar as the conception of emergence for which Calosi's system provides a basis is too weak to distinguish between non-emergent structured entities (of the sort the reductive physicalist accepts) and emergent structured entities (of the sort that Weak and Strong emergentists accept), it remains open that properly accommodating metaphysical emergence might require such additional resources or notions of parthood, after all. That said, it remains unclear to me whether we should be asking our mereological systems to do this work. So far as

[6] Nor is the reductive physicalist's characteristic rejection of there being multiple 'levels' of natural reality (as per, e.g., Heil 2003) based in the supposition that there are no structured wholes. Rather, reductionists as well as emergentists will accept that there are 'levels' of the sort that Calosi offers as 'mirroring' the Weak emergentist conception of multiple levels—though they will then deny that these mereologically-generated levels are tracking what is at issue between them.

I can tell, the conditions I provided on metaphysical emergence in my book don't rely, even indirectly, on any mereological notions.[7]

This brings me to my next topic, which pertains to the question of whether the notions of emergence and mereology are necessarily connected, as in Calosi's claims that if an object is emergent, then it is Irreducible, and (coupled with his supposition that if an object is Irreducible, then it is composite) that if an object is emergent, then it is composite.

Now, it is indeed sometimes claimed that composition is a necessary condition on emergence. For example, Baron (2019) says, "[m]ereological composition is usually thought to be at least a necessary condition on dependence: the emergent entity is composed of the entities from which it emerges" (2210). Calosi (2016a) agrees, saying that "An emergent property is a property which is exemplified by a composite object" (441).

As I see it, however, there are two good reasons to deny that composition is a necessary condition on emergence. First, even if it is granted that an emergent entity must cotemporally depend on a composite entity, as has often (though not universally; see below) been assumed for cases of both Weak and Strong emergence, the bearer of the emergent feature might not be composite. Consider the case of persons and their bodies. It is commonly maintained that persons are emergent, either Weakly or Strongly, in having Weakly or Strongly emergent mental states. But this much doesn't require that persons *themselves* be composite: perhaps they cotemporally depend on composites (bodies, or lower-level aggregates) without themselves having parts.[8] So there can be uncomposed emergent entities, and emergent features (e.g., mental states of non-composite persons) not exemplified by composites. Second, it's unclear that an emergent entity or feature has even to cotemporally depend on anything composite. One sort of possibility here involves a simple entity emerging, Weakly or Strongly, from another simple entity, when the latter is in appropriate circumstances. In the Weak case: perhaps the emergent is a determinable of a more determinate simple entity.[9] Perhaps that's a non-standard case, but it seems coherent to me. Another and quite standard option would involve the emergent entity (feature) cotemporally depending on a plurality (feature of the plurality). In my book I register this possibility, and more generally make room for the base-level goings-on to be pluralities or features of such pluralities (as opposed to, e.g., relational aggregates and features of such aggregates). In any of these cases, there might be emergence of either Weak or Strong varieties in the absence of composition as involving anything like a 'whole.'

[7] Of course, in some cases a given implementation of either Weak or Strong emergence might well involve the supposition that the emergent entities (features) at issue are composed (are features of composed entities); my own degrees-of-freedom-based account of Weak emergence is a case-in-point. But even here, the appeal to mereology is mainly serving as a way of ensuring that the condition on cotemporal metaphysical dependence (encoding the supposition of substance monism generally operative in accounts of emergence), is met; it is not itself serving as the basis for emergent autonomy.

[8] Would persons then not be 'concrete?' I don't see why not, given that they exist in spacetime (see Armstrong 1978).

[9] Note that the determinable-determinate relation is typically not cashed in mereological or related (e.g., conjunctive) terms. Most saliently, to be determinate is not to have a determinable as a proper part: determinates (unlike wholes) do not satisfy anything corresponding to supplementation.

The previous considerations undercut Calosi's necessary conditions on emergence, and more generally suggest that the connection between emergence and mereology might not be as intimate as Calosi and some others have taken it to be.

But what about arguments aiming to show that (the possibility of) emergence is incompatible with reductionist approaches to composition such as composition as identity (CAI), according to which mereological fusions are just identical to the plurality of their parts (see McDaniel 2008, Schaffer 2010, Calosi 2016a and 2016b)? Don't such arguments show that there is a deep connection between emergence and mereology, after all? Though I cannot address all such arguments here, I believe that their conclusions can be resisted, for reasons set out in Bohn 2009 (see also Cornell 2017 for a similar strategy). Bohn focuses his attention on the argument in McDaniel 2008, which Bohn schematically characterizes as follows:

1. Emergent properties are possible
2. If CAI is true, emergent properties are impossible
3. CAI is false

Here the focus is more specifically on Strongly emergent properties. Granting that Strongly emergent properties are possible (a claim with which I agree), why think that such properties would be incompatible with CAI? McDaniel's line of thought is that such an assumption leads to a violation of Leibniz's Law, according to which identicals are indiscernible. To start, let some xx be a plurality of two or more things, and let f(xx) be their compositional fusion. Now, assume that the fusion f(xx) has some Strongly emergent property F, understood (by McDaniel) as fundamentally novel as compared to the intrinsic properties of and spatiotemporal relations between the xx.[10] McDaniel then argues that insofar as F is fundamentality novel as compared to the intrinsic properties and spatiotemporal interrelations of the xx, F can be attributed to f(xx) but not the xx—but in that case, identifying the xx and f(xx) as per CAI would violate Leibniz's Law.

As Bohn correctly notes, however, McDaniel's reasoning here fails to appreciate that there's no problem with taking the plurality xx to have a fundamental collective property. As Bohn puts it, "according to the composition as identity theorist, any emergent property of the fusion should simply be thought of as a terminological variant of a fundamental plural collective property of all the parts, and vice versa. In that way the composition as identity theorist can hold that emergent properties do not violate the principle of indiscernibility of identicals" (221). This seems right to me, and I also agree with Bohn that a similar reply is available in response to those (including Calosi, who in his 2016b argues that a version of CAI is equivalent to mereological nihilism) maintaining that mereological nihilism is incompatible with Strong emergence.

So as it stands I remain unconvinced that emergence of any variety requires that there be composed wholes of the sort that CAI denies exist, or indeed any wholes at all. Pluralities, and even a single object, will do.

All this said, I suspect that there is new work for Calosi's mereology to do, even if it is somewhat different work than that advertised. In particular, and

[10] This characterization of Strong emergence departs in letter but not spirit from my preferred characterization, in ways related to the difference between a one-one and a one-many approach to metaphysical emergence, as discussed in Ch. 1, note 11 of my book; for present purposes nothing turns on the difference.

notwithstanding that debates over reduction and emergence have taken for granted conceptions of levels and their occupants making room, at a level, for structured as well as unstructured entities (and associated features)—such that, e.g., an atomistic physical level would contain not just atoms or pluralities of atoms, but also massively complex combinations of atoms standing in atomic relation—more work needs to be done as regards the details of how the domain of goings-on at a given level are generated. Boolean and classical mereological resources are also typically operative in generating 'lightweight' constructions of entities appropriately placed at a level, as I discuss in Ch. 1, Section 1.4.2, pertaining to the individuation of levels. Calosi and Giordani's system, and Calosi's attention to the difference between Reducible and Irreducible wholes, encode mereological resources which are both new and arguably 'lightweight.' These resources might well be added to the mix of those generating goings-on properly located at a level, and so be indirectly, if not directly, relevant to accommodating emergence, after all.

4. Replies to Emery

Emery's contribution raises a number of important questions stemming from an implemention of Weak emergence in terms of an elimination of degrees of freedom (DOF), of the sort I first offered in my 2010, and which plays a role in my book discussions of the emergence of complex systems (Ch. 5) and ordinary objects (Ch. 6). In this work, a DOF-based account is used to motivate the Weak emergence of certain special science goings-on from lower-level physical (e.g., quantum) goings-on. The overarching theme of Emery's questions concerns the extent to which attention to relations between DOF can be extended to address other cases of emergence—most interestingly, in my view, to cases of purported emergence within physics itself. A full treatment of Emery's unified set of questions deserves its own article; here I'll provide some initial response to what I see as her most pressing questions, and say a bit more about related questions in the footnotes.

Emery wonders, to start, whether a DOF-based implementation of Weak emergence might provide a fully general basis for Weak emergence—and if not, why not? To motivate my response to this question, it's worth recalling that my goal in the book is to consider whether, and ultimately to argue that, certain appearances of metaphysical emergence, drawn from both the special sciences and ordinary experience, can be taken at realistic face value. As such, I am looking to the sciences and to ordinary experience for input into which goings-on are, in those contexts and on the face of it, seemingly both dependent and autonomous in the ways characteristic of metaphysical emergence; and then my goal is to consider whether, and if so how, these appearances of metaphysical emergence can be taken at face value.

Now, a DOF-based implementation of Weak emergence reflects certain facts on the ground, including that certain special science entities are posited as having characteristic features encoded in associated special-scientific laws; that these entities are understood as composed by (systems of) lower-level entities which are also understood as having characteristic features encoded in associated (more) fundamental physical laws; and that the DOF needed to specify certain characteristic states of the former are eliminated as compared to the DOF needed to specify those same characteristic states of the latter. These facts, I argue, enter into a scientific

law-based motivation for thinking that some of the appearances of metaphysical emergence can be understood in terms of an elimination in DOF.[11]

Perhaps there are alternative ways of associating characteristic states of an entity with DOF which don't proceed by attention to scientific laws, in which case a DOF-based approach might be generalized to cover cases of Weak emergence involving such entities (or their features). But, two points. First, the availability of parameter-based accounts of characteristic states of, e.g., mountains, certain conscious mental states, or freely acting persons isn't obvious; in these cases other (e.g., functionalist or determinable-based) implementations of Weak emergence appear to be more naturally implemented.[12] Second, the conception of DOF as closely linked to certain laws plays an important role in my arguments for the conclusion that eliminations in DOF satisfy the conditions in the schema for Weak emergence, both in that the connection between DOF and laws is what blocks the reducibility of special science entities whose characterization involves eliminated DOF (since the lower-level laws require all the relevant DOF in order to operate), and in that insofar as laws express what an entity (system of entities) can or can't do, they also serve to encode what powers the entities have or don't have, in ways that suggest that entities with eliminated DOF as compared to the system of their composing/realizing entities will have fewer token powers than that system. Correspondingly, it's not clear that a DOF-based approach to Weak

[11] On this last, Emery also asks: what is necessary for a degree of freedom to count as eliminated? It can't be that the eliminated (e.g., quantum spin) degrees of freedom are *never* relevant to the behaviour of the entity at issue, for as I note in discussing the Weak emergence of ordinary objects from quantum goings-on, one can set up scenarios (e.g., a variation on Schrödinger's cat case) where quantum phenomena do impact the behaviour of the macro-entity. This question is related, in turn, to the question of which states, with associated DOF, are taken to be 'characteristic' of a given entity. Ultimately, I think that the answer depends on what makes for the sort of non-fundamental joint in nature of the sort plausibly encoded in special science laws. I don't have a general account of what makes for a non-fundamental joint, in part reflecting my view that there are many and diverse metaphysical dependence relations operative in cases of relative fundamentality (following my 2014 and elsewhere, to be given a broad defense in my forthcoming and under contract). But perhaps traction in the present case can be gained by attention to the usual view of special science laws as containing ceteris paribus laws or clauses, which allow for exceptions; and it might also be worth exploring (perhaps drawing on the degree-theoretic variation of the account of metaphysical indeterminacy advanced in Wilson 2013 and Calosi and Wilson 2018) whether Weak emergence comes in degrees, with non-fundamental joints being to some extent fuzzy or metaphysically indeterminate.

[12] A related line of thought applies to Emery's question of whether the DOF-based Weak emergence of ordinary objects might be gained, not by way of the elimination of quantum DOF (as I do in the book), but rather by way of the elimination of broadly statistical-mechanical DOF. Indeed, I appeal to thermodynamic features as having eliminated statistical-mechanical DOF in support of certain complex systems being Weakly emergent, reflecting the applicability of renormalization group methods to such systems when near critical points, which methods track the elimination of such DOF. I focus on the quantum case in the chapter on ordinary objects mainly because, again, there's a clear scientifically endorsed line of thought which (unlike applications of the renormalization group to gasses and other complex systems) targets ordinary objects—and not because a case for Weak emergence needs to involve an absolutely fundamental base; I agree that it doesn't.

emergence can be generalized to cases where no laws are operative without undercutting the motivations for the approach in the first place.[13]

To my mind, the most pressing of Emery's questions pertains to whether and how my schemas for metaphysical emergence might accommodate cases of such emergence within physics itself. As Chalmers (2021) observes:

> Discussion of "emergent spacetime" has exploded, driven largely by theories of quantum gravity—including versions of string theory, loop quantum gravity, and causal set theory—in which spacetime may not appear on the fundamental level. […] The key thesis is that spacetime exists at a nonfundamental level and is grounded in a fundamental level which is nonspatiotemporal. (164)

(See, e.g., Lam and Wüthrich 2018 and Huggett 2021.) Not just quantum gravity (QG), but general relativity (GR) itself (as presupposing relationism; see Rovelli 2007) and quantum mechanics (QM) (if the wavefunction/configuration space is taken as fundamental; see, e.g., Albert 2013, Ney 2021) have been taken to support spacetime as emergent in that spacetime is not fundamental, but is rather completely dependent on more fundamental nonspatiotemporal goings-on. Note that the supposition that emergent spacetime (or its three-dimensional occupants) are nonfundamental indicates that the type of emergence being posited here is of the Weak rather than Strong variety.

Such applications are in prima facie tension with my schemas for metaphysical emergence. One source of tension, observed by Emery, is that the notion of cotemporal material dependence often involves the composition of the entity having the emergent feature by lower-level dependence base entities; but in the cases at issue it is unclear how elements of the more fundamental physical ontology would 'compose' the emergent physical ontology (as Baron 2019 discusses; but see Baron and Bihan 2022 for an attempt to make sense of this). Now, my own view (as I register in my replies to Calosi, above) is that compositional relations aren't required for there to be emergence, but even so, one might be concerned that the dependence condition in the schemas is too restricted to make sense of cases of purported emergence within physics. Let's focus on the purported emergence of spacetime. Recall that the dependence condition encodes substance monism, whereby the only matter is physical matter, along with minimal nomological supervenience, whereby an emergent feature *S* requires and is at least nomologically necessitated by ('minimally supervenes on') cotemporal base-level goings-on *P*. As such, the dependence condition presupposes spacetime: *S* is

[13] Emery also wonders whether attention to DOF might enter into an implementation of Strong emergence, as involving a new DOF—and if not, why not? I didn't advance a DOF-based implementation of Strong emergence mainly because I didn't see clear case studies involving the posit of new DOF. As I discuss in Ch. 5 (182-5), so-called 'order parameters' are sometimes presented as involving new DOF (by, e.g., Morrison 2012 and Lamb 2015), but on closer examination no new DOF are really at issue: either the DOF are present at the micro-level, and what is new is their taking on certain values, or else the order parameters are not genuine DOF, but rather 'phenomenonological descriptions' of a system's order. That said, if there were cases where an apparently new DOF could not be given a reductive or other deflationist treatment, and given that the new DOF was associated with behaviours, law-governed or not, then a DOF-based implementation of Strong emergence might well make sense—though in such a case it's not clear that we would be adding anything new beyond the existing claim that a Strongly emergent feature has a fundamentally novel power.

cotemporal with *P*, and entities possessing these features will typically (per substance monism) share matter, hence spatially overlap.

The autonomy conditions in my schemas also presuppose spacetime: though a power may never be exercised, in any case powers are had by, and causal relations hold between, phenomena which are spatially located; moreover, accounts of causation take this to be either diachronic or synchronic, and so presuppose the notion of time.[14]

My conceptions of metaphysical emergence are not unusual in taking spatiotemporal notions for granted; effectively all standard conceptions do so. Those exploring the status of spacetime as emergent typically recognize that there is a prima facie difficulty in taking ST to be emergent by lights of standard accounts, and in response weaken the notion of emergence by removing references to space or time. There are a couple of different strategies on offer here, but in my view it is not clear that these attempts succeed—effectively, because satisfaction of the weakened conditions is compatible with either reduction or with Strong emergence, contrary to the intended characterization in these suggestions of spacetime as Weakly emergent from more fundamental nonspatiotemporal ontology.

One sort of strategy involves characterizing the dependence and autonomy conditions in ways eliding reference to spatiotemporal notions, as in Crowther's (2022) characterizations in terms of

1. dependence (cashed in terms of asymmetric supervenience correlations)
2. novelty (cashed in terms of qualitative difference)
3. autonomy (cashed in terms of multiple realizability or determination)

Crowther distinguishes 'hierarchical' emergence (a non-ST form of cotemporal emergence) where the base is somehow present, and 'flat' emergence (a non-ST form of diachronic emergence) where ST results from a non-causal 'interaction'.[15] And she argues that on certain accounts of quantum gravity, spacetime satisfies the dependence and autonomy conditions vis-à-vis the specified non-spatiotemporal basis, in hierarchical or flat fashion (and maybe both).

But granting satisfaction of these conditions in some or other versions quantum gravity, the associated weakened conception of emergence is too weak to establish that spacetime is less fundamental than what it depends on:

> A's asymmetrically supervening on B doesn't entail that A is less fundamental than B.[16]
>
> A's being qualitatively different with respect to B doesn't entail that A is less fundamental than B.
>
> A's being multiply realized/determined by B, C, and D doesn't entail that A occupies a less fundamental level than B, C, and D; for if A is identical to the disjunction of B, C, and D (as reductionists typically maintain), A will be as fundamental as the disjunction. That is reduction, not emergence.

[14] "The lack of a metric structure […] seems to result in the loss of causation since, on the face of it, causation requires (at least) time to exist. […] Causation is usually thought to be a relation between events, which are individuated by their spatiotemporal locations" (Baron 2019: 2208). That said, some recent conceptions of causation do not build in the notion of time; see Baron and Miller 2014 and Tallant 2019.

[15] Here the model is something like the occurrence of the big bang.

[16] For example, determinables asymmetrically supervene on determinates; but many think quantum determinables are prior to their determinate values.

> A's being multiply realized/determined by B, C, and D is also compatible with the base phenomena serving as diverse preconditions for something fundamentally novel. That is Strong, not Weak, emergence.

As such, as they stand Crowther's conditions on emergence are too weak to rule out either (identity-based) reduction or Strong emergence. Moreover, on the face of it this weakness reflects the elision of spatiotemporal notions from these conditions. The best shot for establishing genuine autonomy of a Weak emergent variety proceeds by resisting reductionist and Strong emergentist readings by attention to causal considerations, and more specifically via satisfaction of the proper subset of powers condition, which blocks Strong emergence, since such emergence requires a novel power, and blocks reductionism, since disjunctive features are instanced by instancing a disjunct, and each disjunct has more token powers than are had by the Weakly emergent feature. As above, such causal notions appear to presuppose spacetime, and this is true as well on a DOF-based implementation of Weak emergence. That said, in other work Crowther (2018) suggests that a DOF-based implementation of Weak emergence can make sense of the emergence of spacetime from at least some nonspatiotemporal fundamental ontologies:

> Wilson's (2010) weak ontological emergence, where an emergent theory may be characterised by the elimination of degrees of freedom from the underlying theory […] is certainly applicable if spacetime emerges as illustrated by the condensed matter approaches to QG, and it applies to GFT, and any other approaches where spatiotemporal degrees of freedom emerge as collective, low-energy variables, analogous to those of thermodynamics. It also may apply in the context of LQG, where degrees of freedom possessed by the spin foams are eliminated in the approximation and limiting procedures designed to resolve and/or wash-out their discrete nature and quantum properties in the recovery of spacetime. (84)

These are intriguing suggestions. If Crowther is correct, and in a way I hope she is, then I would need to back off, at least for the case of spacetime, from the claim in my book that satisfaction of the conditions in the schemas is 'core and crucial' to metaphysical emergence of the sort connecting special science and fundamental physical goings-on. I'd need to say something more general.[17] Though my arguments that eliminations in DOF suffice to block reductionism and Strong emergence presuppose that DOF are associated with broadly causal laws, perhaps the same line can be implemented using a non-causal notion of information. This is something I'm working on. At present it's not entirely clear to me that there is a workable conception of Weak emergence—one which ensures dependence with autonomy—that abstracts away from causal or other spatiotemporal considerations.

A second strategy aimed at accommodating the emergence of spacetime involves appealing to a specific relation as holding between spatiotemporal and non-spatiotemporal ontology, suitable for seeing the former as dependent yet autonomous from the latter. Here the most popular suggestion appeals to something like functional realization:

> On a functionalist picture, whether an entity (a structure, object or property—from now on I will just say "structure") counts as spatiotemporal is determined by its

[17] Or disjunctive—but that would be unsystematic.

> functional role. The functional role of a physical structure is its role in the physical laws, which often boils down to its implications about the motion of material objects. (Baker 2020: 278)

This suggestion is subject to the sort of considerations I discuss in my book when discussing functional realization in special-scientific contexts. To start, we must distinguish between 'realizer' functionalism, on which functionally implemented goings-on are identified with the realizer of the role, and 'role' functionalism, on which functionally implemented goings-on are identified with the role itself, usually understood as a kind of higher-order property. Realizer functionalism is compatible with (indeed, is a form of) identity-based reductionism, and so is unsuited for purposes of vindicating the metaphysical emergence of spacetime from nonspatiotemporal ontology. Role functionalism potentially does better; and here (following the literature in metaphysics of mind/science), what's needed is some reason to think that there exists such a second-order feature. And the usual means of doing this is by appeal to the multiple realizability of spacetime. But as I'm at pains to highlight in my book, a mere appeal to multiple realizability does not suffice to establish the irreducibility of the multiply realized feature. In particular, work must be done to rule out a disjunctive treatment of the multiple realizability at issue. And again, the main strategy for doing this (mine) appeals to causal considerations, so won't work here—though it may be that looking to eliminations of DOF is the best bet here.

But suppose it turns out that no implementation of a (nonspatiotemporal) variation of my schemas for metaphysical emergence can make sense of the purported emergence of spacetime (or its occupants). In that case, I'll here register that there are alternative, and to my mind more natural, ways of thinking about some of the relations between nonspatiotemporal and spatiotemporal goings-on than in terms of metaphysical emergence. In particular, we have in hand certain metaphysical conceptions of how concrete goings-on are related to comparatively abstract goings-on, including ones on which abstract universals (not in space and time) come to be concretely instantiated, and ones on which among the space of abstract possibilities (not in space and time), just one comes to be actualized. This last seems especially relevant to the present case; for if (following Allori) the wavefunction represents possible ways the world or objects in the world can be, then configuration space is properly seen as a modal space, with concrete goings-on being best understood as instantiations or actualizations of these possibilities. These relations—instantiation, actualization—deserve further investigation and attention. For present purposes, what is important is that there is no clear sense in which the instantiation of a universal, or the actualization of a possibility, is any less fundamental than the universal/possibility. So why think that the relation between configuration space and ordinary spacetime and its occupants entails that the latter is less fundamental than the former? Either way, the relation isn't one of metaphysical emergence per se—in which case the inability of an account of metaphysical emergence to apply to these cases doesn't pose a problem for the account. But again, as with other of the questions Emery raises, there is more work to be done in arriving at a considered answer.

5. Replies to Gozzano

Gozzano's comments address the interesting question of whether the common supposition that Weakly emergent mental features are multiply realizable—or as

he puts it, are 'realization indifferent'—is compatible with the plausible supposition that mental features are 'systematic', in entering into patterns of dependencies. Gozzano expresses the potential threat to mental features' being Weakly emergent in the form of an argument:

> (i) Mental features are systematic;
> (ii) (In many cases) Emergence entails realization indifference;
> (iii) Systematicity entails that realization indifference cannot hold;
> (iv) Therefore, (in many cases) mental features can't be emergent. (271)

(Gozzano puts aside Strong emergence, as implausible; hence here and elsewhere his references to 'emergence' are more specifically to Weak emergence.) Each premise in this argument, Gozzano maintains, can be defended; and the conclusion therefore follows.

The focus of my response in what follows is on premise (iii), but let me start by saying a bit more about (i) and (ii).

First, in re the claim that mental features are systematic. Gozzano doesn't offer a definition of 'systematicity', but does offer a number of illustrations, including cases where increases in the intensity of a perceptual stimulus are (e.g., logarithmically) systematically associated with the intensity of the phenomenal state, and cases where changes in the intensity of a phenomenal state (e.g., pain) are systematically associated with an increase in some other phenomenal state (e.g., anxiety). Though I'm not sure about the status of these particular examples, I think that what Gozzano has in mind here is that there might be relations—better, to avoid ambiguity, 'mappings'—between (to speak loosely) families of mental feature types whereby members of one family are systematically related with members of the other family. I'm happy to grant that various special science laws, including those of psychology and neuropharmacology, will at least sometimes encode these sorts of systematic mappings between (families of) mental features.

Second, in re the claim that many cases of emergence entail realization indifference, three observations. To start, I'd prefer 'involve' over 'entail', since whether a given higher-level feature is multiply realizable is an empirical, not logical, matter. Next, Gozzano's discussion involves a characterization of 'realization indifference' as building in the possibility of 'wildly different' realizers; this goes beyond the usual appeals to multiple realizability as motivating Weak emergence, which appeals often involve realizers being only 'mildly' different, as when, e.g., my belief that Paris is beautiful is realized by different neurological states, or the shape of a flock of birds is realized by different configurations of its constituent birds, will do. As such, in what follows I will usually revert to the usual terminology of multiple realizability, but will revisit whether the possibility of 'wildly different' realizers makes any difference down the line. Finally, as Gozzano notes, I don't take multiple realizability to be either necessary or sufficient for weak emergence: not necessary, since there are cases to be made that some singly realized features satisfy the proper subset condition on powers; and not sufficient, since reductionists have strategies for accommodating multiple realizability in disjunctive or other terms, which must be blocked before multiple realizability can be assumed to involve emergence. All this said, Gozzano is right that many cases of Weak emergence, of mental features in particular, are initially and primarily motivated by multiple realizability; so it is definitely worth considering

whether these suppositions are in tension with the also-plausible assumption that mental states enter into systematic mappings.

I now turn to the key premise (ii) in Gozzano's main argument—namely, the claim that 'Systematicity runs against realization indifference'.

The underlying motivation for Gozzano's endorsement of this claim appears to be a supposition that if special science properties enter into systematic mappings, then the lower-level properties upon which the special science properties cotemporally materially depend must also stand in systematic mappings. As he puts it,

> if we consider the causal relations in which [systematic special science feature] *S* is involved [as encoded in] high-level laws of the sort discussed by special sciences, we may require a sort of systematic counterpart of supervenience: there cannot be systematic variations at a high level without systematic variations at a low level. (272)

In this sense, Gozzano supposes, the existence of a systematic mapping between (families of) higher-level features places constraints on the realizers of these features—constraints which, he maintains, are not in place in cases of multiple realizability.

Gozzano offers a specific subargument in support of this claim and the associated premise in his main argument. In the interest of efficiency I will focus my critical attention primarily on a key premise (2) in that subargument, according to which (and consonant with the previous line of thought), if a property *S* is systematic,

> (2) The *P*s on which *S* cotemporally materially depends (CMD), should follow the same pattern of systematicity shown by *S*. (274)

Now, it is unclear why we should accept this. As Gozzano observes:

> One may wonder why the emergentist should accept [this] premise […]. The emergentist can stress that each "level of reality" […] is characterized by its laws […] and on which *S* cotemporally materially depends. So, what consequences would bear [on] having different systematic relations, if any at all? (274)

The complaint here seems to me to be apropos, as far as it goes. Even granting that systematic mappings between (families of) higher-level features requires systematic mappings between (families of) lower-level features, why would these mappings have to 'follow the same pattern'? Indeed, it's not clear that higher-level systematicity mappings require lower-level systematicity mappings. All that ultimately seems required to accommodate systematic mappings involving realized features is that their lower-level realizers enter into laws compatible with those higher-level systematic mappings. Maybe those lower-level features and laws will also fall into 'systematicity patterns', but at the end of the day all that's required is that any given realizer of any given higher-level feature *S* provide a suitable basis for *S*'s having the powers it needs to have to conform to whatever systematicity mappings are in place.

So, Gozzano's premise is better expressed as requiring not that realizers enter into the 'same pattern of systematicity' as *S*, but just that (at most) whatever laws are in place as regards *S*'s realizer on a given occasion serve as an appropriate

basis for accommodating the systematic mappings into which S enters. To assess whether systematicity runs against multiple realizability, then, the question is whether there are reasons to think that a feature's being multiply realizable somehow poses a problem for its realizers' accommodating the systematic mappings into which S enters.

I answer in the negative; I don't see any problem here. Since at issue are cases where multiple realizability ends up motivating Weak emergence, let me put the point in my favoured terms. To fix ideas, suppose that mental features M_1, M_2, and M_3 are systematically causally connected to mental features M'_1, M'_2, M'_3; suppose also that each of these six types of mental features is multiply realizable; and suppose that (after undercutting reductionist strategies) this multiple realizability is taken to support these features' satisfying the conditions on Weak emergence vis-à-vis whatever features realize them on a given occasion. Here the systematic mapping (like Gozzano's illustrative cases) causally connects certain mental features with certain others; hence to accommodate this mapping just requires that any realizer of M_1 has among its powers the power to cause M'_1, any realizer of M_2 has among its powers the power to cause M'_2, and so on. But on the operative understanding of realization, this follows automatically, since any token power of a realized (Weakly emergent) feature on a given occasion is identical to a token power of the feature that realizes it on that occasion. So the treatment of M_1 as both multiply realized and Weakly emergent is compatible with M_1's entering into the systematicity mapping; and similarly for M_2 and M_3. So systematicity is here accommodated, notwithstanding the multiple realizabilities of the features at issue.

Note also that we were able, in this narrative, to remain neutral on whether the realizers of the mental features themselves enter into a systematicity mapping, whether similar to or different from those into which the mental features enter. Whether this is so will depend on further details about the powers and power profiles of the realizers. This bears on premise (5) of Gozzano's subargument according to which "If [the realizers] have different projectability patterns and support different counterfactuals, they do not establish the same systematic relations" (274). To be sure, the realizers can be expected to enter into different projectability patterns and support different counterfactuals (it is precisely this difference that provides a basis for thinking that Weak emergents are distinctively efficacious, in spite of not having any new powers), and let's even grant that the realizers themselves don't enter into systematic mappings at all, much less 'the same' ones into which mental states enter. None of those further details matter for whether multiple realizers can accommodate higher-level systematic mappings, as the previous case illustrates. All that matters is that the realizers have the requisite powers—as they will do, on my account of Weak emergence.

This seems to me to be a coherent narrative, indicating that there is no in-principle problem with there being systematic, multiply realizable mental (or other) features.

It remains, however, to consider two strategies for defending Gozzano's claim to the contrary. The first reflects Gozzano's characterization of multiple realizability as realization 'indifference', such that the diverse realizers at issue may be 'wildly' different—so different that they might share nothing in common:

> Let's consider pain: supposedly, in humans, it is realized by C-fiber firing, but it could be differently realized in other sentient beings and the realizers form an open set. So, we may take the property of being in pain as one that at a very high level

> can be shared by different entities, from human beings to other mammals, to other animals up to potentially extra-terrestrial individuals. At a finer level of detail, being in pain is multiply realized by structures that may have nothing in common. (270)

Supposing it were the case that diverse realizers of a single feature might have 'nothing in common'—in the case of systematically related mental features, in particular—then I can see how Gozzano might conclude that systematicity runs against realization indifference. My response here is simply that I reject the supposition that realizers might 'have nothing in common', since that supposition leaves it unclear how or in what sense one feature might realize another. If, as I argue is the case for the broadly scientific (including mental) features that are the target of my book, the feature whose realization is at issue has a distinctive power profile, then at a minimum any realizer of a feature *S* has to have, among its powers, the powers of *S*. (And as I also argue, a wide range of accounts of realization, including functional realization and the determinable-determinate relation, agree.) On such an understanding of realization, effectively encoded in the schema for Weak emergence, this much will be 'in common' among multiple realizers of a feature, and as per the case above, that much seems sufficient unto the task of accommodating systematicity.

The second strategy pushes in a different direction, and is suggested by Gozzano's discussion of what powers should be taken to be in the power profile of a given feature:

> According to the subset strategy a property is individuated by the set of its causal powers had by all its instances […] But the causal powers defining the set do have causal relations to other powers. Say, a rubber band is elastic and green. Elasticity is shared among all elastic entities no matter their color. But elasticity determines fragility in cold conditions. Should we consider this as a condition on other elastic entities? […] Should the elasticity also involve a specific ratio between, say, thickness and length of stretchability? If so, then it could be the case that only a specific realizer fits the bill. But if this is the case, then it seems Kim was right after all: each disjunct has its own merits and the high level is just a measure of our ignorance. (275)

Here one can see Gozzano as maintaining that closer examination of the powers associated with a given property indicates that powers are much more finely individuated than is usually recognized, to the extent that the claim that features, including those entering into systematic mappings (which impose yet further constraints on powers) are not appropriately seen as multiply realizable. My response starts by observing that, although this is often mainly left tacit for simplicity, talk of 'powers' in these contexts is intended as talk of 'conditional powers', such that powers are individuated not just by their effects, but also by the intrinsic and extrinsic conditions required for the powers to be manifested or exercised. Hence any given property will be associated with massively many conditional powers—not just 'the power to stretch without breaking', but 'the power to stretch without breaking if instantiated in warm conditions', and so on. All these conditional powers are had by any instance of a feature, even if the conditions of manifestation of the power do nor or even cannot obtain (as when a plastic knife has the property of being knife-shaped, which includes among its powers the power to cut wood if made of steel). This understanding strikes me as unifying and systematic,

and in line with the connection between (in particular) scientific taxonomy and laws, so I am inclined to stick with it, rather than adopting such a fine-grained conception of powers that hardly any features turn out to be multiply realizable.

To return to Gozzano's primary argument: since I can reasonably deny that 'Systematicity runs against realization indifference', I can resist Gozzano's conclusion that considerations of multiple realizability don't support the Weak emergence of mental features—especially those entering into systematic mappings.

That said, I want to close by registering that Gozzano has called something important to attention—namely, that broadly holistic considerations may turn out to be relevant to discussions of metaphysical emergence. Discussions of metaphysical emergence have tended to focus on individual cases—this mental feature, that thermodynamic feature, and so on. But how do systematicity mappings and other more global considerations bear on this topic? For example, in the case above, might M_1 and M_2 be Weakly emergent and M_3 Strongly emergent, or is there some reason to think that systematically related features should, or even must, have the same status? This is a new question, and deserves further attention.

6. Replies to Onnis

In *Metaphysical Emergence*, I motivate my powers-based schemas for Weak and Strong metaphysical emergence by attention to Kim's problem of higher-level causation, which I present as "the most pressing challenge to taking the appearances of emergent structure as genuine" (39). Onnis's contribution is aimed not at directly problematizing the schemas themselves, but at calling into question their underlying motivation in Kim's problem of higher-level causation. She aims to argue that Kim's argument proceeds against certain metaphysical presuppositions—each associated with 'Alexander's Dictum', according to which to be real is to have causal powers—which, if rejected or differently interpreted, would render the argument less of a challenge so far as accommodating emergence is concerned. As she summarizes:

> [T]here are three issues that need to be addressed. The first one concerns the Dictum itself: one may want to reject it and assume other criteria about existence. The second one is about the power-based interpretation of the Dictum: one may want to accept the latter, while considering its power-based interpretation as too strict. The third one is about the metaphysical underdetermination of the powers involved in the power-based interpretation: one may want to accept the Dictum and its power-based interpretation, while requiring a differentiation between microscopic physical powers and macroscopic emergent powers. (296)

Since the problematic presuppositions at issue concern powers, one can see Onnis here as pushing back not just on the stated motivations for my schemas, but more pressingly on my claim that the powers-based schemas are 'core and crucial' to metaphysically accommodating the appearances of emergence.

The considerations that Onnis raises are well worth attention. Even so, as I will now argue, at the end of day the metaphysical presuppositions she identifies as underpinning Kim's problematic are not required for this problematic to put pressure on the viability of metaphysical emergence; hence the motivation for my powers-based schemas as indeed 'core and crucial' to accommodating such emergence remains.

6.1 Alexander's Dictum

As discussed in the *Précis*, I set out Kim's overdetermination problem as involving six premises, four of which (Reality, Distinctness, Efficacy, and Dependence) encode certain assumptions about the seeming higher-level features at issue, and two of which (Physical Causal Closure and Non-overdetermination) encode certain assumptions about causation. The basic concern is that any purported effect of a (real, distinct, dependent) higher-level feature is (per Closure) already brought about by the lower-level physical goings-on upon which the higher-level feature depends, and so is (contra Non-overdetermination) overdetermined. As I observe, standard responses to Kim's argument are associated with certain views, denying some or other premise. Of these views, only those denying Physical Causal Closure (i.e., British emergentism) or Non-overdetermination (i.e., non-reductive physicalism) accommodate metaphysical emergence, understood as coupling cotemporal material dependence with ontological and causal autonomy (distinctness and distinctive efficacy); and the strategies encoded in these two views motivate my schemas for emergence, whereby a higher-level feature has a fundamentally novel power as compared to its dependence base feature on any given occasion (Strong emergence), or a higher-level feature has a proper subset of the token powers of its dependence base feature on any given occasion (Weak emergence).

Now, Onnis maintains that Kim's overdetermination argument presupposes Alexander's Dictum (after British emergentist Samuel Alexander), commonly spun (e.g., by Kim 2006: 557) as the thesis that 'to be is to have causal powers'. To start, Onnis observes, Kim takes Alexander's Dictum to motivate the Efficacy premise in his argument (perhaps given the Reality premise in his argument), insofar as he registers that "to be a mental realist […] mental properties must be causal properties" (1998,: 43). Moreover, in his (2006), Kim goes further, saying "Properties that are lacking in causal powers—that is, whose possession by an object makes no difference to the causal potential of the object—would be of no interest to anyone" (557), again connecting this thesis to Alexander. Onnis goes on to claim that Kim's problem requires and gets traction only under the assumption of Alexander's Dictum:

> If the principle is rejected, entities can have a legitimate existence even without exerting causal efficacy. If the nonreductive physicalist has to give up her nonreductionism, therefore, it is because of Alexander's Dictum. (292–93)

I respond that it isn't clear either that Kim accepts Alexander's Dictum, or that Kim's problem gets traction only if one assumes this Dictum. As regards Kim's own proclivities, it is worth noting that his expressions of claims in the ballpark of Alexander's Dictum (as in his 2006, above) are uniformly offered in a context within which he is presenting the emergentist's point of view, as opposed to his own. In any case, Alexander's Dictum is very broad; it aims to provide a general necessary condition on the existence of goings-on of any ontological category whatsoever. As such, one might reject the Dictum in full generality—perhaps because one believes that platonic universals or numbers exist, but don't have causal powers—yet still maintain that for scientific or concrete entities and features, to be is to have causal powers. Indeed, Kim's focus in his discussion of overdetermination is squarely on broadly scientific features, so it isn't obvious that he intends

to advance anything as strong as Alexander's Dictum, understood as a general criterion of existence.

That said, Onnis is correct that Kim's problematic takes as a premise that mental (more generally: special scientific) features have powers, as per Efficacy. But we don't need Alexander's Dictum to motivate this premise. Independent of that Dictum, the efficacy of special science features is motivated by their entering into special science laws which standardly express causal regularities (chemical reactions, geological forces, biological processes, predator-prey relationships, neurological and psychological interactions, and so on). And we moreover have direct experience of the seeming efficacy of the qualitative mental features that are Kim's primary focus, as is reflected in nomological truisms such as that (ceteris paribus) being in pain causes avoidance behaviour, being hungry causes one to seek out food, and so on.

These independent motivations for taking the higher-level features at issue in Kim's problematic to be efficacious would remain even if one rejected Alexander's Dictum, either in full or in part, perhaps on grounds (as Onnis suggests) that certain motivations for thinking that some goings-on exist don't explicitly require the efficacy of said existents.[18] It would remain that there are theoretical (law-based) and experiential reasons for thinking that mental and other special-scientific features are efficacious; and given the other premises in Kim's argument, his challenge for there being emergent higher-level goings-on would unfurl accordingly. To be sure, the epiphenomenalist responds to Kim's problematic by denying Efficacy; but to offer an epiphenomenalist response to Kim's problematic is not to say that there was never a problematic there in the first place. On the contrary, in the dialectical course of events the burden is on the epiphenomenalist to explain away the science-based and experience-based motivations for Efficacy—a burden not easily discharged, which may account for the relative paucity of epiphenomenalists.

6.2 A Heavyweight Notion of Powers?

I next turn to Onnis's claim that, even granting Alexander's Dictum (at least as applied to mental and other scientific features), Kim's interpretation of the Dictum presupposes a conception of efficacy as involving powers that are real in some metaphysically heavyweight sense. As Onnis interestingly argues, such a conception appears to be at odds with Alexander's own comparatively lightweight correlational conception of efficacy. She moreover suggests that a heavyweight conception of powers "seems to already carry anti-emergentist implications", insofar as such powers are a ready target of reductionist strategies. For example, Onnis observes that on one implementation of Taylor's (2015) 'collapse' objection to the viability of Strong emergence, any purportedly fundamentally novel powers at the

[18] By way of such alternative motivations, Onnis considers being introspectively accessible (as I suggest provides defeasible motivation for our taking libertarian free choice to exist) or being indispensible to our best science. Introspection of free will seems to me to satisfy Alexander's Dictum twice over, insofar as a free choice causes both the awareness of the choice and the outcome of the choice. Indispensibility considerations look better by way of a genuine alternative motivation for existence—perhaps causally inert mathematical entities are required for our best theories. In any case, the availability of such alternative motivations doesn't undercut the specifically causal considerations which motivate mental and other special-scientific goings-on.

higher level can be traced to dispositional properties of base-level constituents. Onnis suggests that less committal conceptions of the efficacy at issue "seem to make the problem of higher-level causation less challenging".

It is true that Kim frames his problematic in terms of powers, as in his Causal Inheritance principle and elsewhere. So far as I can tell, however, all that Kim has in mind in his talk of 'causal powers' associated with a given property is that the having of the property 'makes a difference' to the causal potential of an object—that is, to what the feature (or an object having the feature) can cause when in certain circumstances. Such an understanding is in line with the metaphysically neutral understanding of powers operative in *Metaphysical Emergence*, according to which talk of powers is talk of the contribution that the having a property can make, when instanced in appropriate circumstances, to the production of a given effect. This neutral understanding does not require that powers be understood as dispositions or in any other heavyweight terms; as I argue (33), even a contingentist categoricalist Humean could accept powers in the sense operative in the schemas.

In any case, suppose that the operative notion of efficacy/causation and associated use of 'power' is given a weak—say, Humean—reading in Kim's problematic. Would Kim's argument then pose less of a threat to accommodating the appearances of higher-level reality, as involving emergent special science features? One motivation for a positive answer might proceed as follows. To start, consider the sort of scenarios that are not supposed to be good models for making sense of higher-level causation: namely, firing squad or double-rock-throw cases. Why think that it would be problematic if mental or other special science causation were overdetermined like this? The concern seems to reflect a kind of 'oomphy' understanding of efficacy, where different causes directed at the same effect would, like different substances trying to occupy the same space, get in each others' way. And perhaps such an 'oomphy' understanding is more naturally associated with a heavyweight notion of powers, as real dispositions or the like.

But even supposing a more metaphysically substantial notion of efficacy or power provides one route to finding causal overdetermination problematic, it isn't the only way. Another route simply lies in observing that, whatever the right account of causation, and whatever (in particular) is going on in firing squad and double-rock-throw cases, it remains that mental causation is *not that kind of case*—the relation between the mental goings-on and their physical dependence base is just different from those sorts of overdetermination cases. And yet certain of the premises in Kim's argument suggest that higher-level features would overdetermine the effects of their lower-level bases. That's really all that the 'Non-overdetermination' premise is registering; and Humeans as well as non-Humeans can and typically do agree that this is enough to get the problematic going.

Moreover, just because one accepts a Humean or other lightweight understanding of causation and associated talk of 'powers', or prefers to dispense with talk of powers altogether (even as shorthand for saying what can cause what), it isn't clear that the problem of higher-level causation thereby becomes less challenging. As I observe in Wilson 2002, if causal power is understood just as a matter of nomological sufficiency (in the circumstances), then insofar as base-level properties are nomologically sufficient for higher-level properties, and nomological sufficiency is transitive, then any power purportedly had by the higher-level

property will also be had by the base property.[19] Hence a version of the Collapse objection against Strong emergence attaches even to a lightweight conception of efficacy/powers.[20] And as I also observe in Wilson 2002, if causal power is understood just as a matter of nomological necessity (in the circumstances), then in any case where the higher-level property is multiply realizable, then the physical base-level property will be ruled out as efficacious.[21] In that case it would appear that Physical Causal Closure is violated, and Kim's problematic again comes into play, illustrating a prima facie challenge in reconciling higher-level causation with a broadly physicalist world-view.

So the force of Kim's problematic overdetermination argument does not hinge on commitment to a heavyweight conception of efficacy or powers. Luckily, or so I argue in my book, physicalists and non-physicalists alike have the resources, either in general or via appropriate implementations of the schemas for Weak and Strong emergence, to respond to the full range of ways in which Kim's challenge may be brought to bear.

6.3 Microscopic vs. Macroscopic Emergent Powers

Finally, I turn to Onnis's claim that taking there to be a "difference in kind" between higher-level and lower-level powers "might be able to weaken the problem of high-level causation":

> By examining the nature of causal powers, for instance, it might be discovered that higher-level powers cannot really collapse, while lower-level ones cannot really emerge. Emergent and non-emergent causal powers, in other words, might simply be non-interchangeable powers of a different kind. (300)

Onnis goes on to offer a preliminary characterization of the difference between 'emergent' and 'non-emergent' powers. The latter, she suggests, are associated with properties of micro-objects (e.g., the mass of an electron), and are commonly thought to be "fundamental, essential, intrinsic, intrinsically active, and productive". The former are associated with properties of macro-objects (e.g., the hardness of a diamond), and "are often conceived as nonfundamental, extrinsic, context-sensitive, and constraining", as on Gillett's (2016) understanding of 'machresis' as a form of non-productive 'role-shaping' determination. Onnis speculates that "the most striking difference between micropowers and emergent powers would therefore be the intrinsic activity and productivity of the former and the

[19] As I there illustrated: "[S]uppose one of my brain properties necessitates one of my mental properties, and the mental property bestows some causal power on me. [If] causal power bestowal is just a matter of nomological sufficiency, my brain property will, in virtue of necessitating the mental property, also bestow this causal power on me" (Wilson 2002: 64).

[20] I respond to this and other versions of the Collapse objection in my book (drawing on Wilson 2002 and Baysan and Wilson 2017), but the present point is just that the threat of Collapse does not hinge on a heavyweight conception of efficacy/powers.

[21] As I there illustrated: "The general idea is this: suppose either of two of my brain properties is sufficient for one of my mental properties, and the mental property bestows some causal power on me. Since we're assuming that causal power bestowal is a matter of nomological necessity, as well as sufficiency, and since neither brain property is necessary for the effect in question, neither brain property will bestow this causal power on me" (Wilson 2002: 65).

extrinsic non-productive constraining capacities of the latter" (300). And re the Collapse concern, she suggests that

> differentiating between micropowers and macropowers might make this collapse more difficult. For instance, let's suppose that the macroscopic causal powers exerted by a biological complex system require a biological complex bearer. In that case, a nonbiological system or a biological isolated component could not instantiate those macropowers, which would therefore become non-collapsible. (300)

Onnis notes that these suggestions are preliminary, but even so let me say why I'm not inclined to take on board any such distinction in kinds of powers. To start, I don't speak of 'emergent powers' (or non-emergent powers); it is features, or perhaps entities having the features, which are emergent (or not) on my view. And as above, the conception of 'power' operative in my book encodes just that (talk of) powers associated with a given feature is (talk of) what contributions the having of the feature may make to the production of certain effects, when in certain circumstances. Such a neutral characterization makes sense, so far as I can tell, whatever sort of feature or entity is at issue. Nor would I be inclined to endorse a conception on which emergent and non-emergent features (or associated powers) differ in fundamentality status, both because Strongly emergent features (powers) are just as fundamental as whatever fundamental physical features (powers) there might be, and because the physical features (powers) serving as a co-temporal dependence base for higher-level features (powers) will themselves typically be features of highly complex micro-configurations, and so not themselves be fundamental. I would also resist any general characterization of emergent features (powers) as 'constraining', not just because cases of Strong emergence needn't involve constraints, but also because cases of Weak emergence needn't do so (as on a determinable-based implementation); and even when Weak emergence does involve constraints, it is lower-level goings-on, not higher-level powers, which impose the constraints (as on the degrees-of-freedom-based implementation discussed in §5.2.4 of my book).

That said, I agree with Onnis that further investigations into the nature of powers might open the door to new strategies for responding to at least some concerns about emergent features. Indeed, Onnis's suggested response to the Collapse objection is quite similar to the 'new bearers' strategy which I discuss in Ch. 4 (135), which appeals to Baysan's (2016) view that features have their powers derivatively on the powers of their bearers. But note that whether one wants to go this route to avoid Collapse will depend on whether one is inclined to accept Baysan's view (which as it happens, I'm not). Moreover, the question will remain of whether the macrofeatures (powers) at issue in a given case are or are not in line with physicalism—which brings us back to the terrain of Kim's problematic.

To sum up: while it's worth asking whether Kim's problematic is generated by Alexander's Dictum or related controversial assumptions, my general answer is 'no, it isn't so generated'; and similarly for the Collapse concern for Strong emergence. Rather, these problematics are surprisingly robust across heavyweight and lightweight conceptions of efficacy and powers. As such, for those aiming to realistically accommodate the appearances of metaphysical emergence, the powers-based responses encoded in the schemas for Weak and Strong emergence remain the only game in town.

7. Replies to McLaughlin

In his contribution, McLaughlin raises several important questions about or concerns for my views. My responses here will focus on the following: first, whether my 'no fundamental mentality' account of the physical needs to embrace further constraints; second, whether satisfaction of the conditions in the schema for Weak emergence is either necessary or sufficient for physical acceptability; and third, whether Strong emergence, understood as involving fundamental powers or associated interactions which come into play only at certain levels of compositional complexity, is compatible with quantum field theory.

I start with a quick clarification. McLaughlin describes my account of the physical as one according to which the physical "[…] is whatever would be posited by the completed physics in fact true of our world, with the following caveat: A mental feature is not to be counted as a physical feature even if that physics would posit it" (280); and he describes the associated constraint on physicalism as one according to which "any doctrine deserving of the name "physicalism" should be incompatible with the physics in fact true of our world having to posit mental phenomena" (280). If by a 'posit' of physics we just have in mind the (most) fundamental entities or features treated by that theory, then these descriptions coincide with my account of the physical and the associated constraint on physicalism, respectively. But since physics also in some sense posits non-fundamenta (e.g., protons and other particles composed of quarks) and more generally treats certain non-fundamental complexes (e.g., pluralities or relational aggregates), it's worth being clear that what I rule out as 'physical' are any goings-on that are (as I put it) 'fundamentally mental', in being both (a) fundamental and (b) individually such as to have or bestow mentality, of the sort, e.g., that panpsychists suppose exist—hence the 'no fundamental mentality' (NFM) constraint. The NFM account is compatible, e.g., with physics treating non-fundamental physical states (consisting of some massively complex combination of fundamental physical goings-on) that are either identical with (as on a reductive physicalist view) or which realize (as on a non-reductive physicalist view) mental features.

Now, in re my NFM account of the physical, McLaughlin considers whether I would accept further constraints on the physical—e.g., a 'no fundamental chemical' and 'no fundamental biological' constraints—and speculates that I would do so:

> I think [Wilson] would […] accept such additional constraints. It is clear, for instance, that if the physics in fact true of our world would have to posit entelechies or a fundamental vital force, she would take physicalism to be false. (280)

McLaughlin doesn't present the potential need to introduce further constraints as an objection, but other things being equal, I would prefer not to introduce such further constraints, since it seems to me that doing so would be unsystematic. As I earlier put it:

> One might wonder whether imposing the NFM constraint leads to an unsystematic account of the physical. The NFM constraint is motivated by […] intuitions to the effect that physicalism would be falsified if there turned out to be fundamentally mental entities. But intuitively, physicalism would also be falsified if we were to find that entities at relatively low orders of constitutional complexity were moral

> or freely acting agents, or that aesthetic responses involved a new fundamental interaction or force. Similarly (recalling Driesch and Broad) for chemical, biological and other non-mental, seemingly higher-order features of reality. […] So shouldn't those endorsing a physics-based account of the physical impose, in addition to the NFM constraint, no fundamental morality, no fundamental free will, no fundamental aesthetics, no fundamental chemistry, no fundamental biology, and no miraculous powers constraints? But then, the concern goes, the resulting account of the physical will be unsystematic and ad hoc; for what are mentality, morality, aesthetics, chemistry, biology, and miracles supposed to have in common, that rules them out as being physical? (Wilson 2006: 75)

In my 2006, I aimed to avoid such further constraints in a 'divide and conquer' fashion. As regards fundamental chemistry and biology, I said

> Given that chemical and biological features of reality can, in actual fact, be ontologically accounted for in terms of configurations of […] entities that are not themselves chemical or biological (as all parties to the physicalism debates seem generally prepared to agree), there is no need to explicitly rule these out as being […] fundamental […]. (75)

And for the rest, I argued that insofar as each plausibly involves mentality, no constraint beyond the NFM constraint is needed (76).

This divide and conquer strategy still seems to me to work, but in re the potential need for 'no fundamental chemistry' or 'no fundamental biology' constraints, I now think that something more principled can be said—namely, that these constraints are not needed because chemical and biological goings-on, unlike mental goings-on, are essentially such as to be or be features of comparatively compositionally complex phenomena, such that it would make no sense for individual fundamental physical goings-on, which by the definition of physics are comparatively non-complex, to have chemical or biological features. McLaughlin's question made me realize that there is an important difference here as regards the potential threat of non-mental and mental phenomena so far as characterizing the physical is concerned; for while chemical and biological phenomena might be fundamental in being Strongly emergent (since the advent of such emergence is compatible with, and typically involves, compositional complexity), they could not be fundamental in the sense of being or being features of compositionally basic phenomena. Hence it is, perhaps, that no correlates of panpsychism (panchemism, panbiologism) have been advanced for either chemical or biological features of reality.

I turn next to two concerns that McLaughlin raises for my account of Weak emergence. The first has to do with the whether satisfaction of the conditions in my schema for Weak emergence suffices to render Weak emergents physically acceptable (given the physical acceptability of the base level goings-on). McLaughlin thinks not:

> The nomological requirement on Weak emergence is that if a feature *S* Weakly emerges from a physical feature *P*, then *P* is minimally nomologically sufficient for *S*. That condition is compatible with the law linking *S* and *P* being a fundamental law of nature, a law that doesn't hold in virtue of other laws and conditions. […] The existence of fundamental [e.g.,] psychophysical laws is incompatible with physicalism, reductive or non-reductive. […] To avoid this result, the

> condition of cotemporal material dependence must be amended [...] to include the requirement that the law linking *S* and *P* not be a fundamental law of nature; it must be a law that holds in virtue of physical laws and physical conditions (284).

I see McLaughlin's point as in a similar vein to a concern raised by Melnyk (2006). In *Metaphysical Emergence* I present the general concern as follows:

> [W]hatever makes it the case that some proper subsets of token powers of a given lower-level physical feature correspond to (instantiated) higher-level features, while other subsets do not do so, had better itself be physically acceptable if the higher-level features are to be physically acceptable; yet satisfaction of the conditions in Weak Emergence is silent on why a given higher-level feature *S* has the distinctive power profile it has, and so is compatible (one might think) with the instantiation of a higher-level feature's being, somehow or other, the outcome of a physically unacceptable process. (106)

One can develop the concern by noting (as I do in Wilson 2010) that the satisfaction of the proper subset of powers condition is frequently associated with the holding of certain lower-level constraints; as Melnyk correctly observes, if the holding of these constraints ensues as a matter of some physically unacceptable process (say, if the constraints hold as a matter of God's will), then the physical acceptability of the higher-level feature would be thereby undercut. In my 2010, I explicitly require that the constraints at issue be a matter just of physical or physically acceptable processes, and in *Metaphysical Emergence* I register that if an amendment to the schema for Weak emergence is needed, it would likely involve explicitly incorporating this sort of requirement (107).

McLaughlin's comment can be seen as developing the concern in a way that does not specifically advert to constraints, by attention to the possibility that emergent and base features are connected by fundamental laws, as makes sense for Strong but not Weak emergence. And here too I would say that there may well be a case for making the sort of amendment McLaughlin suggests, and requiring that any laws holding between base-level and Weak emergent features hold solely in virtue of physical laws and conditions. That said, rather than bifurcating accounts of the cotemporal material dependence condition which at present is common to the schemas of Weak and Strong emergence, I would prefer to insert any such amendment into the autonomy condition on Weak emergence, to the effect of requiring that any constraints *or laws* operative in making it the case that a given feature is associated with only a proper subset of the token powers of the lower-level base feature be constituted or otherwise determined by lower-level physical processes and/or laws.

McLaughlin also raises the concern that satisfaction of the conditions on Weak emergence is not necessary for metaphysical emergence of a physically acceptable variety. In particular, he suggests that on a 'role-functionalist' view taking higher-level states to be second-order functional states "of being in some state or other that has certain causal effects [where] the first-order states that have those effects realize the functional state" need not be understood as imposing the autonomy (proper subset of powers) condition:

> It is open to a role functionalist to maintain that a functional state, a state of being in some state or other that has certain effects, does not itself cause those effects. Its

> realizers do. That's compatible with functional states figuring in causal explanations of the effects in question. But it is incompatible with Weak emergence. (285)

McLaughlin's suggestion here seems to reflect his position that, while role-functionalism "cannot avoid epiphenomenalism" (McLaughlin 2006: 39), this much does not prevent role-functionalists from adopting "a weaker notion of causal relevance" (one not requiring of a causally relevant feature that it actually cause anything) on which it suffices for a feature to be causally relevant that it be causally 'explanatory'—say, by "providing information about the causal history of an action". Here I'll just say that such a weak understanding of causal relevance is too weak to capture the sense in which we want higher-level features to be efficacious—e.g., as entering into seemingly causal special science laws, or as mental causes of our agential behaviours. Relatedly, such a weak notion of relevance seems ripe for reductive or eliminativist treatment of role-functional features in (mere) conceptual or pragmatic terms (per, e.g., Heil 2003). So on the assumption that role-functionalist features are epiphenomenal, that they don't satisfy the conditions for Weak emergence doesn't pose a problem for my view. That said, it seems to me that role-functionalists can resist the charge of epiphenomenalism, and more specifically can maintain that such properties satisfy the conditions in Weak emergence, for reasons I set out in my book (Wilson 2021: 59–60).

Finally, I turn to McLaughlin's concern that Strong emergence, understood (as on my preferred implementation) as involving a novel fundamental interaction, is incompatible with current physics—in particular, with quantum field theory (QFT), which aims to unify quantum mechanics and special relativity, and is the foundation of the standard model of fundamental particle physics:

> In the field dynamics of quantum field theory, interactions are *local*. They are local in that fields directly interact with other fields only at spacetime points. That is to say, the dynamics of each field at any spacetime point are directly influenced only by the values and derivatives of the other fields at that same point, and not by anything happening elsewhere. That fundamental interactions are local is inextricably baked into the theory. Quantum field theory could, for instance, accommodate new kinds of particles and new kinds of fundamental forces. But the discovery of fundamental configurational interactions would refute the theory. It thus isn't just that quantum field theory doesn't now posit fundamental configurational interactions, it cannot countenance them. Such direct fundamental interactions would involve whole regions of spacetime. That is incompatible with relativity theory. (288)

More specifically, McLaughlin goes on, the enormous success of QFT defeats the considerations I offer for thinking that there is libertarian free will (to wit: that we have direct experience of ourselves as choosing, and that there are presently no good reasons for thinking that we cannot take this experience at realistic face value):

> Quantum field theory has been enormously successful in its regime of applicability, and […] human brains fall well within that regime. The truly enormous empirical support quantum field theory enjoys soundly defeats any intuitions we might have about there being a fundamental force of will. (288)

I offer four lines of response to McLaughlin's objection.

First, it is incorrect that the supposition that fundamental interactions are local, in the sense that fields directly interact only at points, is "inextricably baked into" QFT.[22] To be sure, standard quantum field theory textbooks often claim that interactions are local in this sense, but (as claims in textbook presentations of physical theories often are) this claim is a gloss, which upon closer examination is metaphysically, theoretically, and historically inaccurate.

The usual gloss is metaphysically inaccurate—or at least, metaphysically suspect. To start, field operators are not definable at points unless the theory is fully regulated (rendered non-divergent) in the UV regime. In continuum QFT, field operators must be treated as operator-valued distributions—i.e., one only gets an operator by integrating the distribution against a test function with support on a compact region (i.e., by averaging the field values in a small region around the point), which results in a field observable that is not even gauge invariant. The metaphysical picture encoded in this procedure is murky, and if anything seems to suggest that fields interact not at points, but rather in the compact vicinity of points.[23] Relatedly, the usual means of dealing with UV divergence in local QFT results in a QFT which is an 'effective' field theory, the import of which is precisely to gloss over what exactly is happening at the small-scale limit. Physicists have identified tools (most saliently: renormalization strategies) enabling QFT to be useful for capturing the long distance physics while allowing us to remain agnostic about the short distance physics. But given this understanding of effective QFT, it's clear that there are lots of ways the short distance physics could be. Indeed, there is nothing in QFT itself qua effective theory that demands that what lies below the limit of applicability is even a quantum field theory, much less one that is local (or nonlocal)![24]

The usual gloss is also theoretically and historically inaccurate, since as it happens attention to nonlocal QFT goes back at least to the 1940's and is alive and well today. As Tomboulis (2015) recently put it:

> Nonlocal field theories is a subject with long, albeit spotty, history. Despite the success of perturbative renormalization in QED in the late forties, the idea that local interactions may be a low energy approximation to fundamental underlying nonlocality of interactions continued to be prominent in the fifties and the subject of many investigations [1].[25] Subsequently, nonlocality was considered mostly in

[22] Thanks to Michael Miller and Patrick Fraser for helpful discussion here.

[23] See also the discussion of the 'localization problem' in Saunders 1992.

[24] This is an epistemic point. Interestingly, however, certain metaphysical readings of the effectiveness at issue (say, as involving a lower limit to the precision of the field values, per Miller forthcoming) might also undercut the claim that interactions in QFT occur at points in a continuum.

[25] "[1] R.P. Feynman, Phys. Rev. 74, 939 (1948); A. Pais and G. E. Uhlenbeck, Phys. Rev. 79, 145 (1950); P. Kristensen and C. Møller, Dan. Mat. Fys. Medd. 27, no. 7 (1952); W. Pauli, Nuovo Cimente, 10, 648 (1953); M. Ebel, Dan. Mat. Fys. Medd. 29, no. 2 (1954); M. Chretien and R. E. Peierls, Nuovo Cimento 10, 668 (1953); M Cretien and R. Peierls, Proc. R. Soc. London A223, 468 (1954); C. Hayashi, Prog. Theor. Phys. 10, 533 (1953); ibid, 11, 226 (1954); N. Shono and N. Oda, Prog. Theor. Phys. 8, 28 (1952); F. Bopp, Ann. d. Physik, 42, 573 (1942); H. Mc Manus, Proc. R. Soc. London A195, 323 (1948); G. Wataghin, Z. Phys. 86, 92 (1934)" (26).

> the context of axiomatic field theory [2].[26] In more recent years it has attracted renewed interest in connection with nonlocal theories of gravity [3] - [9],[27] as well as the nonlocality of string field theory vertices and various nonlocal models in cosmology and other areas, see [10][28] and extensive reference list therein. (2)

Others advancing versions of nonlocal QFT include Nobel laureate H. Yukawa,[29] K. Namsrai,[30] G. Fleming,[31] M. Moffat,[32] and R. Landry and J. Moffat.[33] It's clear, then, that physicists do not see the locality of interactions as "inextricably baked into QFT".

There's good reason why nonlocal QFT is of perennial interest as an alternative research program to local QFT. It's not just that local QFT is subject to UV divergence, though that is part of what drives physicists to look elsewhere. As Fleming (1987) observes, the original and continuing motivation for exploring nonlocal QFT reflects concerns "over the internal consistency of a theory requiring infinite renormalization and the long-standing recognition that local interactions generate that requirement". As above, getting any predictions out of QFT requires adopting perturbative methods involving expansions which, unless arbitrarily cut off, give rise to infinities. To be sure, "at the level of comparing renormalized perturbation theory calculations with experiment …[t]he methods work wonderfully!" Still …

> [T]hrough all these years since Dyson, Feynmann, and Schwinger formulated renormalization theory, it has never shed its fundamentally *ad hoc* character. It remains a recipe for extracting finite results from an infinity-plagued formalism by cancelling the infinities against one another systematically. What is wanted is a formulation of non-trivial interacting QFT that never encounters the infinities in the first place. (Fleming 1987: 98–9)

[26] "M. Meyman, Sov. Phys. JETP 20, 1320 (1965); V. Efimov, Com. Math. Phys. 5, 42 (1967); ibid, 7, 138 (1968); M. Z. Iofa and V. Ya. Fainberg, Theor. Mat. Fiz. 1, 187 (1969); M. Z. Iofa and V. Ya. Fainberg,, Sov. Phys. JETP 29, 880 (1969); V. Ya. Fainberg and M. A. Soloniev, Ann. Phys. 113, 421 (1978); V. Ya. Feinberg and M. A. Soloviev, Theor. Math. Phys. 93, 1438 (1992)" (26–27).

[27] "E. T. Tomboulis, arXiv:hep-th/9702146; [4] T. Biswas, E. Gerwick, T. Koivisto and A. Mazumdar, Phys. Rev. Lett. 108, 031101 (2012) [arXiv:1110.5249]; [5] T. Biswas, A. Conroy, A. S. Koshelev and A. Mazumdar, Class. Quant. Grav. 31, 015022 (2014) [arXiv:1308.2319]; [6] L. Modesto, Phys. Rev. D 86, 044005 (2012); [7] L. Modesto, Astron. Rev. 8.2, 4 (2013) [arXiv:11202.3151]; L. Modesto, arXiv:1402.6795[hep-th]]; F. Briscese, L. Modesto and S. Tsujikawa, Phys. Rev. D 89, 024029 (2014) [arXiv:1308.1413]; G. Calcagni and L. Modesto, Phys. Rev. D 91, 124059 (12015) [arXiv:1404.2137 [hep-th]; L. Modesto and L. Rachwal, Nucl. Phys. B889, 228 (2014) [arXiv:1407.8036]. [8] M. Isi, J. Mureika and P. Nocolini, JHEP 1311:139 (2013) [arXiv:1310.8153 [hep-th]]. [9] V. P. Frolov, arXiv:1505.00492; V. P. Frolov, A. Zelnikov and T. de Paula Netto, arXiv:1504.00412" (27).

[28] "N. Barnaby and N. Kamran, JHEP 0802, 008 (2008)" (27).

[29] See in particular Yukawa 1950a and 1950b.

[30] See, e.g., Namsrai 1986.

[31] See, e.g., Fleming 1987.

[32] See, e.g., Moffat 1990.

[33] See Landry and Moffat (forthcoming).

The deeper motivation for exploring nonlocal QFT is that the assumption of locality itself underlies UV divergence. As Tomboulis (2015) puts it:

> It has long been realized, more or less explicitly, that UV finiteness (or at least superrenormalizability in the presence of gauge interactions) can be achieved by nonlocal interactions. (2)

Of course, UV finiteness isn't the only theoretical desideratum. In addition, theorists want QFT to satisfy unitarity and causality, in a way compatible with relativity. Tomboulis goes on:

> [On nonlocal QFT], unitarity can be preserved, at least perturbatively, provided appropriate analyticity conditions can be imposed on the nonlocal interactions. Causality, however, is a central concern whose investigation has remained woefully inadequate, both in the classical theory, where it is inexorably connected with the mathematically proper formulation of the initial value problem (IVP), and in the quantum theory. (2)

In any case, many nonlocal versions of QFT claim to avoid UV divergence while accommodating both unitarity and causality. For example, Namsrai (1986) constructs "a nonlocal theory of quantized fields by means of the hypothesis of *spacetime stochasticity*", and Fleming (1987) formulates a nonlocal QFT involving spacelike hyperplanes:[34]

> Hyperplane dependence of the dynamical variables of quantum theory, and consequently, their eigenvectors, is the minimal generalization of the concept of time dependence that is required to establish a manifestly Lorenz covariant formalism. […] The reason that hyperplane dependence has not previously become a prominent conceptual tool of theoretical physics [reflects that] contemporary fundamental theories of many-particle systems are expressed in terms of basic quantized fields that are themselves associated with simple points of space-time. [But this line of thought] may be unnecessarily restrictive. The experience my students and I have gained, in exploring the possibilities, allowed for interactions of particles with external potentials when hyperplane dependence is explicitly incorporated into the formalism, and suggests the possibility that consistent Lorentz-invariant quantum field theories with nonlocal interactions may be possible if the fields are hyperplane-dependent. I will suggest below a model of such a theory. (97–8).

In discussing Fleming's view, Saunders (1992: 379) suggests that a relaxing of the demand for local covariance, to be replaced in particular by the weaker requirement of hyperplane dependent covariance, may well be "all but inevitable". Yet more recently, Landry and Moffat (forthcoming) say:

> We discuss the nonlocal nature of quantum mechanics and the link with relativistic quantum mechanics such as formulated by quantum field theory. We use here a nonlocal quantum field theory (NLQFT) which is finite, satisfies Poincaré

[34] A spacelike hyperplane is a three-dimensional, metrically flat section of the flat Minkowski space-time continuum, such that any two points in the hyperplane are separated by a spacelike interval, and such that for any such hyperplane, there is an inertial frame of reference in which all the points of the hyperplane are simultaneous, and all points simultaneous with any point of the hyperplane are in the hyperplane.

> invariance, unitarity and microscopic causality. This nonlocal quantum field theory associates infinite derivative entire functions with propagators and vertices. We focus on proving causality and discussing its importance when constructing a relativistic field theory. […] The result is free of UV divergences and we recover the area law.

Suffice to say that nonlocal QFT is a research program with a long history that people are still actively pursuing.[35]

Third, it's not clear that any Strong emergence there might be would violate microcausality. To start, note that any demand for locality in QFT had better be compatible with entanglement; and indeed it is, since the locality characteristic of QFT is one supposed to preserve "microcausality", whereby no causal influences can travel faster than the speed of light. Entanglement phenomena don't violate microcausality, and so don't violate locality in that sense; rather, they violate separability, according to which the wave-function for the system as a whole is factorizable as a product of wave-functions for the system's parts. In this sense, entangled systems are irreducibly holistic, with a common spin (no pun intended) being that entangled particles are not really distinct; hence it is that for one entangled particle to "influence" another does not require faster-than-light (or any) causal connections. (Or so the story goes.) Now return to Strongly emergent phenomena. These are often characterized in terms evocative of failures of separability: a Strongly emergent feature is one which cannot be factored or otherwise reduced to features of its parts. Moreover, the failure of reduction here is one according to which a Strongly emergent feature is holistic, in arising (in this context) under conditions of compositional complexity, with a common spin on such features being that they render the system that has them a unified whole, whose parts are not really distinct. These similarities suggest that on the face of it, Strongly emergent features, like entangled systems, would violate separability, not microcausality.

That said, in my book I argue that entanglement phenomena are not in general clear cases of Strong emergence, since the failure of reduction might be understood as involving Weak emergence from a spatiotemporally extended dependence base. Strong emergence, on my view, involves a fundamentally novel power, which in turn (on my preferred implementation, and as motivated by the case of the weak nuclear interaction; see my 2002 and 2021) involves a novel fundamental interaction which comes into play only at certain levels of compositional complexity. How would this work? Well, whatever is going on here, it won't be a matter of instantaneous causal influences. Rather, on the usual assumption that fundamental interactions are associated with fields, Strong emergence would involve a new fundamental field (or fields) coming into play, which would presumably interact with other fields/interactions in operation, just as standardly posited fields/interactions do. How, exactly, and what theoretical and empirical consequences this would have, would sensitively depend on the nature of the interaction between the standard fields and the new field(s), which as in the case of standard fields/interactions would be an a posteriori, empirical matter. For present purposes it suffices to note that there is no in-principle barrier to understanding Strong emergence in this way,

[35] It may also be worth noting that, as Weinberg (1997) observes, QFT as standardly formulated is not fully either nonlocal or Lorentz invariant: "there are complications when you have things like mass zero, spin one particles for example; in this case you don't really have a fully Lorentz invariant Hamiltonian density, or even one that is completely local" (7).

and indeed (again, see my discussions of the weak nuclear interaction) there is some historical precedent for doing so.

Fourth, though for the reasons above there's no clear conflict between Strong emergence and QFT, it's worth noting that McLaughlin's claim (following Carroll 2021) that QFT "has been enormously successful in its regime of applicability, and […] human brains fall well within that regime" (288) involves a massive and to my mind unjustified extrapolation. As Carroll himself observes,

> Particle-physics experiments typically examine the interactions of just a few particles at a time, so new physical laws that only kick in for complex agglomerations of particles are not necessarily ruled out by data we currently have (2021: 28).

In that case, though, why think that "particles obey the same equations whether they are inside a rock or inside a human brain" (27), contra applications of Strong emergence to mental phenomena such as (in my book) libertarian free will? Here Carroll appeals to the status of QFT as an effective theory targeting low-energy states, which can be interpreted as collections of interacting particles. Insofar as human beings, like rocks, can (under decomposition) be thought of as such collections, they fall in the regime of applicability of QFT. But the true measure of a theory's "applicability" is predictability, not the fact that, as Carroll puts it, the theory "is meant to be accurate" (18) for phenomena in some or other energy regime. And QFT provides no predictive basis for any human behaviour, unlike the remarkably successful predictions we make through understanding our own and others' mental states. On the face of it, then, McLaughlin's extrapolation, like Carroll's, requires assuming that there are no new fundamental configurational interactions or laws—that, as a synchronic variation on Hume's problem of induction, the physical laws of nature "will continue the same".[36] But like Hume's problem, that assumption builds in what the argument from QFT is supposed to show.

For the various reasons above, I conclude that attention to QFT poses no in-principle difficulty for Strong emergence. But no doubt there is more to say here, and I thank McLaughlin (and Carroll) for raising this important question to salience.

8. Replies to Paolini Paoletti

In his contribution, Paolini Paoletti raises two questions pertaining to the metaphysics of properties, as potentially relevant to my schema for Weak metaphysical emergence. The first question presupposes (correctly, in my view) that in general,

[36] Carroll also says that "if there are additional particles and forces, they interact too weakly with the known fields to exert any influence on human behavior; otherwise they would have already been detected in experiments" (2021: 18). But again, as Carroll notes, the experiments that have been so far conducted are limited to examining "the interactions of just a few particles at a time" (28), far below the complexity at which, e.g., Strongly emergent mental features are supposed to exist or be instantiated. To be sure, if Strong emergence involves the coming into play of a new fundamental interaction, then once such an interaction is on the scene it could (in principle) have theoretical or empirical consequences for interactions involving systems at lower levels of complexity; but whether this would be the case would be an empirical matter.

not every proper subset of powers associated with a given physical feature *P* is associated with a Weakly emergent feature. In that case, one can ask:

(1) What makes it the case that a given proper subset of powers associated with a given lower-level physical feature is associated with a Weakly emergent feature?[37]

The second question presupposes that features can be individuated in a way independently of their powers. In that case, one can ask:

(2) What makes it the case that a given feature *S* is associated, with at least nomological necessity, with a given causal profile?

Paolini Paoletti considers certain candidate answers to these questions, and finds them wanting. He then advances essence-based answers to these questions—but, he maintains, an essence-based approach is in tension with the supposition that "everything whatsoever is physical or fully depends on the physical" (311), such that Weak emergence turns out to be "not so weak", after all.

Now, as Paolini Paoletti notes, I don't aim in my book to answer either question. In re the first question: in my book and elsewhere I take for granted what I call the prima facie appearances of metaphysical emergence in the sciences and in ordinary experience, as coupling dependence with ontological and causal autonomy; and then I argue that in various cases we can make sense of these prima facie appearances—most commonly, as satisfying the conditions in the schema for Weak emergence. In cases of broadly scientific properties, for example: what explains why scientists have posited certain higher-level scientific properties as having certain subsets of powers, as is reflected in these properties' entering into certain special-science laws? I discuss certain broadly empirical motivations which seem to be operative in some cases (upon which I'll expand below), but ultimately I take this to be a question for the (natural and social) scientists. My job, as I see it, is just to show that one can make metaphysical good sense of such posits. And in re the second question: as I further discuss below, this question arises only for those holding certain metaphysical views of features (properties and the like) and powers—in particular, those who think that features can be individuated independently of their powers—in the usual case, via a quiddity or primitive identity, which can then be somehow associated or not associated with certain powers. My own view is that there is no reason to think that features of the sort under discussion in my book are associated with quiddities or any other kind of non-causal aspects, in which case the second question doesn't arise, though I also argue in my book that the viability of the schemas for emergence is neutral on whether features are associated with quiddities.

All this said, one way to read the intended import of Paolini Paoletti's remarks is that if one *does* attend to these questions, one will see that they interestingly bear on how Weak emergence should best be understood, and on whether Weak emergence (properly understood) can provide a satisfactory basis for nonreductive physicalism. So in what follows I start by arguing that answers to the first question are plausibly both diverse and empirical, as are answers to the second question as it arises for those accepting quiddities or other non-causal aspects of properties. I'll then follow up by offering reasons to reject a thesis that enters into Paolini Paoletti's critical assessment of certain strategies for answering his

[37] I phrase this and the second question in terms of "what makes it the case" that *P* as opposed to why *P*, in order to sidestep cases where $\neg P$.

questions—namely, Sider's principle of 'Purity' (see Sider 2011: 126–132), according to which the constituents of fundamental facts must themselves be fundamental. Finally, I raise some concerns with Paolini Paoletti's positive "essence-based" answers to the questions, and relatedly, with his claim that if (as on his preferred answers), a higher-level feature and its causal role are in some sense mutually essentially dependent, this poses a problem for physicalism understood as requiring that "everything […] fully depends on the physical" (311).

To start, then: what makes it the case that a given proper subset of lower-level physical powers is associated with a higher-level Weakly emergent feature? This is a question of general interest, whatever one's metaphysics of properties, at least for those who accept that there is or may be Weak emergence. In my book I discuss some of the considerations motivating scientific posits of certain higher-level features having certain causal profiles. One common answer, which I discuss in Ch. 3 in presenting my DOF-based approach to Weak emergence, adverts to there being certain conditions or associated constraints present at the lower level, which serve to eliminate certain microphysical degrees of freedom as required for characterizing the law-governed properties and behaviour of the higher-level feature (which elimination in DOF in turn operates to eliminate certain powers as had by the feature). A different but related consideration, which I discuss in Ch. 5 in motivating the claim that certain complex systems are Weakly emergent, adverts to the suitability for a given complex system to be modeled by the Renormalization Group Method, which in turn reflects that the system ceases to have a preferred length scale—which again serves to eliminate certain lower-level physical DOF and associated powers. So here we have one sort of broadly (lower-level constraint-based) empirical answer to the first question, which Paolini Paoletti considers under the heading of my 'physicalistic solution,' and which he takes to be successful—in particular, "fully compatible with all versions of physicalism" (308), as far as it goes.

As Paolini Paoletti observes, however, my DOF-based account is only presented as a sufficient implementation of the schema for Weak emergence, and so won't work by way of a general answer to his questions; and indeed, as I clarify in my reply to Emery (this volume), other cases of Weak emergence are not clearly ones involving an elimination in DOF; so in these other cases a different answer to the first question might be operative. For a determinable-based implementation of Weak emergence of the sort that seems promising as applied to perceptual mental states, answering the first question would involve exploring why a given determinate has the determinables it does, which would require (among other things) attention to the determination dimensions of the determinate (see Funkhouser 2006). For a functional realization-based implementation of Weak emergence of the sort that seems promising as applied to artifactual features, answering the first question would involve exploring why certain functional roles are salient in our social economy. So here we have different sorts of answers to the first question, but so far as I can tell, these will also be broadly empirical, in depending on complex, broadly contingent facts. As such, even granting the general interest of the first question, I don't see any reason to think that it will have a single or unified answer, much less a single or unified metaphysical answer, of the sort that Paolini Paoletti appears to be seeking.

What about the second question, of what makes it the case that a given feature *S* is associated, with at least nomological necessity, with a given causal profile? Again, it seems to me that this question arises only for those who think that

features can be individuated independently of their powers via quiddities or primitive identities. Paolini Paoletti seems to take such a view for granted in his attempt to answer this question; hence, e.g., in considering whether the connection of a given causal profile to a given property is primitive, he says, "To make sense of this situation from an ontological standpoint, we may hold that there is some irreducible relation *R* that links *S* (and only *S*) with its causal role (and only with it)" (306). He rejects this primitivist answer, for reasons I'll discuss down the line, but the terms of the solution, like the question itself, presuppose that one may refer to a feature in some way independent of its powers—which those rejecting quiddities or the like will deny. Paolini Paoletti suggests that even someone not endorsing quiddities will have to answer a version of the second question. Hence he says of a non-quiddistic view on which properties are mere bundles of token powers that, "one would still need to explain why only certain bundles of token powers (and not others) seem to 'give rise to' or 'be legitimately describable as' token features" (308). But first, one may reject quiddities without embracing a bundle theory (which on the face of it reifies powers in a way that I would resist); one may rather simply think of properties in what I think of as metaphysically adverbial terms, as ways things are.

In any case, the (second) question as directed at the non-quidditist of whatever variety isn't the same as that directed at the quidditist. The question for the non-quidditist can be understood in two ways, depending on whether it is asked against a backdrop assumption of there being lower-level physical features associated with specific causal profiles. If so, then the question collapses into the first question—i.e., what makes it the case that a given subset of physical powers corresponds to a genuine feature? If not—if the question is more generally asking which collections of powers or "ways things are" correspond to genuine properties—the question collapses into the question "Which properties exist?" That's an interesting question, to which whole fields are devoted—but not one that any individual metaphysician has the burden of answering.

Putting my own inclinations aside, it seems to me that proponents of quidditistic accounts of properties typically suppose that the answer to Paolini Paoletti's second question is an empirical matter, even if they disagree over details. Hence, for example, Lewis (1986) supposes that what powers are associated with which (intrinsic, categorical) properties is a matter of the distribution of those properties in the Humean mosaic, which metaphysically contingent distribution determines the laws of nature at the world; and Armstrong (1983) supposes that what powers are associated with which universals is a matter of which metaphysically contingent relations of nomological necessitation hold at the world. Either way, answers to Paolini Paoletti's second question will be both diverse (depending on further commitments of the individual quidditist) and ultimately advert to certain contingent empirical facts.

I want to turn now to a thesis that shows up in Paolini Paoletti's assessment of a primitivist response to the first and second questions. Focusing on a specific instance of the first question, he says "suppose that we claim that it is a primitive and inexplicable fact of the matter that the proper subset made of $p1$, $p2$ and $p3$ (i.e., the causal role of *S*) is the only one that is associated with a higher-level token feature" (305). He goes on:

> [T]hat *R* holds between *S* and its causal role is an irreducible fact of the matter. Thus, it is a fundamental fact. Moreover, this fact constitutively includes a non-

> physical token feature such as *S*. Thus, there are fundamental facts with non-physical token features such as *S*. The constituents of fundamental facts are fundamental [following Sider 2011]. Therefore, non-physical token features such as *S* are fundamental. This conclusion may be hard to swallow for physicalists. (306)

Clearly it would be problematic for physicalists were a given higher-level feature, that was supposed to be Weakly emergent and so (though physically acceptable) not identical to any physical feature, turned out to be fundamental; for physicalists of any variety maintain that lower-level physical goings-on are the only fundamenta there are. Now, as above, I don't think there's any pressure here to embrace primitivism about the first or second questions, since each admits of diverse, broadly empirical answers. That said, Paolini Paoletti's remarks offer me an opportunity[38] to rail against Sider's purity principle (for short: 'Purity')—again, according to which the constituents of fundamental facts must themselves be fundamental.

In brief: I see no reason to accept Purity, and on the contrary good reason not to do so. For the fundamental goings-on—whether these be facts, states of affairs, or some other constituents of reality—are (if nothing else) required to serve as a suitable basis for all of reality, including any non-fundamenta there might be. Everyone agrees on this much, whatever the further details of their preferred account of what makes it the case that some goings-on at a world are fundamental at that world.[39] Hence it is that characterizations of fundamentality often start with the familiar "All God had to do" heuristic, according to which the fundamental goings-on are all God had to create in order to create the world as a whole. But if the world as a whole flows, one way or another, from just the fundamenta, then far from supposing that the fundamenta cannot contain or encode reference to non-fundamenta, it seems on the contrary that the fundamenta must contain or encode reference to non-fundamenta, for otherwise it is opaque how they could bring the non-fundamenta in their wake. Hence Purity is false. A better characterization of fundamental facts, it seems to me, is one according to which a fundamental fact must contain at least one fundamental entity or feature as a constituent; but that's compatible with fundamental facts' containing non-fundamenta as well. In any case, given that Purity is (to my mind: clearly) false, Paolini Paoletti's rejection of primitivist answers to his questions will have to rely on considerations (e.g., parsimony concerns of the sort he discusses) other than their leading to a supposed violation of Purity.

I now want to move on to Paolini Paoletti's preferred essentialist approach to his two questions. He maintains:

[38] Or another opportunity: see Wilson 2018 for an initial salvo.

[39] Among the usual suspects here are independence-based accounts (what makes it the case that some goings-on are fundamental is that they are independent of all else; see Schaffer 2009, Bennett 2017), complete minimal-basis accounts (what makes it the case that some goings-on are fundamental is that they are part of a minimal collection of goings-on which serve as a basis for all else; see Tahko 2018), and primitivist accounts (what makes it the case that some goings-on are fundamental is a primitive matter, not metaphysically analyzable in any other terms—a view which is, by the way, compatible with it being necessary or even essential to the fundamenta at a world that they enter into a basis for all else at the world; see Fine 2001, Wilson 2014 and forthcoming).

> [T]he best way to answer questions (1) and (2) consists in embracing something akin to 'grounding categoricalism,' i.e., the doctrine according to which the causal roles of categorical properties are somehow grounded on those very properties (see, among others, Tugby 2012, 2020, 2022, Yates 2018, Kimpton-Nye 2021, Paoletti 2021). In Paoletti (2021), I have defended the following form of grounding categoricalism: by virtue of its own essence, the causal role C of a categorical property P (i) is the causative role of P, so that it essentially depends (also) on P, (ii) it depends for its origins on P (i.e., it starts to exist as a causal role thanks to P or thanks to the instantiation of P) and (iii) it depends for its continuing to exist (also) on P (i.e., it continues to exist also or only thanks to P or to the instantiation of P). This entails that, as a matter of necessity, the existence of C implies the existence of P: necessarily, C cannot exist without P. And it also entails that, as a matter of necessity, C is the causal role of P and of no other property distinct from P. (308–309)

Here by the "essence" of an entity, Paolini Paoletti means "what that entity nonderivatively is (or could be) in all possible circumstances" (309).

In what follows I'll register certain concerns about an essence-based approach to the questions at issue, and with Paolini Paoletti's claim that such an approach has substantive implications for our understanding of physicalism, and more specifically, of Weak emergence.

First, Paolini Paoletti claims that grounding categoricalism provides attractive answers to the questions he has posed, but I don't see that this is so. Taking properties to be essentially such as to have or be otherwise associated with certain causal roles, which as it happens are comprised of a specific proper subset of lower-level physical powers (per the schema for Weak emergence), certainly provides a mechanism whereby a property and its causal profile go hand in hand, but it doesn't illuminate why (as regards the first question) only certain subsets of lower-level powers are associated with higher-level Weakly emergent properties, or why (as regards the second) a given property is essentially such to have *these* powers, either as a matter of nomological or metaphysical necessity. Again, I'm inclined to think that these questions admit of empirical answers, but Paolini Paoletti seemed to want more—namely, some metaphysical account of why certain collections of lower-level powers, but not others, are associated with some or other feature (the first question), and moreover with a specific feature (the second question). I don't see how grounding categorialism provides such an account, since that causal profiles are essentially tied to properties doesn't tell you what causal profiles and associated properties there are. Rather, grounding categoricalism introduces a slew of new questions, including: which essences are there? and why is a given essence associated with this causal role in this world (worse: at this time) and that causal role in that world (that time)?

If the answers to these questions turn out to be "it's primitive", then Paolini Paoletti's (remaining) concerns with primitivist answers to his question attach also to his account. Now, Paolini Paoletti claims that with his essence-based solution, "we avoid introducing primitive and sui generis connections" (310) between features and causal profiles, but to my mind an appeal to causal profiles as "grounded" in essences just pushes, and indeed multiplies, the primitivist bump(s) in the rug. Paolini Paoletti asserts that the grounding connections are not primitive, since "internal", but even granting that internal relations do not introduce primitive posits, the claim that the relation between essences and causal profiles

is internal doesn't establish this much; since no handle on the relation at issue has been provided sufficient unto showing that the relation is in fact internal.[40]

These considerations provide, in my view, good reason to stick with the usual array of empirical considerations offered by scientists and philosophers as motivating there being these special science features and associated powers/laws, and not others, which methodological strategy provides a generally explanatory and comparatively parsimonious basis for answering Paolini Paoletti's questions (to the extent that one feels pressure to do so, as a consequence of one's independent commitments---e.g., to a quidditistic conception of properties). Here it is also worth noting that one can deny Paolini Paoletti's claim that quiddities are motivated as answering his questions, since as previously discussed, there are available broadly empirical answers to the first question, and the second question doesn't arise unless one posits something like quiddities—in which case a purported need to answer his second question doesn't provide independent reason to posit quiddities.

Second and finally, even if it turns out that properties and their powers are essentially mutually dependent, I don't see that there is a deep problem for physicalism here. Physicalism is the view that all broadly scientific goings-on are "nothing over and above" lower-level physical goings-on, in the way that reductive versions of physicalism (appealing to identity) or non-reductive versions of physicalism (appealing to functional or other forms of realization, the key features of which are encoded in the schema for Weak emergence) aim to capture. It isn't any part of the physicalist project to maintain that mathematical or metaphysical features—e.g., the property of being prime, the relation between a universal and its instantiation, or (if such there be) the relation between a feature and its causal profile—are in any way nothing over and above or completely dependent on lower-level physical goings-on. So even if one is inclined to follow Paolini Paoletti in taking an essence-based approach to the questions he has raised, this in itself poses no tension with physicalism, or so it seems to me.

References

Albert, D., 2013. Wave function realism. *In*: A. Ney and D. Albert, eds. *The wave function: Essays on the metaphysics of quantum mechanics*. New York: Oxford University Press, 52–57.

Armstrong, D.M., 1978. *Universals and scientific realism, Vol i: Nominalism and Realism*. Cambridge: Cambridge University Press.

Armstrong, D.M., 1983. *What is a law of nature?* Cambridge: Cambridge University Press.

Baker, D.J., 2020. Knox's inertial spacetime functionalism. *Synthese*, 199 (S2), 277–298. https://doi.org/10.1007/s11229-020-02598-z.

[40] In addition, Paolini Paoletti appeals to diverse, but unspecified, dependence relations in order to accommodate a purported "circle of dependence" between features and profiles, in which case his approach appears to involve (in addition to its being primitive what essences of features there are, and primitive that certain causal profiles depend on such essences) a third primitive component, tracking that the features at issue depend on the associated profiles.

Baron, S. 2019. The curious case of spacetime emergence. *Philosophical studies*, 177, 2207–2226. https://doi.org/10.1007/s11098-019-01306-z.

Baron, S. and Le Bihan, B., 2022. Composing spacetime. *Journal of philosophy*, 119, 33–54. https://doi.org/10.5840/jphil202211912.

Baron, S. and Miller, K., 2014. Causation in a timeless world. *Synthese*, 191, 2867–2886. https://doi.org/10.1007/s11229-014-0427-0.

Baysan, U., 2016. An argument for power inheritance. *Philosophical Quarterly*, 66, 383–390.

Baysan, U. and Wilson, J.M., 2017. Must strong emergence collapse?. *Philosophica*, 91, 49–104.

Bellazzi, F., Biochemical functions. *Forthcoming. British journal for philosophy of science.*

Bennett, K., 2003. Why the exclusion problem seems intractable and how, just maybe, to tract it. *Noûs*, 37, 471–497.

Bennett, K., 2008. Exclusion again. *In*: J. Hohwy and J. Kallestrup, eds. *Being reduced: New essays on reduction, explanation, and causation*. Oxford: Oxford University Press, 280–307.

Bennett, K., 2017. *Making things up*. Oxford: Oxford University Press.

Bohn, E., 2009. Composition as identity: A study in ontology and philosophical logic. Thesis (PhD). University of Massachusetts.

Calosi, C., 2016a. Composition, identity, and emergence. *Logic and Logical Philosophy*, 25, 429–443. https://doi.org/10.12775/llp.2016.010.

Calosi, C., 2016b. Composition is identity and mereological nihilism. *Philosophical Quarterly*, 66, 219–235. https://doi.org/10.1093/pq/pqv109.

Calosi, C. and Giordani, A., Atoms, combs, syllables and organisms [unpublished manuscript].

Calosi, C., Universalism and extensionalism revisited. [unpublished manuscript].

Calosi, C. and Wilson, J.M., 2018. Quantum metaphysical indeterminacy. *Philosophical Studies*, 176, 2599–2627.

Cameron, R., 2007. The contingency of composition. *Philosophical Studies*, 136, 99–121.

Canavotto, I. and Giordani, A., 2020. An extensional mereology for structured entities. *Erkenntnis*, 87, 2343–2373. https://doi.org/10.1007/s10670-020-00305-5.

Carroll, S.M., 2021. Consciousness and the laws of physics. *Journal of Consciousness Studies*, 28, 16–31.

Chalmers, D.J., 2021. Finding space in a nonspatial world. *In*: C. Wüthrich, B. Le Bihan, and N. Huggett, eds. *Philosophy beyond spacetime*. Oxford: Oxford University Press.

Clapp, L., 2001. Disjunctive properties: Multiple realizations. *Journal of Philosophy*, 98, 111–136.

Cornell, D.M., 2017. Mereological nihilism and the problem of emergence. *American Philosophical Quarterly*, 54, 77-87.

Cotnoir, A. and Varzi, A., 2021. *Mereology*. Oxford: Oxford University Press.

Crowther, K., 2018. Inter-theory relations in quantum gravity: Correspondence, reduction and emergence. *Studies in History and Philosophy of Modern Physics*, 63, 74–85. https://doi.org/10.1016/j.shpsb.2017.12.002.

Crowther, K., 2022. Spacetime emergence: Collapsing the distinction between content and context? *In*: S. Wuppuluri and I. Stewart, eds. *From electrons to elephants and elections: Saga of content and context*. Cham, Switzerland: Springer, 379–402.

Fine, K., 1975. Critical notice of Lewis, *Counterfactuals*. *Mind*, 84, 451–58.

Fine, K., 2001. The question of realism. *Philosophers' imprint*, 1, 1–30.

Fine, K., 2010. Towards a theory of part. *Journal of philosophy*, 107, 559–589.

Fleming, G., 1987. Hyperplane dependent quantized fields and Lorentz invariance. *In*: H. Brown and R. Harre, eds. *Philosophical Foundations of Quantum Field Theory*, Oxford: Clarendon Press, 93–115.

Funkhouser, E., 2006. The determinable-determinate relation. *Noûs*, 40, 548–69.

Gillett, C., 2016. *Reduction and emergence in science and philosophy*. Cambridge: Cambridge University Press.

Guay, A. and Sartenaer, O., 2016. A new look at emergence. Or when after is different. *European journal for philosophy of science*, 6, 297–322. https://doi.org/10.1007/s13194-016-0140-6.

Heil, J. 2003. Levels of Reality. *Ratio*, 16, 205–221.

Huggett, N., 2021. Spacetime "emergence". *In*: E. Knox and A. Wilson, eds. *The Routledge Companion to Philosophy of Physics*. New York: Routledge, 374–385.

Humphreys, P., 1997. How properties emerge. *Philosophy of science*, 64, 1–17.

Kim, J., 1989. The myth of nonreductive materialism. *Proceedings and addresses of the American Philosophical Association*, 63, 31–47.

Kim, J., 1998. *Mind in a physical world*. Cambridge: MIT Press.

Kim, J., 2006. Emergence: Core ideas and issues. *Synthese*, 151, 547–59.

Kimpton-Nye, S., 2021. Reconsidering the dispositional essentialist canon. *Philosophical studies*, 178, 3421–3441. https://doi.org/10.1007/s11098-021-01607-2.

Koslicki, K., 2008. *The structure of objects*. Oxford: Oxford University Press.

Lam, V. and Wüthrich, C., 2018. Spacetime is as spacetime does. *Studies in history and philosophy of science Part B: studies inhHistory andphilosophy of modern Physics*, 64, 39–51. https://doi.org/10.1016/j.shpsb.2018.04.003.

Lamb, M., 2015. Characteristics of non-reductive explanations in complex dynamical systems research. Thesis (PhD). University of Cincinnati.

Landry, R. and Moffat, J., forthcoming. Nonlocal quantum field theory and quantum entanglement. *European physical journal*, https://doi.org/10.48550/arXiv.2309.06576.

Lewis, D., 1986. *On the plurality of worlds*. London: Blackwell.

McDaniel, K., 2008. Against composition as identity. *Analysis*, 68 (2), 128–133. https://doi.org/10.1093/analys/68.2.128.

McLaughlin, B., 2006. Is role-functionalism committed to epiphenomenalism? *Journal of consciousness studies*, 13, 39–66.

McLaughlin, B. and Bennett, K., 2018. Supervenience. *In*: E.N. Zalta, ed. *The Stanford encyclopedia of philosophy*, Winter 2018. https://plato.stanford.edu/archives/win2018/entries/supervenience/; Metaphysics Research Lab, Stanford University.

Melnyk, A., 2006. Realization-based formulations of physicalism. *Philosophical Studies*, 131, 127–155.

Miller, M., forthcoming. Worldly imprecision. *Philosophical studies*, [unpublished manuscript].

Moffat, J.W., 1990. Finite nonlocal Gauge field theory. *Phys. Rev. D*, 41, 1177–1184. https://doi.org/10.1103/PhysRevD.41.1177.

Morrison, M., 2012. Emergent physics and micro-ontology'. *Philosophy of science*, 79, 141–166.

Namsrai, K., 1986. *Nonlocal quantum field theory and stochastic quantum mechanics*. Dordrecht: Springer Netherlands.

Ney, A., 2021. *The world in the wave function*. Oxford: Oxford University Press.

Paoletti Paolini, M., 2021. Emergence and structural properties. *Synthese*, 198, 8755–8778. https://doi.org/10.1007/s11229-020-02599-y.

Rovelli, C., 2007. *Quantum gravity*. Cambridge: Cambridge University Press.

Santos, G., Vallejos, G., and Vecchi, D., 2020. A relational-constructionist account of protein macrostructure and function. *Foundations of Chemistry*, 22, 363–382.

Sattig, T., 2015. *The double lives of objects: an essay in the metaphysics of the ordinary world*. Oxford: Oxford University Press.

Saunders, S.W., 1992. Locality, complex numbers, and relativistic quantum theory. *PSA: Proceedings of the biennial meeting of the Philosophy of Science Association*, 365–380.

Schaffer, J., 2009. On what grounds what. *In*: D. Manley, D. Chalmers, and R. Wasserman, eds. *Metametaphysics: New essays on the foundations of ontology*, 347–83. Oxford: Oxford University Press.

Schaffer, J., 2010. Monism: The priority of the whole. *Philosophical review*, 119, 31–76.

Shoemaker, S., 2000/2001. Realization and mental causation. *In*: *Proceedings of the 20th World Congress in Philosophy*, Cambridge: Philosophy Documentation Center, 23–33.

Sider, T., 2011. *Writing the book of the world*. Oxford: Oxford University Press.

Simons, P.M., 1987. *Parts: a study in ontology*. Oxford: Oxford University Press.

Tahko, T.E., 2018. Fundamentality and ontological minimality. *In*: R. Bliss and G. Priest, eds. *Reality and its structure*. Oxford: Oxford University Press, 237–53.

Tallant, J., 2019. Causation in a timeless world? *Inquiry: an interdisciplinary journal of philosophy*, 62, 309–325. https://doi.org/10.1080/0020174x.2018.1446051.

Taylor, E., 2015. Collapsing emergence. *Philosophical Quarterly*, 65, 732–753.

Tomboulis, E.T., 2015. Nonlocal and quasilocal field theories. *Phys. Rev. D.*, https://doi.org/10.1103/PhysRevD.92.125037

Tugby, M., 2012. Rescuing dispositionalism from the ultimate problem: Reply to Barker and Smart. *Analysis*, 72, 723–731. https://doi.org/10.1093/analys/ans112.

Tugby, M., 2020. Grounding theories of powers. *Synthese*, 198 (12), 11187–11216. https://doi.org/10.1007/s11229-020-02781-2.

Tugby, M., 2022. Dispositional realism without dispositional essences. *Synthese*, 200 (3), 1–27. https://doi.org/10.1007/s11229-022-03554-9.

Weinberg, S., 1997. What is quantum field theory, and what did we think it is? https://arxiv.org/abs/hep-th/9702027.

Wilson, J.M., forthcoming. The fundamentality first approach to metaphysical structure. *Australasian Journal of Philosophy*.

Wilson, J.M., under contract. *Fundamentality and metaphysical dependence*. Oxford: Oxford University Press.

Wilson, J.M., 2002. Causal powers, forces, and superdupervenience. *Grazer Philosophische-Studien* 63, 53–78.

Wilson, J.M., 2005. Supervenience-based formulations of physicalism. *Noûs*, 39, 426–459.

Wilson, J.M., 2006. On characterizing the physical. *Philosophical Studies*, 131, 61–99.

Wilson, J.M., 2010. Non-reductive physicalism and degrees of freedom. *British journal for the philosophy of science*, 61, 279–311.

Wilson, J.M., 2013. A determinable-based account of metaphysical indeterminacy. *Inquiry*, 56, 359–385.

Wilson, J.M., 2014. No work for a theory of grounding. *Inquiry*, 57, 1–45.

Wilson, J.M., 2015. Metaphysical emergence: Weak and strong. *In*: T. Bigaj and C. Wüthrich, eds. *Metaphysical emergence in contemporary physics; Poznan Studies in the Philosophy of the Sciences and the Humanities*. Amsterdam/New York: Brill, 251–306.

Wilson, J.M., 2018. Grounding-based formulations of physicalism. *Topoi*, 37, 495–512.

Wilson, J.M., 2021. *Metaphysical emergence*. Oxford: Oxford University Press.

Yates, D., 2018. Inverse functionalism and the individuation of powers. *Synthese*, 195: 4525–4550. https://doi.org/10.1007/s11229-017-1417-9.

Yukawa, H., 1950a. Quantum theory of non-local fields. Part I. Free fields. *Phys. Rev.*, 77 (January): 219–226. https://doi.org/10.1103/PhysRev.77.219.

Yukawa, H., 1950b. Quantum theory of non-local fields. Part II. Irreducible fields and their interaction'. *Phys. Rev.*, 80 (December): 1047–1052. https://doi.org/10.1103/PhysRev.80.1047.

Advisory Board

Argumenta 10, 1 (2024)
Article Discussion

On Eric Olson's *Parfit's Metaphysics and What Matters in Survival*

The Journal of the Italian Society for Analytic Philosophy

The Fission Argument for the Unimportance of Identity Cannot Be Correct

Harold Noonan

University of Nottingham

Abstract

Eric Olson has made an important addition to the discussion started by Parfit of the argument from the possibility of fission to the unimportance of personal identity. Olson's discussion is challenging. I want, more briefly, to highlight what is the most important consequence of it. This is that it is metaphysically impossible, impossible in the strongest sense, that any version of Parfit's argument from fission can yield his conclusion. Olson argues specifically that this is impossible if what he calls a 'capacious ontology' is assumed. I argue that it is a consequence of Parfit's reasoning that this is so even without the assumption of a capacious ontology.

Keywords: Fission, Identity, What matters, Parfit, Olson.

1. Introduction

Sometimes, occasionally, something new and important is added to a long-running philosophical debate. Eric Olson (2019) has made just such an addition to the discussion started by Parfit (1971) of the argument from the possibility of fission to the unimportance of personal identity.

But Olson's discussion is long, complex and challenging. I want, more briefly, to highlight what is the most important consequence of it—a consequence he does not actually draw out. This is that it is metaphysically impossible, impossible in the strongest sense, that any version of Parfit's argument from fission can yield his conclusion.

The reason for this is that any version of the argument:

(a) has to appeal to the difference between two situations (i) one in which a single brain hemisphere is transplanted (with consequent transfer of psychology) and the other destroyed; and (ii) one in which two hemispheres of a brain are transplanted into distinct skulls (with consequent transfer of psychology)—the fission case, and

Argumenta 10, 1 (2024): 369—373
ISSN 2465-2334

DOI 10.14275/2465-2334/202419.noo
First published: 30 November 2024

(b) must assume (premise 1) that whilst identity is preserved in the first case there is no identity in the second, though (premise 2) everything *that matters* is preserved in the second case as in the first.

Of course, Parfit needs to justify the second premise, that everything in the single hemisphere transplant that matters is preserved in the fission case, as well as the first. He could just insist that only psychological continuity matters. But, as Olson notes, he does not want to do that. I think the best response Parfit has, at this point, is to appeal to our intuition when we think about the possibility *first-personally*: it seems that given a choice between a single hemisphere transplant and fission there is nothing to make it reasonable to choose the former. This seems a good reply (Shoemaker 1984: 119) to a demand for a justification of Parfit's second premise. So, the crux, which Olson is mainly concerned with, is whether the first premise, that identity is not preserved in the fission case though it is in the single hemisphere transplant, can be defended. My claim in what follows is that thinking through Olson's criticism we can see that it cannot be, even if a capacious ontology (as Olson calls it) is not assumed. Note that throughout when I say 'identity' I mean personal identity. A capacious ontologist might say that identity is preserved in fission, but not personal identity. That is, he might say that there is something, one and the same thing, present before the fission and afterwards, but that there is no person present before and after the fission. But that would be implausible, no one does say this and Olson sensibly ignores the possibility.

2. Why the Fission Argument Fails

I now go on to explain all this.

The focus of Olson's argument is, in fact, what he calls "the capacious ontology"—the ontology of a philosopher who thinks that every matter-filled region of space-time contains a material thing which exactly matches its boundaries (Olson 2019: 30). An example of this is the four-dimensional ontology of Lewis (1976) and Quine (1960), in which any shorter-live thing coincident throughout its existence with a longer-lived one is a temporal part of the latter. But Olson uses the term more generally. He makes a convincing case that Parfit accepts the capacious ontology, though without ever arguing for it, but he notes that Parfit is silent on the Lewis–Quine ontology of temporal parts. He also draws attention to Shoemaker (Shoemaker 1984), who also seems committed to a capacious ontology, but vociferously rejects the Lewis–Quine ontology of temporal parts.

Olson then goes on to argue that the defender of the capacious ontology cannot employ Parfit's fission argument to establish the unimportance of identity, in the sense championed by Parfit (so, of course, by assuming the capacious ontology, Parfit has undermined his own argument).

His argument for this claim depends on a careful distinction between what Parfit is arguing for and the (uninteresting) claim he is not arguing for.

Parfit's actual claim Olson expresses as follows:

Strong Unimportance of Identity: What matters in survival is never identity, but only some sort of psychological continuity. Whenever someone has a special prudential reason to care about someone's future, it's not because anyone survives, but only because that future person is psychologically continuous with her.

He distinguishes this from:

> *Weak Unimportance of Identity:* What matters in survival is always identity. Psychological *continuity* is practically important because it secures identity. Whenever someone has a special prudential reason to care about someone's future welfare, it is either because she is the person and thus survives or because someone coincident with her survives. But it is always because someone survives.

According to the weak claim, psychological continuity is not what ultimately matters. What does is identity. But what matters to a person about to fission is not that he, the very same person, exists after the fission. What matters to him is that there is *a* person coincident with him before the fission who exists after the fission, and so persists as one and the same identical thing through the fission. The previously coincident person may or may not be psychologically continuous with himself as he was earlier. This is not important to a person about to fission. What matters to such a person is only that someone coincident with him before exists after. This is not Parfit's claim. It is no one's claim. As Olson puts it, "strong unimportance of identity is a radical challenge to our ordinary thinking about value. The weaker claim is much less interesting. The most likely reaction to it is bafflement. It is unlikely to change our thinking".

But, as Olson explains, Parfit's actual thesis about the unimportance of identity cannot be supported by appeal to the fission argument if the capacious ontology is assumed. According to the capacious ontologist, it is metaphysically necessary that in a case of fission there is survival. So, a thought experiment separating the two factors that might ground what matters—the presence of identity on the one hand (as in the single hemisphere transplant case) and the presence of mere psychological continuity (as in the fission case)—is metaphysically impossible. Granted that nothing is present in the former that matters which is lacking in the latter, we cannot infer that identity is not something that matters since, according to the capacious ontology, there is identity in the latter too.

However, it is obvious that one can think that there is identity in the fission case, i.e., that one can think that someone who exists after the fission in that situation existed before, without endorsing the capacious ontology. One needs not believe that every filled space-time region contains an object which exactly fills it to believe this.

A plausible line of thought that yields the conclusion that if there is someone in the single brain-hemisphere transplant case who survives the transplant then someone who is present after fission in the fission case was there before the fission, goes as follows. First thought. A person cannot go out of existence unless something happens to him. But in the relevant sense something happens to a person only if he undergoes a non-relational change. Nothing thereby happened to Socrates when Theaetetus grew taller than him. Nothing thereby happens to a man when his long-separated wife dies—though he becomes a widower. Nothing happened to the Merry Men when evil Prince John had a sudden change of heart and pardoned them, and the next day, returning to his old ways, reversed the pardon—though the number of outlaws in Sherwood Forest went from 100 to 0 then back up to 100. That a person cannot go out of existence merely because of a relational change is a fact, a necessary fact, about persons. It is not a fact about things generally, it is not, for example, a fact about holes or indentations more

generally.[1] But it is a fact about lots of things other than people: dogs and trees and ships and computers and ashtrays. People are like dogs and trees etc., not like holes. The second thought is simply that if a person (or dog or tree etc.) does not go out of existence at some time in one situation, it cannot go out of existence at that time in any second situation in which nothing happens to it that does not happen to it in the first. This is just part of what it is to be a person or a dog etc.[2]

If this line of thought is accepted, then—even if the capacious ontology is rejected—it must be acknowledged that, in the fission case, there is necessarily someone who exists after the fission who existed before it, if there is a person with such a lifespan in the single hemisphere transplant case. So, we again secure, by Olson's reasoning, that it is impossible for any version of the fission argument to secure Parfit's conclusion, since no thought experiment separating the two factors that might ground what matters is metaphysically possible.

Of course, someone might resist the line of thought just described and insist that a mere relational change can bring a person's existence to an end—persons *are* like holes (he then has to choose whether to say the same of dogs etc., or to accept that persons are unlike dogs). But, apart from a defender of the capacious ontology, who thinks things are constantly going out of existence without any non-relational change happening to them, who would want to say this? This is the line that must be taken by those who endorse a non-branching, no-rival or best candidate, account of personal identity. But those who endorse this are typically capacious ontologists—the most prominent defenders of such an account of personal identity being Parfit himself, and Shoemaker.

I conclude that reflection on Olson's argument should lead to the position that Parfit's fission argument necessarily fails to yield its conclusion. Maybe some other argument will do the job. But Parfit's own additional argument, the argument from below, is much contested, and specifically, as Olson shows, requires the assumption of the capacious ontology and is thus inconsistent with the strong independence of what matters from identity that Parfit believes in. And I know of no other. So, I think that where we are at present is that there is no good reason to accept Parfit's famous claim that identity does not matter.

References

Noonan, H., 2019. *Personal identity*. New York: Routledge.

[1] One can bring a hole into existence by digging. So, a way to cause a hole to cease to exist is to fill it up. But one can also cause it to cease to exist by lowering the ground around it.

[2] This line of thought, of course, derives from Williams (1956–7) and is employed by him in Williams (1970). Noonan (2019: 140) attempts a formulation of the basic principle (as applicable to persons) that can be put as follows in the terminology of this paper: "If two events are parts of the history of a person in one situation they must also be parts of the history of a person in any second situation in which they, and all the events which are part of the history of the person in the first situation, remain present and differ in no non-relational way from the way they are in the first situation". J.R.G. Williams (2013) gives a better formulation which can be put as follows: "If a spatio-temporal region is exactly occupied by a person any duplicate (intrinsically identical) region is exactly occupied by a person or is part of a region exactly occupied by a person".

Olson, E., 2019. Parfit's metaphysics and what matters in survival. *Argumenta*, 5(2), 21–39.

Parfit, D., 1971. Personal identity. *Philosophical review*, 80(1), 3–27.

Quine, W.V.O., 1960. *Word and Object*. Cambridge, MA: MIT Press.

Shoemaker, S., 1984. Personal identity: a materialist's account. *In*: S. Shoemaker and R. Swinburne, eds. *Personal identity*. Oxford: Basil Blackwell.

Williams, B.A.O., 1956–7. Personal identity and individuation. *Proceedings of the Aristotelian society*, 70, 229–52.

Williams, B.A.O., 1970. The self and the future. *Philosophical review*, 79(2), 169–80.

Williams, R., 2013. Part-intrinsicality. *Noûs*, 47(3), 431–52.

Is What Matters Present in a Fission Scenario? A Conventionalist Response to Noonan

Alfonso Muñoz-Corcuera

Complutense University of Madrid

Abstract

In a recent paper, Olson (2019) returned to Parfit's argument from the possibility of fission to the unimportance of identity to claim that it is inconsistent with Parfit's ontological commitments. Picking up Olson's claim, Noonan (2024) argues that one consequence of this is that Parfit's argument necessarily fails to yield its conclusion. Here I show that Noonan's ontological stance is similar to Parfit's in one significant sense, thus diminishing the scope of his claim. As a result, I hold that if we want to defend that personal identity is what matters, we should reject that what matters may be present in a fission scenario.

Keywords: Personal identity, Fission cases, Conventionalism, Derek Parfit, Eric Olson, Harold Noonan.

1. Introduction

Among the many contributions that Derek Parfit made to the debate on personal identity, his argument from the possibility of fission to the unimportance of identity stands out for its relevance (Parfit 1984: 253–266). Recently, Eric Olson revisited this argument, claiming it does not work if one accepts Parfit's ontology (Olson 2019). In this regard, Olson argues that Parfit's argument about the unimportance of identity lacks support.

In response, Harold Noonan argues that Olson's point leads to an important consequence: any version of the fission argument fails to yield the conclusion that personal identity does not matter, regardless of whether one accepts Parfit's ontology (Noonan 2024). Noonan contends that it is metaphysically impossible to have a fission case where the persons existing after the fission have what matters in survival but they are not the same as someone existing before the fission. Thus, he concludes there is no good reason to accept that personal identity does not matter.

Noonan's discussion is illuminating and thought-provoking. However, his ontological stance aligns more closely with Parfit's than he may realise. As a result,

Argumenta 10, 1 (2024): 375—382
ISSN 2465-2334

DOI 10.14275/2465-2334/202419.mun
First published: 30 November 2024

the scope of his paper is more limited than he states. It only shows that the fission argument fails if one shares one ontological commitment with Parfit. In this regard, I argue that if we want to defend that personal identity is what matters, we should deny that what matters may be present in a fission scenario. I will aim to show this succinctly.

2. Parfit's Argument

We can begin by considering Parfit's argument, which can be outlined as follows:

(Premise 1). There are fission cases where a person *a* exists at t_1 and two persons *b* and *c* exist at a later time t_2, such that both *b* and *c* would be the same person as a if the other one did not exist but, since *b* and *c* are not the same person, per the transitivity of identity, both cannot be the same person as *a*.

(Premise 2). In such fission cases, both *b* and *c* have what matters in survival regardless of whether they are the same person as *a*.

Therefore,

(Conclusion). Personal identity is not what matters in survival.

3. Noonan's Argument

Olson argues that Parfit cannot endorse his argument because Premise 1 contradicts his own ontological commitments (Olson 2019: 35–38). The details of Olson's criticism are complex, but we do not need them now.

What matters for our discussion is Noonan's interpretation of Olson's criticism. Noonan claims that Parfit's argument fails not only because Premise 1 is incompatible with his ontology, but because it violates two necessary facts about persons (Noonan 2024). These facts are related to what Noonan elsewhere calls "the only x and y principle" (Noonan 2019: 33). Here, I will refer to them as

The Non-Relationality Principle: A person can only cease to exist due to a non-relational change

and

The Comparative Principle: If a person would survive a given situation involving non-relational changes, they would survive in any other situation where the same non-relational changes occur.

If we accept these two principles, Premise 1 in Parfit's argument collapses. The existence of *c* cannot be the reason why *b* fails to be the same as *a*, and *vice versa*. Therefore, Noonan concludes that, in the absence of any convincing argument in its favour, "there is no good reason to accept Parfit's famous claim that identity does not matter" (Noonan 2024).

At first sight, Noonan's account seems compelling. His only explicit commitments are:

(1) the Non-Relationality Principle,

(2) the Comparative Principle,

(3) the claim that personal identity is what matters; and

(4) Premise 2 in Parfit's argument, which Noonan takes to be quite unproblematic.

However, although these four claims may seem intuitive, they place Noonan in a difficult position. First, they entail an awkward interpretation of fission scenarios that depletes all the intuitive appeal of his proposal. Second, they align Noonan more closely with Parfit's ontology than he likely intends, thus making Noonan's claim more modest than he believes.

Let us see what I mean.

4. The Main Ways to Account for Fission Cases

Let us start by examining why I think that Noonan's four explicit commitments lead to an awkward interpretation of fission cases. To do this, we first need to see that his commitments are incompatible with the main approaches used to explain these scenarios.

The main problem with fission cases lies in the transitivity of identity. If a person *a* has two continuers *b* and *c* such that both could be the same person as *a*, we have to face the fact that the continuers are obviously distinct from each other. But if they are distinct, they cannot both be the same person as *a*, as this would violate the transitivity of identity.

There are two main ways to deal with this problem. The first one consists in taking the possibility of fission as evidence that we are relying on a wrong account of our persistence. This is the approach that Williams championed (Williams 1957, 1970). As both *b* and *c* are distinct, and our criterion does not allow us to pick one as the right continuer, we must conclude that neither of them is the same person as *a*. Moreover, if neither of them is *a*, we should also conclude that even if they had been their sole continuer, they would have neither been *a* because the identity between two persons cannot depend on the existence of a third person. This is a consequence of the Non-Relationality and Comparative Principles above.

The second approach is based on rejecting the Non-Relationality and Comparative Principles. We could think that our criterion of personal identity is mostly right but incomplete. It has to be amended to resolve fission cases. In this regard, we could hold that both *b* and *c* would be *a* had the other one not existed, but since they both exist, none of them is *a*. Alternatively, we could say that even if *c* would have been *a* had *b* not existed, since *b* exists then they are *a*, as they are a better candidate than *c*. The former approach, known as the non-branching view, is favoured by Parfit among others. The latter approach, known as "the best candidate view", has its best-known supporter in Nozick (1981).

Noonan, however, cannot endorse either of these approaches without dropping one of his key commitments. Both views entail that either *b* or *c* (or both) would not be the same person as *a*. For Noonan, this means he would have to abandon either his belief (3) that personal identity is what matters or (4) that both *b* and *c* would have what matters. Since he explicitly endorses both claims, he cannot address fission scenarios in any of these ways.

5. Noonan's Way to Account for Fission Cases

There is a third way to approach fission scenarios that would let Noonan keep his four explicit commitments. One can hold that *b* and *c* are indeed distinct persons, but that they already existed before the fission. In this view, we must accept that at t_1 there were at least two persons who shared *a*'s body and mind. The fission scenario simply provided separate bodies and minds for each of them. Hence, we

can easily uphold both (3) that personal identity is what matters, and (4) that *b* and *c* have what matters. They have what matters because they have survived.

The acceptance of the possibility of two or more persons sharing one body and mind is what Noonan calls the Multiple Occupancy Thesis (Noonan 2019: 14–15). It is a position commonly associated with those who accept a four-dimensionalist view, like Lewis (1983). However, it can also be accepted by others who reject four-dimensionalism. For instance, Parfit probably did not endorse four-dimensionalism, but according to Olson, he is committed to the Multiple Occupancy Thesis (Olson 2019: 29). Similarly, it seems this thesis best characterises Noonan's position in his paper, as it is the only way he can consistently hold his four commitments.

However, while Noonan's four commitments may seem intuitively true, the Multiple Occupancy Thesis is quite the opposite. Few people would be willing to endorse a theory that entailed the Multiple Occupancy Thesis. In this regard, the intuitive appeal of Noonan's proposal appears to vanish.

Before we delve further into the problems of the Multiple Occupancy Thesis, there is a significant point I would like to make.

6. The Limited Scope of Noonan's Argument

At the end of section 3, I argued that Noonan's four explicit claims committed him to an ontological stance significantly similar to Parfit's. This similarity makes the scope of his paper more limited than he realizes. We are now in a position to see why.

In his paper, Olson claims that Parfit needs "*some* sort of capacious ontology" to hold that questions about personal identity are empty (Olson 2019: 28). Olson's discussion is quite detailed about what exactly Parfit needs in his ontological view to support his account. However, one of the most important pieces of that "sort of capacious ontology" is the Multiple Occupancy Thesis.

Olson shows that Parfit's argument ultimately conflicts with his ontological stance (Olson 2019: 35–38). If we accept the Multiple Occupancy Thesis, *b* and *c* must both exist before and after the fission. As a result, Parfit cannot claim that personal identity is not what matters because even if *b* and *c* have what matters, they are still the same person as someone who existed before the fission.

In this regard, we can see that Noonan's paper does not go much further than Olson's. He does not demonstrate that "any version of the fission argument" necessarily fails to prove that personal identity is not what matters. Rather, he only shows that the fission argument fails if we accept the Multiple Occupancy Thesis. Certainly, Parfit's capacious ontology is much more demanding than the Multiple Occupancy Thesis. Thus, Noonan's paper still has the value of extending Olson's conclusion. However, the reach of his claim is much more modest than he thinks, as it only applies to those who endorse the Multiple Occupancy Thesis.

Noonan might argue that we all should endorse the Multiple Occupancy Thesis. But what reasons could we have to do so?

7. Reasons to Endorse the Multiple Occupancy Thesis

The Multiple Occupancy Thesis is a highly contested claim. It is far from being an obvious truth about persons, and it is neither intuitive nor particularly plausible. In fact, it seems that there is only one reason why anyone would

endorse it: we would if it were entailed or necessitated by some other fact whose truth we wanted to preserve.

This seems to be the case with Parfit. As Olson convincingly shows, Parfit needs the Multiple Occupancy Thesis to sustain his claims about the emptiness of questions about personal identity (Olson 2019: 28–29). However, Parfit's claims are as contested and counterintuitive as the Multiple Occupancy Thesis. Thus, Noonan is not likely to find any additional support for his account in Parfit's view.

Four-dimensionalism may entail the Multiple Occupancy Thesis too (see Lewis 1983). However, I doubt that the decission to accept four-dimensionalism hinges on the debate on personal identity. It is likely to be the other way around. In this regard, Noonan may find some support for the Multiple Occupancy Thesis if he can convince us that we should endorse four-dimensionalism as a general ontological framework. In any case, if he would like to do so, I think the ball is in his court.

That leaves us with the only reason that Noonan gives to support the Multiple Occupancy Thesis: he presents it as if it were an entailment of the Non-Relationality and Comparative Principles. Furthermore, he argues that these principles are necessary facts about persons. Thus, Noonan claims, it is necessarily true that *b* and *c* already existed before fission. And that can only be true if the Multiple Occupancy Thesis is true as well.

However, the Multiple Occupancy Thesis does not follow from the Non-Relationality or the Comparative Principles. In fact, these claims are irrelevant to the issue at hand. We can see evidence of it in Williams' reasoning. When addressing fission scenarios, Williams relies on the Non-Relationality Principle to argue that any account of our persistence that entailed that *b* and *c* would be the same person as *a* had the other one not existed must be wrong, as any other conclusion would be absurd. Then he goes on and, relying on the Comparative Claim, concludes that any such account of our persistence must be wrong in non-fission scenarios too (Williams 1957: 239; 1970: 178). Williams avoids the conclusion that the Multiple Occupancy Thesis is necessary, despite endorsing both the Non-Relationality and Comparative Principles.

What really forces Noonan to accept the Multiple Occupancy Thesis are his other two explicit commitments: (3) that personal identity is what matters, and (4) that in fission scenarios both *b* and *c* have what matters. If these two claims are true, then *b* and *c* must be identical to someone existing before the fission. But since they are clearly distinct persons after the fission, they must have been distinct persons before the fission as well. And this can only be possible if one accepts the Multiple Occupancy Thesis. Whether one also holds the Non-Relationality or the Comparative Principles does not affect this conclusion.[1]

In sum, the acceptance of the Multiple Occupancy Thesis does not depend on whether we accept the Non-Relationality and Comparative Principles. Instead, it hinges on whether we think (3) that personal identity is what matters;

[1] This does not mean that accepting the Non-Relationality and Comparative Principles does not have consequences for fission cases. If Noonan denied them, he should claim that before the fission there were two persons sharing *a*'s body and mind, both of which survived. As he accepts it, then he must accept that before the fission there were three persons, one of them died and the other two survived. In any case, this does not affect Noonan's commitment to the Multiple Occupancy Thesis (see Noonan 2019: Ch. 12).

and (4) that both *b* and *c* have what matters. Either way, if one rejects the Multiple Occupancy thesis, Parfit's argument remains viable.

8. The Real Choice About Fission Cases

At this point, we may interpret fission cases as revealing a conflict between the claims (3) that personal identity is what matters and (4) that both *b* and *c* have what matters. One can endorse (3) and reject (4), as Williams does. Alternatively, we may side with Parfit and accept (4) while rejecting (3). Or we could stick with Noonan and accept both (3) and (4).

We should remember that the whole debate stems from Parfit's argument against (3). Thus, if we wish to preserve the intuition that personal identity is what matters, we are left with two choices. We may either accept or reject (4).

Noonan argues that we should accept (4). However, as we have seen, this commits him to hold the Multiple Occupancy Thesis too. As this cannot be seen as a desirable outcome, the rationale behind Noonan's acceptance of (4) should be especially compelling.

Unfortunately for Noonan, his justification falls short. He relies on Shoemaker and argues that, first-personally, "given a choice between a single hemisphere transplant and fission there is nothing to make it reasonable to choose the former" (Shoemaker 1984: 119–120). However, I can easily think of numerous reasons to choose the former.

From my first-person perspective, many things matter to me in my survival, and it's unclear whether a fission scenario could provide these to both of my continuers. For example, going on vacation with my wife; enjoying quality time with my kids; getting the satisfaction of being congratulated by my students after a challenging course; blending anonymously in the city to have some time for myself… Would both of my continuers be able to enjoy any of these things if I underwent fission? What would my family think of them? Would we all live together in the same house? And who, if anyone, would get to keep my current position at my university? Would they pay one of us, both, or would they resolve my contract and hire someone new? Would I be all over the news thus making it impossible to take a walk without being constantly asked to take a selfie with someone? I am not claiming that fission would outright prevent my continuers from having what matters to me, but I do argue that it is far from clear whether it would.[2]

The debate here is complex and would merit much more space than I have left. As I have noted elsewhere, the notion of "what matters" lacks sufficient philosophical precision. Thus, it is difficult to know whether what matters would be present in any given scenario (see Muñoz-Corcuera 2023). Nonetheless, I think that what I have said at least supports the following conclusion: Claim (4) is not obviously true and has the highly undesirable consequence of committing us to the Multiple Occupancy Thesis. Thus, if we want to claim that personal identity

[2] One could think that all these difficulties could be imagined away if, for example, one of my continuers were transported to Australia, while the other one remained at home. However, while this could grant what matters to my stay-at-home continuer, my Australian counterpart would not be capable of enjoying any of those things. And what I am trying to dispute here is that a fission scenario could grant what matters to both of my continuers at the same time. I am thankful to one anonymous referee for making me think about this.

is what matters in survival, we would be better off siding with Willaims and asserting that, in fission scenarios, *b* and *c* would not have what matters.

9. A Final Thought

Even though the Non-Relationality and Comparative Principles do not affect the Multiple Occupancy Thesis, they are still substantive claims that can make a difference in debates about the persistence of persons. One such difference lies in how we could defend that personal identity is what matters.

Noonan would probably want to hold both the Non-Relationality and Comparative Principles, as he takes them to be necessary facts about persons. In this regard, he would probably argue that if personal identity were to depend on relational properties, it would lose its significance. If I could cease to exist simply because someone else exists, then personal identity could not be considered of much importance.

On the contrary, I think that rejecting the Non-Relationality and Comparative Principles is perfectly compatible with the view that personal identity is what matters. Conventionalist accounts, which give social properties a significant role in personal identity, are a good example (see e.g. Braddon-Mitchell & Miller 2004; Wagner 2019, and Muñoz-Corcuera 2021). For instance, according to Schechtman (2014), a person is an entity defined by a cluster of biological, psychological and social properties. As such, persons persist over time as long as enough of these properties still hold together.

In fission cases usually biological and psychological continuity are disrupted significantly. A person may survive such disruption if their social properties remain unaltered. However, in fission cases, social properties would likely be affected as well. Hence, the mere existence of two continuers instead of one, combined with the diminished degree of biological and/or psychological continuity, may cause a person to cease to exist (Schechtman 2014: 159–166).

This view rejects both the Non-Relationality and Comparative Principles. However, it does not entail that personal identity does not matter. It only would if it entailed a rejection of another principle which, like the Non-Relationality and Comparative Principles, traces back to Williams' work on personal identity:

> *The Non-Arbitrariness Principle*: An arbitrary convention cannot cause a person to cease to exist.

As Williams pointed out, if personal identity depended on arbitrary conventions, it could not bear the ethical significance that we attribute to it (Williams 1970: 178–179). Williams thought that the mere existence of a third person was a trivial fact that could only affect personal identity if we relied on such arbitrary conventions. The non-branching and closest continuer views seemed to justify his opinion.

However, conventionalism does not treat the mere existence of a third person as a trivial fact. Instead, it views this as a significant factor, because it disrupts the continuity required for the post-fission persons to live the same life as the original person (Schechtman 2014: 166). In this regard, conventionalism does not make personal identity depend on arbitrary conventions, but on non-arbitrary ones. And a non-arbitrary convention can carry ethical significance and support the claim that personal identity is what matters (for a discussion, see Muñoz-Corcuera 2021: 732–737).

Again, there is much room for debate here. However, I do not have space to fully address this issue which surely merits further thought.

10. Conclusion

Noonan claimed that Parfit's argument from the possibility of fission to the unimportance of identity necessarily failed to yield its conclusion, as its first premise was false. However, we have seen that Noonan's ontological stance diminishes the scope of his claim. Parfit's argument only fails if one accepts the Multiple Occupancy Thesis. And this thesis is far from being an uncontroversial or widely accepted view.

As a result, I have argued that if we aim to defend that personal identity is what matters, it would be more fruitful to focus on the second premise of Parfit's argument. Namely, that what matters would be preserved in a fission scenario. This is difficult to do at the moment, as the notion of what matters remains philosophically imprecise. I have suggested that conventionalist accounts of personal identity might be of help here. In any case, there is still much to be discussed.[3]

Bibliography

Braddon-Mitchell, D., Miller, K., and The Hegeler Institute, 2004. How to be a conventional person. *Monist*, 87 (4), 457–474.

Lewis, D.K., 1983. Survival and identity. *In*: *Philosophical Papers Volume I*. New York: Oxford University Press, 55–77.

Muñoz-Corcuera, A., 2021. Persistence narrativism and the determinacy of personal identity. *Philosophia*, 49 (2), 723–739.

Muñoz-Corcuera, A., 2023. The transplant intuition as an argument for the biological approach. *Argumenta*, Online first.

Noonan, H., 2019. *Personal identity* (Third Edition). London: Routledge.

Noonan, H., 2024. The fission argument for the unimportance of identity cannot be correct. *Argumenta*, 19, 369–373.

Nozick, R., 1981. *Philosophical explanations*. Cambridge, MA: Harvard University Press.

Olson, E.T., 2019. Parfit's metaphysics and what matters in survival. *Argumenta*, 9, 21–39.

Parfit, D., 1984. *Reasons and persons*. Oxford: Clarendon.

Schechtman, M., 2014. *Staying alive: personal identity, practical concerns, and the unity of a life*. Oxford: Oxford University Press.

Shoemaker, S., 1984. Personal iodentity: a materialist's account. *In*: S. Shoemaker and R. Swinburne, eds. *Personal identity*. Oxford: Basil Blackwell, 67–132.

Wagner, N.F., 2019. Against cognitivism about personhood. *Erkenntnis*, 84(3), 657–686.

Williams, B., 1957. Personal identity and individuation. *Proceedings of the Aristotelian Society*, 57, 229–252.

Williams, B., 1970. The self and the future. *The Philosophical Review*, 79 (2), 161–180.

[3] This research was supported by the Ministry of Science and Innovation of the Spanish Government through the Research Project "Institution and Constitution of Individuality: Ontological, Social, and Legal Aspects" (PID2020-117413GA-I00/AEI/10.13039/501100011033).

Advisory Board

Agency without Action: On Responsibility for Omissions

Sofia Bonicalzi and Mario De Caro***

* *Roma Tre University*

** *Roma Tre University and Tufts University*

Abstract

In the last few years, there has been a growing philosophical interest in the problem of moral responsibility for omissions. Like actions, however, omissions are not all-of-a-kind. Recently, most of the research effort in this field has been devoted to the so-called *unwitting omissions*. However, in some cases, people make clear-eyed, or quasi-clear-eyed, decisions about not interfering with a given course of action potentially having unethical consequences (let's call these decisions *witting omissions*). In this paper, we abstract away from the epistemic concerns that typically refer to unwitting omissions to discuss the problem of moral responsibility for omissions *as omissions*, i.e., as non-events that may contribute to the occurrence of a state of affairs without necessarily being their *primary* cause. In particular, we call attention to how to define the set of omissions we are accountable for. Indeed, even narrowing the scope to witting omissions, there is an awful lot of morally undesirable events that we could contribute to preventing if we just wanted to do so. Thus, the question is: in which cases are we responsible for our witting omissions? In this perspective, we first consider the proposals of referring to derivative, role, or vicarious responsibility for arbitrating between the relevant cases. Although not mistaken, these proposals are helpful only in a limited subset of situations. Employing the example of a witness witnessing a crime by chance, we discuss a more encompassing strategy. Siding with those who see omissions as causes, we defend a counterfactual approach based on identifying when people could intervene and are normatively required to do so.

Keywords: Backward-looking responsibility, Witting omissions, Circumstantial luck, Counterfactual reasoning.

1. Introduction: Responsibility and Omissions

In the last few years, there has been a growing philosophical interest in the problem of moral responsibility for omissions. Like actions, however, omissions are not all-of-a-kind. Recently, most of the research effort in this field has been devoted to the so-called *unwitting omissions* (Clarke 2017, Fitzpatrick 2017, Murray

Argumenta 10, 1 (2024): 385—401
ISSN 2465-2334

DOI 10.14275/2465-2334/202200.bon
First published: 7 March 2022

and Vargas 2020, Wieland 2017). These are omissions that result from failures of attentiveness and vigilance, negligence, mistaken beliefs, or poor judgment,[1] for which the agent can sometimes be held 'culpably ignorant' (Rosen 2003, Smith 1983, 2011) and thereby (morally and sometimes legally) accountable.[2] Typical examples include surgeons leaving the surgical instruments into the patient's body or spouses forgetting to celebrate anniversaries or buy groceries on the way home (Amaya 2011, Clarke 2017).

In such cases, omissions may compromise responsibility, not *as omissions* but because they are instances of behavioural types that violate some standards for responsibility, i.e., the agent did not meet some epistemic requirements (Clarke 2017) or was not animated by the intent to harm or ill will (see Sripada, 2015, Talbert 2017). Correspondingly, one's willingness to forgive or excuse implies coming to terms with the observation that average humans often navigate the environment in the autopilot mode, cannot be expected to constantly meet even basic epistemic standards, and are often just lucky in avoiding pitfalls (Raz 2010, Sher 2009). In this respect, the problem of responsibility for unwitting omissions is partly analogous to that of responsibility for action-based non-deliberative patterns (*unwitting wrongdoing*)—including *habitual and automatic* actions (Lumer 2017) or even *inadvertent* actions whereby the agent unknowingly, or without having a corresponding intention or plan, causes an unethical consequence (Mele and Moser 1994).

There can be little doubt that many of our culpable omissions are unwitting (i.e., had we known better, we would have behaved differently). However, in other relevant cases (*witting omissions*), people make clear-eyed, or quasi-clear-eyed, decisions about not interfering with a given course of action, potentially having negative consequences. Unlike unwitting omissions, in witting omissions, the agent is animated by a direct intent to harm (Rachels 1975) or at least has the occurrent knowledge that some negative event, which she does not try to prevent, will likely occur (Pereboom 2015). This paper focuses on this latter case and abstracts away from the epistemic concerns that typically refer to unwitting omissions. We thus discuss responsibility for omissions *as omissions*, i.e., as non-events that may contribute to the occurrence of a state of affairs without necessarily being their 'primary cause'.[3]

[1] Due to failures of rational agency, unwitting omissions can be distinguished from *unintended omissions* that "never surfaced in the agent's mind" (Raz 2010: 449), e.g., calling "the person whose name is first in the Munich telephone directory today" (*Ibid*.). For an analogous distinction, see Brand 1971.

[2] According to the so-called 'control-based theories', accountability is grounded in the recognition that the agent satisfies some epistemic and agential requirements and is therefore an appropriate target of moral considerations, notably blame and praise (Björnsson 2017, Fischer and Ravizza 1998, McKenna 2012). Control-based accounts of responsibility are typically contrasted with the so-called 'quality of will' or 'deep-self' approaches, according to which agents are responsible for the actions that reveal their practical identity, i.e., their valuational system and the attitudes they reflectively endorse (Frankfurt 1971, Sripada 2016, Talbert 2017). Here, we mostly assume a control-based approach to responsibility and focus on accountability for the negative consequences of one's omissions.

[3] We define the notion of 'primary cause' as roughly referring to the most obvious cause, i.e., the event whose occurrence is more straightforwardly related to the effect—adopting

Here, we only briefly address whether people can be responsible for omissions *in general* (Clarke 2014) and set aside the issue of whether there is a morally relevant distinction between *doing* and *allowing harm* (Moore 1993, Rachels 1975, Scheffler 2004).[4] We instead call attention to how to define the set of omissions we are morally responsible for. Indeed, even narrowing the scope to witting omissions, there is an awful lot of unethical events that we could contribute to preventing if we just wanted to do so. For example, one can contemplate the idea, be aware that she could, feel tempted to (although without necessarily forming the intention or plan to) offer emotional support to all her Facebook friends, donate more money to charities after watching the daily news, feeding her friends' cat so that they can enjoy a free weekend, writing down her wedding anniversary on a calendar to avoid forgetting, etc. As a result, it remains to be clarified on what basis we should legitimately hold the agent accountable (i.e., blameworthy) in some, but probably not all, of such cases, i.e., what plausible normative expectations can restrain the set of omissions for which one can be held morally accountable.[5]

In the paper, we articulate three complementary strategies to define the set of witting omissions we are responsible for. The first one (§ 3) begins with the suggestion that one is morally responsible for given witting omissions in virtue of some prior specific commitment one has undertaken. This is partly analogous to the *tracing* strategy, often discussed in relation to unwitting omissions. We will conclude that it can deal with a limited type of omissions and move forwards by focusing on the set of omissions that can be attributed to no prior commitment. Our paradigmatic case is that of the witness to a crime, who *by chance* finds herself in a situation—i.e., she is in an appropriate spot in the proper moment—that occasions an action or an omission. The rest of the paper examines two other strategies to solve the problem. One consists in seeing people as responsible for witting omissions whenever at least one of two features is realised: some forms of vicarious responsibility, on the one hand, or the role—e.g., that of the witness—that is assigned to them, on the other. In § 4, we discuss this idea and conclude that, again, it can help address only a limited number of situations. In § 5, we present, and argue for, a counterfactual strategy similar to the capacitarian account for unwitting omissions, according to which we are responsible if and only if we *could* and *should* have intervened. We will articulate this proposal further in terms of spelling out what it means that, in a given scenario, an agent could and should have intervened.

Before getting started, however, it is important to stress that our primary concern here is backward-looking responsibility, i.e., the kind of responsibility

a counterfactual account of causation, the event without which the effect would have, most likely at least, not occurred (Lewis 1973).

[4] We side, however, with those who maintain that the distinction between doing and allowing harm does not necessarily collapse into the distinction between actions and omissions (see Foot 1967).

[5] Discussing unwitting omissions, Randolph Clarke analogously says that omissions identify instances of absent actions whereby the action would have been required by a norm, standard, or ideal (2014: 33). Others have suggested that even the standard that defines what omissions are causes must be normative (McGrath 2005). Here, we rather work out the standard that defines the causally relevant omissions for which we can be held accountable.

that implies a moral assessment of past events. Backward-looking responsibility is distinct from forward-looking responsibility, i.e., the type of responsibility that plays a functional role in shaping one's future behaviour (Pereboom 2015, Pereboom and Caruso 2018). Indeed, there might be many forward-looking reasons why it is appropriate to hold a person responsible for omissions. For example, we may blame a lazy witness to elicit a more collaborative spirit in the future. Less clear, though, is whether and why this practice is also acceptable in the backward-looking sense. Here, we set aside the broader question about whether it is *ever* fair to hold people responsible in a backward-looking sense—a matter that has to do with the long-lasting debate concerning free will (see, however, Bonicalzi 2019a, De Caro 2020). We will only assume, then, that there is a sense in which this is fair, and, in this light, we will discuss whether there may be responsibility for omissions in the backward-looking perspective.

2. Control and Causation by Omission

Primarily, people are held responsible for the consequences caused by their intentional actions, i.e., according to classic causal theories, actions that are caused by conscious mental states, such as intentions or plans (Bratman 2007, Davidson 1978, Mele and Sverdlik 1996). The underlying reason is that intentional actions tend to be the actions that agents can control (Shepherd 2014, but see Raz 2010).[6] Accordingly, jointly with the knowledge of the actions' circumstances, control is indeed often indicated as a necessary condition for responsibility (Bonicalzi 2019a).[7] So, within this framework, in generating responsibility causation comes into play twice. First, a conscious intention has to cause an action we perform, such that we are in control of this action; second, we are responsible for the outcome we have caused through that action. The question is whether this may also be true in the case, instead of an action, we wilfully perform an omission.

It has to be noted that the responsibility literature, broadly conceived, already presents notions of control that may work well for both actions and omissions. For example, Peter van Inwagen's well-known 'Consequence argument' (1983) considers control in terms of what is up to us and lack of control in terms of what is not up to us: I am not in control of the laws of physics because it is not up to me whether they are valid or not, but I can be in control of most of my behaviours because it is generally up to me whether I behave in a certain way (action) or not (omission). Joshua Shepherd (2014: 397) discusses control in terms of deploying "behavior in service of an intention": control is achieved when the representational content of the intention—e.g., a plan (see also Mele 1992)—matches the actual behaviour. Framed in these terms, witting omissions can be appropriately caused by our relevant mental states. For example, one can say that it is up to the witness of a crime to intervene or not (following van Inwagen), or that her plan of not act-

[6] The possible detachment between intention and control has prompted the discussion on the so-called 'deviant causal chains' (Davidson 1973, Mele 1992). Here we are only committed to the claim that, in standard cases, the presence of guiding, conscious intentions is necessary for an agent to control her actions.

[7] Conversely, uncontrolled bodily movements, such as hitting your partner while you are asleep, are not conducive to responsibility (Rumbold et al. 2016).

ing matches her behaviour (following Shepherd).[8] Analogously, Fischer and Ravizza's model of 'guidance control'—exercised when the action stems from one's reasons-responsive mechanism—deals equally well with actions and omissions (1998).

While it is clear that omissions can be caused by conscious intentions (thus satisfying the first causal element of responsibility), more problematic is whether they can cause some consequences (thus perhaps failing the second causal condition of responsibility). Consider, for example, a situation in which I witness a crime and decide not to get mixed up in it. My missing intervention is appropriately caused by, and can be traced back to, my decision (i.e., a mental action (Proust 2001)) of not intervening. However, if I omit to intervene and the perpetrator's action brings about the consequence, no action of mine has caused the crime. So, people can indeed be said to be responsible for their conscious decision of not intervening; however, it is less clear whether they can be responsible for the crime occurrence (an event) through their inaction (a non-event).

The problem of whether omissions can count as causes of events is not new but remains deeply controversial. Some philosophers have denied that omissions can enter causal relations (Armstrong 1999, Beebe 2003, Moore 2009, Varzi 2006): if these accounts are correct, we are not responsible in many situations in which we ordinarily think that we are, at least to some degree. However, others have claimed that omissions can play a causal role (Lewis 1987, McGrath 2005, Montminy 2020).[9] For example, in the context of his pragmatic defence of causal pluralism, Hilary Putnam presents a plausible view according to which omissions can be considered genuine causes to the extent that they count as *appropriate explanations* of events (Putnam 1999). There is not enough space here to discuss this issue in detail, so it will suffice to say that we stand with those who argue that omissions can be causes (De Caro 2021).

However, assuming that omissions can be causes raises another problem, i.e., how to distinguish between the omissions for which we are responsible and the ones for which we are not. *When* we can be held responsible depends on the answer to this question. In this case, the problem is that if we are taken to be responsible for the results of all our omissions, we are responsible in a much greater number of cases than we ordinarily think (Bernstein 2013, Henne et al. 2019, McGrath 2005).

3. Responsibility as Grounded in a Prior Action-Bound Commitment

At first glance, the fact that people can be held responsible for omissions looks uncontroversial. We are customarily held responsible for not taking good care of our children, not keeping our promises, etc. However, if we look closely, some

[8] In suggesting that omissions, like actions, are sometimes under our conscious control (i.e., we can deliberate about what we will not do), we are not assuming that *ipso facto* all omissions are forms of negative agency or that all witting omissions result from a conscious decision to refrain (see Clarke 2014).

[9] It's worth noticing that in our everyday practices, "omissions are as likely as actions to be judged as causes" (Clarke et al. 2015: 27), although they might be perceived as less causally relevant than actions (Baron and Ritov 2004, Bonicalzi 2019b, Bonicalzi et al. 2022).

puzzles emerge. How do we select the witting omissions we are responsible for in the accountability sense?

One first solution consists in suggesting that the witting omissions one is responsible for are those that derive from an action-bound prior commitment one has willingly undertaken. Consider a situation in which I am held responsible for not keeping my promise of feeding my friend's cat and, more interestingly for our discussion here, for the cat's subsequent death. I am held responsible because I willingly made this promise in the past: the promise is the action-bound (e.g., feeding the cat) responsibility-grounding prior commitment.

In this light, witting omissions are treated analogously to cases of derivative responsibility for unwitting actions or omissions (Rosen 2004). Let's briefly expand on this idea: a well-established view sees responsibility for behaviours we are not in control of as located in, or traced back to, some prior event that we could control (Fischer and Ravizza 1998, Smith 1983, Vargas 2005). Original or basic responsibility occurs when the agent is responsible for the event that directly led to some consequence. Derivative responsibility occurs when the agent is responsible for a prior event that eventually led to the result. The key distinction lies in whether the event that directly caused the consequence was under the agent's control or not. To give a standard example, I am non-derivatively responsible for wilfully drinking a glass of wine at a party while still sober. By contrast, I am derivatively responsible for hitting a pedestrian while drunkenly losing control of my car (and temporarily losing track of the relevant moral concerns for people's safety).

Analogously, the witting omissions I am responsible for could be conceived of as something similar to episodes of derivative responsibility. Some omissions do not happen in a vacuum but result from a prior action-bound commitment. In making a promise, I am expected to foresee that I could find myself, at some point in the future, in the position of being not so willing to keep it. Promises work as commitments for the future: their role consists in forestalling reconsideration (Cupit 1994). If I decide not to keep a promise—that is, if I omit to stand by it—I can be blamed because of my prior commitment, i.e., the promise I made.

However, 'tracing' (i.e., because of the promise I made, I am blameworthy for forgetting to feed the cat) has problems on its own as a general solution to the puzzle of unwitting omissions. In particular, it is unclear whether and how the epistemic requirement for accountability can be truly satisfied by features acquired "in circumstances that are epistemically remote from our current decisions" (Vargas 2005: 287).[10] Moreover, barring exceptional cases, witting omissions do not suffer from a constitutive lack of occurrent control and awareness. Therefore, it seems more evident that responsibility must be conceived as basic rather than derivative. One could then try to defend the idea that people are responsible in a basic or non-derivative manner for omissions explicitly forbidden by a prior action-bound commitment.

[10] See also Clarke 2014, Graham 2012, Rudy-Hiller 2017. The tracing strategy has also been criticised for tracing back *all* unwitting behaviours we are responsible for to acts of *clear-eyed akrasia* in which the agent decides that, despite knowing that she is going to commit wrong, she does not take any countermeasure (see Rosen 2004).

However, there seem to be omissions for which we want to hold people responsible and that do not rest on an action-bound prior commitment. Consider, for example, the case of the witness who chancily observes a crime from her window. Let's consider a real example: the infamous murder of Kitty Genovese and the debate it sparked. In March 1963, Kitty Genovese was murdered by Winston Moseley near the apartment she shared with her girlfriend, Mary Ann, in Queens, NY. Two weeks after the murder, *The New York Times* famously reported that 38 respectable citizens, who watched the three separate attacks against Kitty that ended in her death, remained indifferent to her repeated cries for help. While the report's authenticity has been put into question, the case received the attention of social scientists and generated research on the diffusion of morally reprehensible instances of the so-called 'bystander effect' (Beyer et al. 2017, Darley and Latané 1968). However, our focus is not on the psychological and societal implications of the diffusion of responsibility but on the normative question of whether and why the witnesses of that murder should be held responsible.

In Kitty's case, there was no prior action-bound commitment involving, among its foreseeable consequences, the possibility that one must prevent a crime. Therefore, if the witnesses had to be blamed for the crime occurrence, this could not depend on something they had done previously. As a preliminary conclusion, it seems that the link with a prior action-bound commitment may only explain a limited subset of cases. Therefore, in the rest of the paper, we will focus on the more relevant issue of the moral responsibility for witting omissions that do not rest on a prior action-bound commitment.

4. Responsibility as Grounded in Role and Vicarious Responsibility

As said, there are situations in which one can be blamed for violating commitments that are not action-bound, as in the case of feeding a friends' cat during the weekend. This typically happens when people have forms of role and vicarious responsibility.

Role responsibility identifies the obligations a person has in virtue of occupying a specific societal position (Cane 2016). A parent's obligations towards her children can be framed in terms of role responsibility. Similarly, a head nurse will have specific responsibilities towards her patients. Moral reproach for faulty behaviours can be grounded in the violation of the obligations associated with one's role: a parent can be blamed for not taking care of her children and a nurse for not giving a life-saving medicine to her patients. Since roles come with obligations, violations of such obligations—whether through actions or omissions—lead to blaming attitudes.

Vicarious responsibility (or liability) characterises situations in which a person is held responsible for the faulty behaviour of others (May 1983). The justification often depends on the fact that the person is expected to prevent the defective behaviour of another person. Role and vicarious responsibility often, but not always, go together.[11] In fact, the expectation that a person exerts some con-

[11] Vicarious liability is usually invoked in the context of employer-employee relationships, e.g., in the English tort law (Mulheron 2016). We can extend the notion to other

trol may quickly arise because she has a specific role. For example, a parent can be held vicariously responsible for the faulty behaviour of her children or the head nurse for the crimes of a subordinate. Role and vicarious responsibility do not necessarily lead to moral reproach: a parent can be role responsible for their children and always fulfil the obligations associated with this so that she never actually deserves moral reproach. And moral reproach for the misdemeanours of a subordinate might not even be appropriate for the vicariously responsible head nurse. This suggests that role and vicarious responsibility are not necessarily connected to an ongoing moral appraisal.

In omission cases where there is no prior action-bound commitment on which responsibility depends, we may deploy the tools of role and vicarious responsibility. Consider role responsibility as applied to the case of the witness. Being a 'witness' can be framed as a role assigned to a bystander. Like many other societal roles, being a witness implies specific obligations, e.g., trying to prevent a crime when it is possible to do it safely. There are, however, some problems with this strategy. First of all, people usually are not automatically assigned roles. *Prima facie*, it seems that people must voluntarily consent to it to be given a specific role (see Murray and Vargas 2020). Consent can be explicit, as in the case of the head nurse signing a contract, or implicit, as in the case of the parent deciding to have children. In both cases, responsibility is grounded in this prior consent.

Ideally, the person who gives her consent is also expected to be aware of the foreseeable consequences of assuming the role. By contrast, before witnessing a crime, there is no time when a witness knowingly accepts to take up the role of witness. If consent is necessary for a role to be binding, a witness cannot be obligation-bound to do something. Consent plays a role in vicarious responsibility as well. Usually, to be vicariously responsible for the crimes of her subordinates, the head nurse must consent to assume a role that implies bearing responsibility for their behaviour. Moseley's boss—supposing, for the sake of argument, that he had one—is not vicariously responsible for his crime insofar as she never consented to supervise him during her spare time.

However, one may object that there are situations in which role and vicarious responsibility come without consent. If this is the case, it might be fair to hold the witness responsible after all: think, for example, of being a soldier in cases of enforced conscription. If a soldier has an obligation-bound role despite her lack of consent, this counts as an exception to the norm. A reply might be that the soldier forced to take up the role finds herself in a situation whose exceptionality legitimises forcing people to assume roles they would not be willing to take up otherwise. There might be situations in which the rule of consent is overridden by other considerations whereby people are forcefully assigned offi-

contexts, however. For an example of vicarious responsibility that does not collapse into role responsibility, consider the following. Suppose that an adult finds herself in a dangerous situation involving some minors, as happens to Léon, the character played by Jean Reno in the movie *Léon: The Professional*. Léon is a hitman who, more or less willingly, gets to become friend with his 12-year-old neighbour Mathilda and teaches her to use weapons. Mathilda wants to avenge the death of her brother, murdered by an agent of the Drug Enforcement Administration. In this case, it seems that, while Léon has no role responsibility towards Mathilda (e.g., he is not Mathilda's father), he might be vicariously responsible for her misdemeanours.

cial roles without consent. However, this answer does not readily apply to the case of the witness (which is an unofficial role, if any), even though she may also find herself in exceptional situations that impose an extra burden on her. At the same time, there are more familiar, unexceptional roles that seemingly come without consent: we all share the role of being a son or a daughter without us consenting to that. Whereas parents have a role responsibility towards their offspring (because, and perhaps only when, they decided to have children or engage in sexual intercourses that might lead to pregnancy), it may seem that the offspring also have role responsibility towards their parents. However, if we look closely at the case, it becomes doubtful whether the offspring's responsibility is genuinely grounded in a role. Offspring may assume explicit role responsibilities towards their parents at some point in their life, e.g., when they are old or sick. Setting this aside, it is unclear that offspring have role responsibilities by default, e.g., an unbounded obligation to care for their parents.

There are indeed asymmetries between the respective obligations of parents and offspring: parents are expected to take care of their offspring independently of the offspring's behaviour and attitude. By contrast, offspring are not expected to take care of their parents in *all* possible situations. Blaming the offspring for not taking care of their parents might be more easily justified by something with little in common with a specific role, such as the obligation to honour special or agent-relative obligations to subsets of people, including family members or friends (Parfit 1984). Thus, a thorough examination of the two alleged exceptions to the rule of consent—e.g., the soldier and the offspring—shows that they are not central to our discussion (the former because it is rooted in exceptional conditions, the latter because offspring's obligations do not seem to depend very much on their role). Given this, we suggest that the rule of consent applies to most relevant cases of role responsibility.

The next question is whether we can have vicarious responsibility without consent. It is arguably hard to find any case of the sort. Unless this is made explicit, offspring are not vicariously responsible for their parents' faulty behaviours. Forcefully enlisted soldiers are not usually vicariously responsible for their comrades. Vicarious responsibility without consent, i.e., voluntary membership or agreement, is very rare and controversial.[12] It might sometimes be brought into play but usually, we believe, in metaphorical terms. Consider, for example, the debate about whether the adult male German population in the 1930s/1940s was vicariously responsible for the Nazi crimes (Darcy 2007). Proper vicarious responsibility might come from party membership or vote in the election. Still, it would be a stretch to hold all Germans vicariously responsible just for being German and independently of their individual behaviour.

To sum up, we have suggested that both role and vicarious responsibility are usually grounded in consent. Still, we are not committed to claiming that consent is necessary for them to arise. Although this notion remains unsatisfactorily vague, there might be exceptional circumstances in which the need for consent is overridden by other considerations, especially in the case of role responsibility (as in the case of the forcefully enlisted soldier). However, most everyday omissions

[12] It might be debatable whether, in a case such as that of Léon, there might be vicarious responsibility without consent or whether Léon's attitudes towards Mathilda imply at least a form of implicit consent.

for which we bear responsibility concern ordinary things like not supervising one's employees or handling one's children. At least in situations of these kinds, the rule of consent easily applies. From a practical point of view, renouncing consent would lead to the untameable proliferation of responsibilities, with as many roles and obligations as possible interpersonal conditions. If every situation in which we find ourselves (e.g., being a bystander) were recast as a role (e.g., being a witness), we would end up having no clear definition of roles anymore. Thus, in § 5, we will discuss the challenges and the prospects of a complementary, more encompassing treatment of responsibility for omissions in the absence of a prior action-bound commitment (§ 3) or consent (§ 4).

5. Responsibility in Chancy Circumstances

In this section, we will consider whether a witness can be held responsible for the occurrence of a crime she is present at, assuming that she has unluckily found herself in the situation we described and that she could do something to prevent the crime with no serious harm for herself. Before doing that, however, we should briefly reflect upon the role of luck in assigning responsibility. As we said, we hold people responsible, first and foremost, for situations in which they willingly put themselves (see § 3) or to which they consented (see § 4). Let's assume that the witness found herself in a situation where she did not willingly put herself or did not consent. Here, the notion of *circumstantial luck*, e.g., luck in the circumstances one finds oneself in, comes at hand (Nagel 1979). Might the presence of circumstantial luck per se suffice to make it inappropriate to hold the witness responsible? We doubt that this is the case.

The first thing to notice is that this problem of luck looks orthogonal to the distinction between actions and omissions. Most of our behaviours are driven by chancy circumstances while being willingly chosen. Usually, this simple fact does not prevent us from holding the relevant agent responsible.[13] Consider the following as an example of a faulty action driven by chancy circumstances. Butch Cassidy, one of the most successful thieves in history, was famous for robbing banks and trains. Should we blame him any less if he had robbed a bank after happening to find the vault open? Obviously, in some cases, finding oneself by chance in a situation where one can prevent harm, as often happens with omissions, coincides with lacking a direct intent to harm. While Moseley intended to harm Kitty through his action, the witnesses plausibly did not intend to harm Kitty but merely intended not to get mixed up in the crime. This may lead to diminished blame compared to that attributed to Moseley, without luck playing any specific role in modifying one's responsibility.[14]

Let's recap the problem. First of all, if we accept that omissions can be causes, we might blame the witness for the crime, but other events could be indicated as causally relevant such that the witness would not have to be blamed. For example, had Kitty not gone to work on the 12th of March 1963, she would

[13] Although in some cases agents might be additionally blamed for actively creating the circumstances from which specific actions and omissions will likely derive.

[14] Evidence from cognitive science has shown that we have a widespread tendency to attribute less responsibility for omissions than actions, often assuming that the former are less intended than the latter (Bonicalzi et al. 2022, Spranca et al. 1991). However, this is not necessarily the case (see Rachels 1975).

not have been killed. This is known as the 'causal selection problem' (Bernstein 2013), which consists in the difficulty of individuating the causes of an event against the background of the causal conditions that have made that event possible. Per se, this problem is not restricted to omissions (Menzies 2004). At any rate, Henne and colleagues (2019) discuss the 'profligate causes problem' as a version of it by targeting omissions. This indicates that it is hard to identify the causal status of an omission, given that other omissions could have caused the same result. For example, had Mary Ann walked Kitty's home (knowing that they lived in a dodgy area), this might have helped prevent the assault. Since there are many more omissions than actions, holding that omissions can be causes produces an overflow of potential causes.[15] How do we decide which omissions count as causes and are thus relevant for responsibility?

The selection and the profligate causes problems have been extensively discussed (Hesslow 1988). Here, we will just say a few words concerning how they relate to responsibility. As said, we consider the notion of 'primary cause' as referring to the event whose occurrence is more straightforwardly associated with the effect—adopting a counterfactual account of causation, the event without which the effect would have, most likely at least, not occurred (Lewis 1973). The primary cause of Kitty's death is that Moseley assaulted her, not that the witnesses did not intervene. Nonetheless, we may also blame the witnesses for Kitty's death. This suggests that, for someone to be held responsible for an event, their behaviour need not be the primary cause of that event. Indeed, the witnesses of Kitty's homicide are thought to be blameworthy even though their omitted intervention is not the primary cause of her death.

For the agent to be held responsible, their omissions must thus count as causes without necessarily being the primary causes of an event. Whereas this helps explain how relevant omissions lead to responsibility attributions (e.g., had the witnesses intervened, Kitty might have been saved), it can also make the selection and the profligate causes problems intractable (since even omissions that are seemingly minor can be causally relevant). To avoid a causal explosion, in the absence of a prior action-bound commitment (§ 3) or consent (§ 4), we must then identify some criteria that explain when an agent is morally responsible for her omissions.

A plausible suggestion is that an agent is responsible only when she could and should have intervened. But how do we know what an agent could and should have done? Counterfactual reasoning can indeed be muddy. To address this point, it is helpful to consider an adjusted version of the 'capacitarian' approach to responsibility, which so far has been mainly discussed in relation to unwitting omissions where people fail to meet given epistemic standards (rather than the control condition) for responsibility (Clarke 2014, Murray and Vargas 2020, Rudy-Hiller 2017, Sher, 2009). Capacitarian accounts emend the basic epistemic condition by suggesting that, at least in some cases of unwitting (but culpable) omissions, the agent could and should have known better—a condition that is potentially sufficient for responsibility. For instance, one might plau-

[15] Another potential issue one might consider is that, if we consider omissions as causes of an event X, the notion of 'sufficient cause' becomes intrinsically problematic insofar as all the innumerable omissions that did not prevent the event X could be listed as part of the sufficient cause of the occurrence of X.

sibly think that the forgetful spouse could and should have written the date of the anniversary on her calendar to avoid forgetting. Deciding whether the agent could or should have known better must avoid arbitrariness and be grounded in plausible considerations about their standard capacities, available information, or professional training. This decision cannot depend only on statistical regularities about the reasonable or average person's prototypical performances; by contrast, one must also consider the agent's specific capacities, e.g., cognitive functioning and ability to retrieve information (Sher 2009).[16]

In the case of witting omissions, the problem does not necessarily involve epistemic considerations concerning factual awareness. In our toy example, the witness is aware that a crime is happening and decides not to get involved, and we can stipulate that she makes up her mind before it is too late to intervene. However, (fallible) *agential* and *normative* counterfactual considerations can be helpful to select the omissions we are accountable for. Such concerns must be grounded in the witness's capacities, information, and training, on the one hand, and in the existing social and moral norms, on the other hand.

First, agential considerations are necessary to determine whether the witness could have had a reasonable opportunity for successfully intervening without endangering herself. This judgment must be partially relativised to the specific agent's capacities rather than solely determined by statistical regularities about what an ideal or average person is expected to do. Furthermore, the judgment must be relativised to the specific social context in which one is operating, granted that different contexts may allow for other actions to be done safely, e.g., the same agent might safely or non-safely intervene depending on whether she lives in a residential or dodgy part of town. In our example, a necessary but not sufficient condition for the witness to be blameworthy—in a basic, non-derivative way—is that she had a reasonable expectation that she could have safely prevented the crime through her intervention (e.g., shouting to Moseley or calling the police).[17] Agents have such reasonable expectations when their mental states match mind-independent states of affairs in the world. In this sense, the witnesses could be blamed for not rescuing Kitty, given their reasonable expectations of being safely able to do so. By contrast, as an equally endan-

[16] Obviously, the capacitarian view does not deal equally well with all scenarios. Many cases remain problematic due to difficulties in determining whether the individual could have known better. Even relativising the judgement to the agent's cognitive capacities, occasional lapses, which deviate from the specific agent's standards, remain a problematic and concrete possibility (Amaya 2011). Whereas it is trivial to say that the person should have known better in general, it remains dubious whether she could have known better in these specific circumstances.

[17] Some philosophers maintain that one can omit to do X only if one is able or has the objective opportunity to do X. For example, the witness cannot omit to call the police if her phone is broken unbeknownst to her. In this case, she can only omit to *try* to call the police (Clarke 2014). However, although the missing actions for which the witness is blameworthy can be different (failing to call or failing to try to call), the witness can be equally blameworthy in both cases, assuming that she was unaware of whether the phone was broken (see Frankfurt 1988, van Inwagen 1983). Additionally, we suggest that a witting omission could count as culpable only when the witness had a reasonable expectation that she would not have endangered herself by intervening.

gered young woman walking Kitty home, Mary Ann could have been spared the blame for the analogous omission of not rescuing Kitty.[18]

Second, even when there are reasonable expectations that they could have made the difference, agents remain nonetheless blameless in the absence of a normative standard of some sort suggesting that they should have intervened, e.g., the moral obligation to rescue that a witness may have. This requirement may not depend on prior commitments or roles but on the existence of basic interpersonal obligations to act. Walking Kitty home would be a nice gesture, but, although Mary Ann might be aware that they live in a dodgy part of town, there is no moral obligation for her to do so insofar as deciding not to is not associated with the violation of any foreseeable moral requirement. Some courses of action (e.g., feeding the cat or walking someone's home) are made obligatory, so that the corresponding omission can be blameworthy, only when there is a specific action-bound prior commitment we must uphold or when we consented to take up some roles or vicarious responsibilities. Others (e.g., the obligation to rescue someone in immediate distress) apply independently of prior commitment or consent. Even in such cases, however, what one can normatively be expected to do may vary as a function of the social environment and its structures (see Hurley 2011, Rudy-Hiller 2019). For instance, witnesses have an obligation to intervene, i.e., by calling the police, whenever they happen to be in a social context where they have a reasonable expectation that involving the police is the best course of action to get help without running into significant risk.

Of course, holding a witness responsible for the occurrence of a crime does not imply that she is as blameworthy as the perpetrator. One's degree of blameworthiness depends on the balance between various causal, normative, and agential considerations. Furthermore, no default cut-off point allows us to establish whether the witness is morally required to intervene,[19] or whether the risk is too high. Indeed, the counterfactual *could* and the moral *should* are meant to specify necessary, but not sufficient, conditions for responsibility, establishing a morally relevant connection between a non-event and some state of affairs. Whereas it is unfair to blame an intellectually disabled or a defenceless witness for not fighting an armed aggressor, there might be borderline cases where responsibilities remain to be decided.[20] Nonetheless, whenever they are jointly sat-

[18] Obviously, agents might be wrong in their evaluations of what reasonable expectations are in place or in their assessment of how much time they can spend deliberating. For example, the witnesses might have wrongly assumed that it would have been dangerous to intervene or could have spent too much time deliberating. In this case, however, the more classic capacitarian approach could help explain when ignorance for mistaken beliefs or unwitting wrongs more generally counts as culpable.

[19] It might be difficult even to decide when acting is morally required or supererogatory. If we had the obligation to act morally whenever possible, this would imply that we have the responsibility to engage in all sorts of helping behaviours constantly. For example, do we have the obligation to be part of the Global Kidney Exchange program (Minerva et al. 2019) in virtue of the fact that we could do so?

[20] Especially in situations in which the agent seemingly fulfils basic epistemic considerations but still fails to act. Consider an example inspired by Berofsky (2002): if an agent suffers from arachnophobia, she might be unable to remove a spider from the wall even though she would do so if she wanted to. Unfortunately, given her condition, she cannot be wanting to remove the spider. Similar considerations might apply when a witness de-

isfied, these necessary conditions allow us to address the profligate causes and the causal selection problem by adequately restricting the range of omissions for which we could be plausibly held accountable.

6. Conclusions

We are usually keen on drawing a thick line between what we are responsible for and not. In this paper, we have focused on responsibility for witting omissions, first considering whether derivative, role, or vicarious responsibility can help arbitrate between relevant cases. Although not mistaken, we found that these solutions are helpful only in a limited subset of situations. Employing the example of the bystander witnessing a crime by chance, we thus discussed a more general strategy. Siding with those who see omissions as causes, we defended a counterfactual approach based on identifying when people could intervene and are normatively required to do so. Of course, to adjudicate individual cases, further work has to be done to refine this necessary condition, particularly to explain how the different agential and normative requisites come together.

References

Amaya, S. 2011, "Slips", *Noûs*, 47, 3, 559-576.

Armstrong, D. 1999, "The Open Door", in Sankey, H. (ed.), *Causation and Laws of Nature*, Dordrecht: Kluwer, 175-85.

Baron, J. and Ritov, I. 2004, "Omission Bias, Individual Differences, and Normality", *Organizational Behavior and Human Decision Processes*, 94, 2, 74-85.

Beebee, H. 2003, "Causing and Nothingness", in Collins, J., Hall, N., and Paul, L. A. (eds.), *Causation and Counterfactuals*, Cambridge (MA): MIT Press, 291-308.

Bernstein, S. 2013, "Omissions as Possibilities", *Philosophical Studies*, 167, 1, 1-23.

Berofsky, B. 2002, "Ifs, Cans, and Free Will: The Issues", in Kane, R. (ed.), *The Oxford Handbook of Free Will*, 1st ed., New York: Oxford University Press, 181-201.

Beyer, F., Sidarus, N., Bonicalzi, S., and Haggard, P. 2017, "Beyond Self-Serving Bias: Diffusion of Responsibility Reduces Sense of Agency and Outcome Monitoring", *Social Cognitive and Affective Neuroscience*, 11, 12, 138-45.

Björnsson, G. 2017, "Explaining Away the Epistemic Condition on Moral Responsibility", in Robichaud, P. and Wieland, J.W. (eds.), *Responsibility: The Epistemic Condition*, Oxford: Oxford University Press, 146-62.

Bonicalzi, S. 2019a, *Rethinking Moral Responsibility*, Milano-Londra (IT/UK): Mimesis International.

Bonicalzi, S. 2019b, "Agire o non agire. Il ruolo dell'omission bias nei giudizi morali", *Notizie di Politeia*, XXXV, 136, 124-36.

Bonicalzi, S., Kulakova, E., Brozzo, C., Gilbert, S.J., and Haggard, P. 2022, "The Dynamics of Moral Judgment: Joint Role of Dependence and Transference Causal Explanations on Responsibility Attribution", *Philosophical Psychology*, 35, 1, DOI 10.1080/09515089.2021.2021165.

cides not to intervene because she is paralysed by some irrational fear preventing her from acting.

Brand, M. 1971, "The Language of Not Doing", *American Philosophical Quarterly*, 8, 45-53.

Bratman, M. 2007, *Structures of Agency: Essays*, Oxford: Oxford University Press.

Cane, P. 2016, "Role Responsibility", *The Journal of Ethics*, 20, 178-298.

Clarke, R. 2014, *Omissions: Agency, Metaphysics, and Responsibility*, Oxford: Oxford University Press.

Clarke, R. 2017, "Ignorance, Revision, and Commonsense", in Robichaud, P. and Wieland, J.W. (eds.), *Responsibility: The Epistemic Condition*, Oxford: Oxford University Press, 233-51.

Clarke, R., Shepherd, J., Stigall, J., Waller, R., and Serpentine, C. 2015, "Causation, Norms, and Omissions: A Study of Causal Judgments", *Philosophical Psychology*, 28, 2, 279-93.

Cupit, G. 1994, "How Requests (And Promises) Create Obligations", *The Philosophical Quarterly*, 44, 177, 439-55.

Darcy, S. 2007, *Collective Responsibility and Accountability under International Law*, Leiden: Brill.

Darley, J. and Latané, B. 1968, "Bystander Intervention in Emergencies: Diffusion of Responsibility", *Journal of Personality and Social Psychology*, 8, 4, 377-83.

Davidson, D. 1973, "Freedom to Act", in Honderich, T. (ed.), *Essays on Freedom of Action*, London: Routledge and Kegan Paul, 139-56.

Davidson, D. 1978, "Intending", in Yovel, Y. (ed.), *Philosophy of History and Action*, Philosophical Studies Series in Philosophy, Dordrecht: Springer, 41-60.

De Caro, M. 2020, "*Actus non facit reum nisi mens sit rea*. The Concept of Guilt in the Age of Cognitive Science", in D'Aloia, A. and Errigo, M.C. (eds.), *Neuroscience and Law*, Cham: Springer, 69-79.

De Caro, M. 2021, "The Antinomy of Omissions", *Synthesis*, 1, 99-112.

Fischer, J.M. and Ravizza, M. 1998, *Responsibility and Control: A Theory of Moral Responsibility*, New York: Cambridge University Press.

FitzPatrick, W.J. 2017, "Unwitting Wrongdoing, Reasonable Expectations, and Blameworthiness", in Robichaud, P. and Wieland, J.W., *Responsibility: The Epistemic Condition*, Oxford: Oxford University Press, 29-46.

Foot, P. 1967, "The Problem of Abortion and the Doctrine of Double Effect", *Oxford Review*, 5, 5-15.

Frankfurt, H.G. 1971, "Freedom of the Will and the Concept of a Person", in Id. 1988, *The Importance of What We Care About: Philosophical Essays*, New York: Cambridge University Press, 11-25.

Frankfurt, H.G. 1988, "What We are Morally Responsible For", in Id. 1988, *The Importance of What We Care About: Philosophical Essays*, New York: Cambridge University Press, 95-103.

Graham, P.A. 2012, "A Sketch of a Theory of Moral Blameworthiness", *Philosophy and Phenomenological Research*, 88, 2, 388-409.

Henne, P., Pinillos, Á., and De Brigard, F. 2019, "Cause by Omission and Norm: Not Watering Plants", *Australasian Journal of Philosophy*, 95, 2, 270-83.

Hesslow, G. 1988, "The Problem of Causal Selection", in Hilton, D.J. (ed.), *Contemporary Science and Natural Explanation: Commonsense Conceptions of Causality*, New York: New York University Press, 11-32.

Hurley, S. 2011, "The Public Ecology of Responsibility", in Knight, C. and Stemplowska, Z. (eds.), *Responsibility and Distributive Justice*, Oxford: Oxford University Press, 187-215.

Lewis D.K. 1973, *Counterfactuals*, Oxford (UK): Blackwell Publishers.

Lewis D.K. 1987, "Postscripts to 'Causation'", in Id., *Philosophical Papers II*, Oxford: Oxford University Press, 172-213.

Lumer, C. 2017, "Automatic Actions – Agency, Intentionality, and Responsibility", *Philosophical Psychology*, 30, 5, 616-44.

May, L. 1983, "Vicarious Agency and Corporate Responsibility", *Philosophical Studies*, 43, 1, 69-83.

McGrath, S. 2005, "Causation by Omission: A Dilemma", *Philosophical Studies*, 123, 1/2, 125-48.

McKenna, M. 2012, *Conversation and Responsibility*, New York: Oxford University Press.

Mele, A. 1992, *Springs of Action*, New York: Oxford University Press.

Mele, A. and Moser, P.K. 1994, "Intentional Action", *Noûs*, 281, 39-68.

Mele, A. and Sverdlik, S. 1996, "Intention, Intentional Action, and Moral Responsibility", *Philosophical Studies*, 82, 3, 265-87.

Menzies, P. 2004, Difference-Making in Context", in Collins, J., Hall, N., and Paul, L.A. (eds.), *Causation and Counterfactuals*, Cambridge (MA): MIT Press, 139-80.

Minerva, F., Savulescu, J., and Singer, P. 2019, "The Ethics of the Global Kidney Exchange Programme", *The Lancet*, 394, 10210, 1175-1778.

Montminy, M. 2020, "Omissions and Their Effect", *Journal of the American Philosophical Association*, 6, 4, 502-16.

Moore, M.S. 1993, *Act and Crime: The Philosophy of Action and Its Implications for Criminal Law*, New York (NY): Oxford University Press.

Moore, M.S. 2009, *Causation and Responsibility: An Essay on Law, Morals, and Metaphysics*, Oxford: Oxford University Press.

Mulheron, R. 2016, *Principles of Tort Law*, Cambridge: Cambridge University Press.

Murray, S. and Vargas, M. 2020, "Vigilance and Control", *Philosophical Studies*, 177, 3, 825-43.

Nagel, T. 1979, "Moral Luck", in Statman, D. (ed.), *Moral Luck*, New York: SUNY Press, 57-72.

Parfit, D. 1984, *Reasons and Persons*, Oxford: Oxford University Press.

Pereboom, D. 2015, "Omissions and Different Senses of Responsibility", in Buckareff, A., Moya, C., and Rosell, S. (eds.), *Agency, Freedom, and Moral Responsibility*, London: Palgrave Macmillan.

Pereboom, D. and Caruso, G. 2018, "Hard-Incompatibilist Existentialism: Neuroscience, Punishment, and Meaning in Life", in Caruso, G. and Flanagan, O. (eds.), *Neuroexistentialism: Meaning, Morals, and Purpose in the Age of Neuroscience*, New York: Oxford University Press, 193-222.

Proust, J. 2001, "A Plea for Mental Acts", *Synthese*, 129, 105-28.

Putnam, H. 1999, *The Threefold Cord. Mind, Body, and World*, New York: Columbia University Press.

Rachels, J. 1975, "Active and Passive Euthanasia", The New England Journal of Medicine, 292, 78-86.

Raz, J. 2010, "Being in the World", *Ratio*, 234, 433-52.

Rosen, G. 2003, "Culpability and Ignorance", *Proceedings of the Aristotelian Society*, 103, 1, 61-84.

Rosen, G. 2004, "Scepticism about Moral Responsibility", *Philosophical Perspectives*, 18, 1, 295-313.

Rudy-Hiller, F. 2017, "A Capacitarian Account of Culpable Ignorance", *Pacific Philosophical Quarterly*, 98, 1, 398-426.

Rudy-Hiller, F. 2019, "Moral Ignorance and the Social Nature of Responsible Agency", *Inquiry*, 1502-3923.

Rumbold, J., Morrison, I., and Riha, R. 2016, "Criminal Law and Parasomnias: Some Legal Clarifications", *Journal of Clinical Sleep Medicine*, 12, 8, 1197-98.

Scheffler, S. 2004, "Doing and Allowing", *Ethics*, 114, 2, 215-39.

Shepherd, J. 2014, "The Contours of Control", *Philosophical Studies*, 170, 395-411.

Sher, G. 2009, *Who Knew? Responsibility Without Awareness*, New York: Oxford University Press.

Smith, H.M. 1983, "Culpable Ignorance", *The Philosophical Review*, 92, 4, 543-71.

Smith, H.M. 2011, "Non-Tracing Cases of Culpable Ignorance", *Criminal Law and Philosophy*, 5, 115-46.

Spranca, M., Minsk, E., and Baron, J. 1991, "Omission and Commission in Judgment and Choice", *Journal of Experimental Social Psychology,* 27, 1, 76-105.

Sripada, C. 2015, "Moral Responsibility, Reasons, and the Self", in Shoemaker, D. (ed.), *Oxford Studies in Agency and Responsibility*, 3, Oxford: Oxford University Press, 242-64.

Sripada, C. 2016, "Self-Expression: A Deep Self Theory of Moral Responsibility", *Philosophical Studies*, 173, 5, 1203-32.

Talbert, M. 2017, "Akrasia, Awareness, and Blameworthiness", in Robichaud, P. and Wieland, J.W. (eds.), *Responsibility: The Epistemic Condition*, Oxford: Oxford University Press, 47-63.

Van Inwagen, P. 1983, *An Essay on Free Will*, New York: Oxford University Press.

Vargas, M. 2005, "The Trouble with Tracing", *Midwest Studies in Philosophy*, 29, 1, 269-91.

Varzi, A. 2006, "The Talk I Was Supposed to Give", in Bottani, A. and Davies, R. (eds.), *Modes of Existence: Papers in Ontology and Philosophical Logic*, Frankfurt: Ontos Verlag, 131-52.

Wieland, J.W. 2017, "Introduction", in Robichaud, P. and Wieland, J.W. (eds.), *Responsibility: The Epistemic Condition*, Oxford: Oxford University Press, 1-28.

I Don't Feel like That! A Phenomenology-Free Approach to Moods

Daniele Cassaghi

University of Milan

Abstract

People in moods usually claim that they feel in a certain way, and yet they also say that moods are undirected states. If one takes these reports at face value, moods are a counterexample to representationalism, namely the doctrine of a necessary connection between phenomenal character and content. The standard representationalist answer is to deny moods' undirectedness in order to capture the phenomenal character of moods. I go in the opposite direction: I will deny moods' phenomenal character and secure moods' undirectedness instead. I will show that both our folk-psychological usage and our introspective based reports favour this proposal over standard representationalism.

Keywords: Moods, Emotions, Representationalism, Intentionalism, Functionalism.

1. Introduction

Being in a bad mood is something that everybody has experienced at least once in a lifetime. Considering the world to be a terrible place, holding beliefs about negative things, etc. are all experiences that anybody in a depressive mood may have had. Like emotions, moods are a common occurrence in our mental lives. Usually, people's introspective reports on moods highlight two features. Firstly, being in a certain mood is feeling in a certain way. In other words, we report that moods are mental states with a phenomenal character, we report that there is something *it-is-like* to be in a certain mood. Elation, for instance, is feeling extremely positive. Secondly, we report that moods are undirected.[1] In contrast to states like beliefs and desires, we take moods to be contentless. Someone who is

[1] For the purposes of this paper, I will use the term "representational", "directed" and "content" interchangeably. Indeed, the heart of the dispute is whether moods are about, *i.e. refer to,* something else than themselves. And this minimal notion of reference is common to any account of representations, directedness and contents. See also footnotes 2, 4 and 5.

Argumenta 10, 1 (2024): 403—420
ISSN 2465-2334

DOI 10.14275/2465-2334/202200.cas
First published: 3 March 2022

elated, for example, reports a sort of positive and diffuse affection, but also that this affection seems not to be directed. In contrast, emotions seem to have specific contents. For example, a person feeling happy may report that she is happy about *her friend* or, alternatively, happy that *her friend enjoys a certain fortune.*[2]

This overall picture causes a lot of troubles for philosophers attracted to the doctrine of phenomenal character called *Representationalism*:

> Representationalism: It is necessary for any phenomenal state to be also a representational state (Siewert 2017).[3]

Detractors of Representationalism (*e.g.* Voltolini 2013, Bordini 2017) usually argue that (*a*) moods are phenomenal states; (*b*) since moods are undirected, they lack a representational content; (*c*) Representationalism predicts that any phenomenal state has a representational content; (*d*) thus, anti-representationalists conclude, Representationalism is false.[4] Call this the *argument from moods.*[5]

The doctrine I call *standard representationalism* tries to answer to the argument from moods by providing moods with a representational content. Thus, standard representationalism rejects premise *(b)* of the argument from moods. It does so in two ways. The most common one is to assume that moods are directed to an object, contrary to what it seems. Usually this object is very general, like the *whole world* (Solomon 1993, Goldie 2000, Crane 2007), *everything* (Seager 2002), *our total environment* (namely, the totality of relations that a subject holds with things in the world, including past, present, future and possible relations) (Mitchell 2018a, 2018b), *things in general* (Tappolet 2018, Kriegel 2019). A recent proposal is by Rossi (2019), who claims that moods are directed to *undetermined objects*, namely something that the subject is not able to identify. A less popular view is Tye's (1995), according to which moods are about *changes in our bodily equilibrium*. The second way to reject premise *(b)* is Mendelovici's (2013a, 2013b). According to her, moods are not directed to any object, but they are directed to *sui generis, unbound, uninstantiated, evaluative properties* of the kind emotions usually attribute to their objects (for example "*being dangerous, being wonderful etc.*").[6] Standard representationalists aim to treat moods with the same analysis they adopt for emotions. Indeed, they start from the observation that emotions represent certain objects

[2] According to your conception of emotions, you may consider emotions either to be propositional states or directed only to objects (see Kriegel 2017), or both.

[3] This definition of representationalism is minimal. It is accepted by philosophers with very different representationalist accounts (Tye 1995, Dretske 1995, Horgan and Tienson 2002, Byrne 2001, Crane 2007, Mendelovici 2013a, 2013b, Kriegel 2019).

[4] The choice of the label 'representationalism', instead of the more common 'intentionalism', is to avoid any confusion with another doctrine called 'intentionalism': the thesis that all mental states are intentional states. I am not interested in defending this latter thesis, which can be accommodated also by phenomenal states that are representational only in a *contingent way*. For the same reason, I avoid speaking about intentionality *tout court*.

[5] Other scholars believing that moods are undirected are Armon-Jones (1991), Frijda (1994), Deonna and Teroni (2012), and Searle (1983).

[6] I believe also Mendelovici (2013b), who takes object-undirectedness at *face value*, is happy with this definition of directedness. Indeed, she assumes that moods *refer to* unbound properties. See Bordini (2017) for further remarks on this point. Moreover, further discussion on Mendelovici's unbound representationalism can be found in Kind (2013), and Hatzimoysis (2017).

under certain evaluative lights, and all but Mendelovici think that this is true also for moods. For the purposes of this paper, I will assume that the standard representationalist analysis is correct for emotions, but, as I will argue, it should not be extended to moods.

The aim of my paper is to show that standard representationalism barks at the wrong tree in respect to the argument from moods. Indeed, the best account of moods for Representationalism should deny premise *(a)*, and reject both moods' phenomenal character and moods' directedness. My overall strategy is the following: I will turn the argument from moods into an argument for phenomenology-free moods, and defend its most controversial premise, *i.e.* that moods are contentless. The conclusion that moods are phenomenology-free comes from accepting Representationalism while denying that moods are representational.

The primary aim of this paper is not to provide a knock-down argument for Representationalism in front of anti-representationalists' criticism. More modestly, it is to show that the *phenomenology-free approach* is the best shot a representationalist has for moods. My conclusion can be accepted also by anti-representationalists, albeit in a conditional form: *if* we assume Representationalism, the best way to account for moods is the phenomenology-free approach. Only Lormand (1996) has explicitly claimed that moods are not phenomenal states so far. This paper is purported to give new life to this currently discarded idea. For this reason, I will tentatively sketch a positive proposal, which marries Representationalism with mood functionalism.

I will address some unconvincing arguments for phenomenology-free moods in section 2 and I will propose my own one. In section 3, I will show that both folk-psychological explanations and introspection-based reports support contentless moods. In section 4, I will sketch a way to account for phenomenology-free moods. In so doing, I will explore the idea that moods are functional states. This, I will show, vindicates the idea that moods are both phenomenology-free and undirected (See Lormand 1996). Finally, some objections will be met in section 5.

2. Arguments for Phenomenology-Free Moods

Phenomenology-free moods seem *prima facie* implausible. So far, only Lormand (1996) has advanced an argument for the thesis that moods do not have a phenomenal character. He holds that it is necessary for any phenomenal state to be liable either to the "image illusion" or to the "appearance illusion". Since moods are liable to none, moods are phenomenology-free, Lormand concludes. Unfortunately, this argument is unconvincing. Lormand defines the "image illusion" as the illusion in which subjects take mental objects to have properties of nonmental objects (we do not take beliefs of a yellow banana to be banana-like, nor yellow). It is easy to see how standard representationalists can argue that moods produce the "image illusion": in elation we may take both the world and the mood itself to be wonderful. This is enough to dispel Lormand's argument. However, things are even worse. Lormand defines the "appearance illusion" as the illusion in which subjects experience nonmental objects as having properties that are proper of mental objects. But a representationalist about emotions may think that my fear of a dog is experiencing the dog *as dangerous,* but claim that the dog itself does not instantiate the property of *being-dangerous*. It may be my mental activity that projects the mental property of *being-dangerous* onto the dog. And this form of ap-

pearance illusion, the representationalist concludes, may constitute the phenomenal character of my fear.[7] So, the claim that the appearance illusion prevents mental states from being phenomenal seems unmotivated.

However, there is a simple argument for phenomenology-free moods to pursue. It is sufficient to notice that, if moods are undirected, moods cannot be phenomenal states. Thus, we can transform the argument from moods framed in the introduction into the following *argument for phenomenology-free moods*:

Argument for Phenomenology-Free Moods

1) *Representationalism is true* [assumption]
2) *Moods are undirected* [undirectedness is distinctive of moods]
3) *If Representationalism is true, undirected states cannot have a phenomenal character* [by definition of Representationalism]
4) *Therefore*, *moods do not have a phenomenal character* [conclusion].[8]

This argument is valid. However, its soundness crucially relies on premise (2). In the next section I will give some arguments in support: both our folk-psychological practice (section 3.1) and our introspective reports (section 3.2) suggest that moods are undirected rather than directed.

3. The Contentless-Approach

The task of this section is to provide two arguments in favour of undirected moods. Much of the discussion will be a confrontation with standard representationalism, whose main assumption is that moods are directed rather than undirected. The defence goes into two steps. Firstly, I will show that our folk-psychological usage of moods favours contentless moods over standard representationalism (section 3.1). Secondly, I will show that undirectedness of moods is favoured by the fact that standard representationalism has no clear account of some of our introspective reports (section 3.2).

3.1. Folk-Psychology and Moods

The *distinctive explanatory role* of moods in folk-psychology suggests that moods are contentless. I offer two Observations to make this point.

Observation n°1: Let us assume that Benny ran away from the room. Here is a list of possible folk-psychological explanations of her behaviour:

1) Benny ran away from the room because she believed there was a fire in it.
2) Benny ran away from the room because she desired to avoid the fire in it.
3) Benny ran away from the room because she feared the fire in it.
4) Benny ran away from the room because she was anxious.

These are all folk-psychological explanations (as certified by the "because-clause" in each sentence) of the same behaviour by Benny. Obviously, explanations (1-3) necessarily involve a content for the mental state. Indeed, let us compare (1-3), with the following:

[7] This position is known as *projectivism*. Mendelovici's (2013a, 2013b) theory is an example.
[8] Curiously Lormand (1996) writes in a footnote that representationalism suggests phenomenology-free moods. He did not go for the full-blown conclusion that moods *are* phenomenology-free in a representationalist framework.

1*) Benny ran away from the room because she believed.
2*) Benny ran away from the room because she desired.
3*) Benny ran away from the room because she feared.

(1*-3*) are incomplete explanations of Benny's behaviour. We are tempted to ask "what" the mental state is about in each explanation. Without this piece of information, Benny's flight would remain unintelligible. The same is not true for (4), which is a complete explanation despite the state's lack of content. The first conclusion is obvious enough: (4) is an explanation involving a mood (anxiety). (1-3) involve contentful states (beliefs, desires and emotions respectively). We do not need content to make sense of folk-psychological explanations involving moods.

There are some replies to Observation n°1. For example, some may notice that an explanation like (5) is fully intelligible:

5) Benny ran away from the room because she was anxious about the fire.

So, is (5) a "moody" explanation involving contents? I am not sure that (5) is a counterexample to my position, and no standard representationalist theory accepts that anxiety in (5) is a genuine mood. Indeed, no standard representationalist proposal assumes moods to be directed toward particulars like the fire. And for a good reason: if moods were about particulars, they would be considered by standard representationalists as emotional episodes.[9] Standard representationalists consider (5) to involve a mood term 'anxiety' picking up a contentful emotion.[10] An alternative suggestion comes from Mendelovici (2013b) and Stephan (2017). They claim that there are different kinds of affective states, corresponding to the kind of directedness involved. For example, moods *senso strictu* are genuinely undirected, moods *senso latu* may have a content (either a general one or a particular one). We may restrict our analysis only to the former kind of mood.

The second reply to Observation n°1 is to notice that (4) may be an abbreviation of (4*):

4*) Benny ran away from the room because she represented X *as dangerous*. [Where "X" may stand for a general object, an undetermined object, or simply marks that dangerousness occurs unbound].

In other words, the standard representationalist may complain that our folk-psychological explanations only superficially treat moods as contentless. Rather, mood terms are just abbreviations hiding a reference to mood contents. Finally, she says, also folk-psychological explanations involving contentful moods are adequate to explain Benny's behaviour.

I think this is false: Benny's flee in (4*) is unintelligible. Indeed, following the reconstruction by the standard representationalist, Benny is motivated to run away from the room because she takes a general object, her total environment for example, to be dangerous. But why should she run away *from the room*, if she takes *her total environment* (instead of the room) to be dangerous? The obvious move is to consider that the room is part of her total environment and claim that Benny is

[9] Tappolet (2018) might disagree. According to her modal account, there is a genuine difference between the emotion of fear, which represents the fire to be dangerous, and the mood of anxiety, which represents the fire as *likely* to be dangerous. However, the endorser of such a proposal should provide an answer to the following question: "If moods have a so clear object like the fire, why do we misreport their lack of directedness?". It is a difficult task, as shown in section 3.2.

[10] This is suggested also by Lormand (1985).

ultimately running away from her total environment of which the room is part. But her total environment also includes outside the room. So why should she go outside the room, since it is also dangerous *there*? In general, it makes no sense to run away from our own total environment to reach again our own total environment. Actually, it makes no sense to run away from our total environment *at all*. (Ditto for the other proposed general objects.) The point easily generalises to unbound dangerousness and to undetermined objects, which should be located anywhere by Benny's lights.

Standard representationalists may be tempted to appeal to other mental states to adjust the explanation. For example, Benny may still *believe* that it is safer outside the room, even if this contradicts the information provided by the mood. This complicates the structure of the folk-psychological explanation of her behaviour. There are *parsimony reasons* against this solution. Folk-psychological explanations treating moods as contentless do not need to pose other mental states to make Benny's behaviour intelligible. Moreover, in the same vein, it is more parsimonious *per se* to treat moods as contentless, rather than inflating our ideology of moods and assume they enjoy the property of directedness.

Observation n° 2: As Lormand (1985), Sizer (2000), DeLancey (2006), Deonna and Teroni (2012), and Rossi (2019) maintain, our moods seem *arational*, that is moods are insensitive to reasons and norms of rationalisations. A cluster of folk-psychological features points to this direction. Among these features, we find moods' inability to both rationalise behaviours and transmit justifications (Lormand 1985, Sizer 2000, Deonna and Teroni 2012); their inability to be derived from practical reasonings (Lormand 1985); their usage for asking for mitigating circumstances (Goldie 2000, Deonna and Teroni 2012); the idea that moods, in contrast to emotions, do not provide subjects with goals to act toward objects (the point is advanced by Lormand 1985, Sizer 2000, and DeLancey 2006. In a slightly different vein by Price 2006 and Rossi 2019. Tappolet 2018 makes it part of *pervasivity*, namely the ability of moods to influence a greater number of mental states than emotions). So, it seems that folk-psychological explanations involving moods are much more like causal explanations connecting two events (*e.g.* "Mary gave the wrong answer because her concentration dropped"), than to explanations involving rationality (*e.g.* "Mary made this choice because it was the best chance to achieve her goal").[11]

The arational character nicely fits the idea of contentless moods. Rationalization crucially relies on contents: it is all about explaining and predicting what a person *ought to* do, think and feel *in virtue of* previous information in her possession (*e.g.* her goals, the different ways in which goals and means are delivered to the subject etc.). It follows that arationality is a necessary feature of contentless states, but it is not sufficient for establishing that moods are not representational. Indeed,

[11] This clarification is important to avoid a certain ambiguity. One may claim that "Mary gave the wrong answer because her concentration dropped" is an explanation that makes it rational, *i.e.* makes it intelligible to a third party why Mary behave that way. This is not the sense of rationality involved here, since in my sense rationality norms are those governing deliberative practical reasoning. They are applied only to the subject whose behaviour has to be explained. Moods seems to escape rationality in this sense, so they are arational.

there are contentful states which arguably do not obey norms of rationality. Perceptions may be a significant example. I take Observation n° 1 to support the claim that moods are ultimately contentless and they differ from perception in this respect. Here I defend the arational character of moods against recent criticisms.

The standard representationalist Mitchell (2018a) argues that moods are "rationally intelligible". According to him, subjects feel a causal relation between their total environments and their moods. So, subjects interpret their moods as an *appropriate response* to their total environments. This sense of appropriateness is Mitchell's "rational intelligibility", and it is very different from the sensitivity to reasons employed in Observation n°2. Hence, it is not a counterexample to the arational character of moods. Moreover, in Mitchell's view, moods are close to perceptions: they make subjects aware of their own (total) environment, and are not dependent on previous information. This is further evidence that Mitchell's "rational intelligibility" is not in tension with arationality, which may be true for perceptions.

Rossi challenges the idea that moods are not employed in rationalising behaviour by offering the following cases:

> 6) She decided she needed a change, something more stimulating in her life, because she was assailed every day by an endless boredom.
> 7) She decided to call a taxi, as she felt quite anxious in the street alone at such a late hour of the day (Rossi 2019: 18).

Rossi points out that in explanations (6) and (7) moods are not just causal pulls for the subjects: the subjects *are informed* by their moods that something is wrong, and they decide to act accordingly. It is not just a belief of being in a certain mood that motivates their behaviour, Rossi concludes. I do not think Rossi's cases are conclusive. The persuasive force of his claim relies on the usage of the verb "decided" in both explanations. The assumption by Rossi is that the final behaviour is the product of a practical reasoning. However, it does not immediately follow that the mood provides information about the environment to the subject within this practical reasoning. Both explanations (6) and (7) make perfectly sense if the subject desires to avoid the unpleasant mood itself and deliberates consequently. In other words: the belief of being in a mood is sufficient to make sense of both (6) and (7).

Rossi claims that moods themselves may be rationalised. He offers the next two explanations as a case study:

> 8) "Why are you so irritable? There is nothing to be upset about!"
> 9) "Susan was in a good mood for no particular reason" (Rossi 2019: 18).

Again, these cases seem unconvincing. (8) seems to pick up an emotion (being upset) called with a mood term ('irritability') rather than a mood. In (9), Susan's elation does not rely upon information already in her possession. This is in line with the arational character of mood.[12]

[12] Rossi (2019: 20) is ultimately trying to show that moods can be evaluated like perceptions with these examples. Therefore, the same answer we provided to Mitchell holds: the kind of rational character envisaged in examples (8) and (9) is not the same of that of Observation n°2.

Kostochka (2020) claims that moods are sensitive to beliefs. She provides some cases of moods starting to change as soon as our beliefs change. In one of these examples, Kostochka suggests that a depressed person may start feeling better after positively re-evaluating what she has done during the day. This person undergoes a change in her beliefs: she does not believe that the day was negative anymore, she now believes her day is positive. And the mood changes accordingly. Again, this example does not jeopardise the arational character of moods. Indeed, if successful, Kostochka (2020) offers a case of correlation between change in beliefs and change in mood. But it is doubtful that this sensitivity to belief variation takes place in virtue of moods' contents, as we should expect if moods were genuinely part of practical reasoning. Kostochka does not offer an account of what the object of a mood may be. So, we have no clue about how the content of beliefs interacts with the alleged information provided by the mood. Thus, in this example by Kostochka, it may be the case that moods are still arational after all: they can be automatic reactions *caused* by belief change.

To sum up, our usage of moods terms in folk-psychological explanations and their arational character support the view that moods are contentless. This motivates premise (2) of the argument for phenomenology-free moods in turn. Other evidence comes from our introspective reports, as it will be shown in the next section.

3.2. Introspective Reports

As Bordini (2017) suggests, introspective reports are part of the reasons why we take moods to be undirected. People report that their moods are undirected after introspection. A simple explanation of why people speak this way is that moods *are* undirected. According to this view, people' introspective reports should be taken at *face value*. However, this is not the explanation a standard representationalist may give, since she believes that moods have objects. Thus, endorsers of standard representationalism are in charge to provide an alternative explanation of people introspective reports about moods. If moods are directed, why people (mis)report that moods lack directedness?

Standard representationalists are not explicit, but they seem to assume that the object the mood is about plays the trick: in representing the world in general, we are not representing its parts (individual dogs, telephones, trains, etc.) distinctly. Our experience presents us with an indefinite whole with no discrimination among its parts. Given that we usually notice that our experience is directed towards x *by noticing* how x distinguishes itself from the other things, we misjudge this lack of discrimination as lack of directedness. Thus, the standard representationalist concludes, we misreport.

At a closer look this proposal is untenable. In general, it is not obviously true that we mistake the lack of discrimination as lack of content. I may hold beliefs about the whole world (*e.g.* the belief that the world is of an infinite size), and probably I do not discriminate the proper parts of it. But I do not fail to recognise that my belief is directed upon the world, and this is what I am prone to report. Another example may be the hallucination of an undetermined, shapeless blob in front of me. I see no reason why the very same presentation of such a thing should prevent me to report that I am experiencing at least *something.* But reporting that I am experiencing *something* is tantamount to report that my experience is directed. There is little or no reason to assume that things are any different for moods.

Other standard representationalist theories are not on a better ground. Representing undetermined objects is to represent that *an undiscriminated object* is in a certain way (Rossi 2019: 2). So, it is still the case that we represent that *there is something*, and the same problem remains. If moods are about unbound properties, we should still report about the occurrence of *some* properties. The same is true for representations of bodily changes. Please notice that my point holds even if the content is nonconceptual (Tye 1995). In perceiving red things, we effectively report that we are perceiving something, even if we cannot report which kind of red shade we are presented with. The same should apply for moods.

To sum up, a view that treats moods as undirected has a simple explanation of why people report moods as having no object. Reports should be taken at *face value*. Standard representationalism has no clear account of why people report that moods are undirected. Introspective reports about moods' lack of object favour a view according to which moods are contentless.

4. The Phenomenology-Free Approach

The argument for phenomenology-free moods in section 2 is valid, my defence of premise (2) makes it also sound. We are now in position to sketch how a representationalist, phenomenology-free theory of moods might look like. The aim of this section is not to defend a certain account of phenomenology-free moods over the others. More modestly, it is just to show that at *least one* phenomenology-free account of moods is viable. The starting point of this inquiry will be the next three questions. I take them to be the most common worries a proposal involving phenomenology-free moods might rise. The answers will shed lights on the positive view of moods I am advancing in this paper. These are:

1) How is it possible that we misreport about the phenomenal character of moods?
2) What kind of states are moods?
3) Is there a tension between the representationalist framework and phenomenology-free moods?

Without an answer to the first question, the phenomenology-free approach would be obviously incomplete. This is the topic of section 4.1. The second question arises because moods are neither phenomenal nor contentful states in my view. So, one might wonder what kind of mental states they are. In section 4.2, I will explore the possibility that moods are functional states. Although functionalism and Representationalism are usually considered rivals, they can be compatible in respect to moods. The third question arises because some might consider contentless moods at odds with the very representationalist project to account for mental states in terms of representations. The question will be assessed in section 4.3.

4.1. The Phenomenological Error

Why do we misreport about moods' phenomenal character? Before answering this question, we should notice that there is consensus on the fact that moods are responsible for the occurrence of certain congruent emotions (Lormand 1985, Sizer 2000, Chomanski 2017, Tappolet 2018). Therefore, even if moods are mental states with no "specific" phenomenal character by themselves, they systematically come with an associated phenomenal character: that of the emotions caused by the mood itself. Crudely put, according to the phenomenology-free theory, we

misattribute the phenomenal character of the emotions the mood generates to the mood itself. How is it possible to *mistake* the phenomenal character of the emotions as if it belongs to the mood, then?

We can understand the phenomenal character and the content as two different *aspects* of the same thing: the emotions related to the mood. A great array of emotional states is generated when we are in a certain mood. It is impossible to pay full attention to our affective states all the time. Therefore, we tend to devote just a small amount of peripheral attention to the phenomenal character of the emotions generated by the mood. The relevant "part" of the phenomenal character of emotions is linked to our bodily changes, which are still maintained in a mood state. Since we do not direct all our attention to a single emotion in this state, we do not attend to its outward content. When we are in the mood of anxiety, for example, token emotions of fear occur. So, we should expect that also bodily changes preparing the subject to fight or flee are in place (Deonna and Teroni 2012). These bodily changes are those constituting part of the phenomenal character of anxiety. Suppose that I am anxious and there is a dog in front of me: token episodes of fear of that dog occur in me because of my mood. My suggestion is the following: it is possible to pay peripheral attention to our bodily changes, without paying attention to the dog. And we misattribute this phenomenological element of fear to the mood.[13] Finally notice, peripheral attention is directed toward emotions' features. It's directed to nothing regarding the mood.

This goes along with the idea that the phenomenal character we misattribute to the mood is reported to be unitary, not a mere juxtaposition of emotional characters. Let us take the case of anxiety again. Many different emotions are generated when we are in this mood: fear about particulars in the environments, worries about possible situations in the future, anger about both offensive and innocuous things. As long as we are not attending to any particular content, we are not able to make a distinction between the felt characters of these emotions: we are not attending to their external contents. So, we do not discriminate among the different emotions, and we may report a sort of unitary phenomenal character for the mood.[14] However, we know anxiety makes us much more sensitive to what goes on outside: a strange noise would be soon the focus of my attention. In that moment we can start being afraid, and we can single out our fear from the "over-

[13] This nicely fits attitudinalism about emotions (Deonna and Teroni 2012, Kriegel 2017), according to which we feel our body as an attitude toward an external content: peripheral attention would be directed to attitudinal features. Pure representationalism (Tye 2008), according to which bodily changes are represented in the content, can accommodate this view by assuming that peripheral attention is directed to *part* of the emotion content: the one representing bodily changes.

[14] Does it mean that we cannot distinguish two similar emotions (*e.g.* anger and fear) if they are directed to the same object? Nope. According to the customary analysis of emotions by representationalism, when we attend to the content of our fear of Darth Vader and to the content of our anger against Darth Vader, we attend to two different contents. Anger and fear attribute different evaluative properties to Darth Vader (*e.g. as dangerous* vs. *as despicable* respectively).

all" phenomenal character by attending to its content, the strange noise. This attention trick obviously explains why sometimes we feel an emotion "flowing" into a mood or *vice versa*.[15]

This account has one further bonus. Mendelovici (2013a, 2013b), Tappolet (2018), and Rossi (2019) suggest the phenomenal character of moods to be similar to emotions. For example, we are able to mark out a corresponding mood for each emotion: happiness/elation; sadness/depression, anger/irritability, etc. Arguably, this is due to which kind of emotion is prevalently generated by each mood. The reason for this similarity is obviously that we become aware of being in a certain mood in virtue of the phenomenal character of the generated emotions. On the other hand, any putative difference we report between the phenomenal character of the mood, and the phenomenal character of the corresponding emotion, can be easily explained in virtue of the fact that, in making such a contrast, we have to put emotions into focus. Thus, we get access to those contents which were previously neglected. And the experience of contentful state feels different from a (putative) experience of a contentless state.[16] Finally, relying on attention has another advantage: it explains why we do not misattribute the phenomenal character of an emotion generated by, for example, a belief to the belief itself. Beliefs are contentful states: we can put our focal attention to their contents. This is tantamount to single out the belief from our train of thought, and understand that the putative phenomenology of beliefs, if any, is different from that of emotions. The same cannot be not true for moods, which are contentless states and cannot be singled out in the same manner.[17]

4.2. What Are Moods?

I assume Representationalism is true, but according to the phenomenology-free approach, moods are not representations. So, we need a nonrepresentational metaphysical account to explain their nature. The main rival of standard representationalism, unpopular nowadays, is the *functionalist account of moods*. This approach has been developed in length by Lormand (1985), Griffiths (1997) and Sizer

[15] The transformation of emotions into moods and *vice versa* as a feature of moods is discussed by Deonna and Teroni (2012) and Rossi (2019). The choice of attention as responsible for the phenomenological error, rather than any other faculty, is driven by the assumption that attending to our inner or outer environment is necessary to perform judgments (including introspective reports) in the first place.

[16] Moreover, strictly speaking, the putative phenomenal character of elation includes happiness, the prevailing emotion, and the phenomenal character of a bunch of emotions of different kinds. For this reason, I take the phenomenal character associated to elation to be *similar* but *not entirely indistinguishable* from that of happiness. I thank an anonymous referee for pushing me to clarify this.

[17] This proposal is fully compatible with Chomanski's (2017) Manifestation principle. According to this principle, *what it is like* to be in a mood is to be aware that other kinds of mental states feel differently from how they usually do, and that this modification is somehow coherent. These states do not limit themselves to emotions, but they also include perceptions and thoughts (Chomanski 2017: 107). However, we should be careful to accept Manifestation. Full-blown Manifestation can probably be endorsed only by accepting both a sort of cognitive penetrability for perception and that thoughts have a phenomenal character. These are open options, but they need philosophical defence. Therefore, I prefer to be neutral. So, I just focus on the phenomenal character of emotions.

(2000). If functionalism is viable, it is possible to account for the nature of moods while considering moods as undirected.

Functionalist theories of moods are designed to account for a distinctive feature of moods, namely their *pervasivity*.[18] When we are in a mood our mental life undergoes a deep change: moods alter the standard functioning of our mind. Among the other things, we tend to undergo certain emotional episodes, thoughts and beliefs and avoid certain others. For example, in elation we enjoy positive thoughts about the joy of life, and we do not entertain beliefs about how painful our illness once was. Moreover, according to the empirical literature reviewed by Sizer (2000: 764), positive moods tend to generate mental states focused on a wider range of information, creative thoughts, and unusual associations of ideas. They also reduce the number of thoughts focused on details, which are peculiar of some negative moods. Moods have effects also on attention, memory, and people tend to interpret ambiguous situations according to the mood they are in (see also Eysenck and Keane 2010 for a review). Plausibly, given that their primary function is to alter our mental lives in a systematic way, they might have evolved to make the subject more responsive to the environment.[19]

Hence, the main idea behind moods functionalism is that moods are best described as *functions*. Moods are those states causing (and caused by) the occurrence of congruent emotions, beliefs and thoughts, and hampering (and hampered by) certain others. The *functional description* of a mood is the list of states systematically causing and caused by the mood. Emotions play a key role in this respect. Indeed, a certain mood would not be the mood it is, if it did not cause the related emotions. In other words, being responsible for the generation of certain emotions, but not of certain others, is part of the mood's functional description. And this fact explains why we systematically misattribute the same kind of phenomenal character to the same kind of mood. In other words, we do not feel "saddish" when we are in elation, because elation always causes happiness, joy etc. and hampers sadness. Finally, an additional reason to adopt a functional interpretation of moods is that the functional description may be deduced by our usage of moods in folk-psychological explanations. As a result, the functional role of any mood matches the way in which we use that mood in folk-psychology. We describe elation as that mood causing positive thought and hampering sadness, because this is the role elation plays in our folk-psychological explanations.

For our purposes, the main advantage of functionalism is that the functional description is the only thing that matters to identify moods: neither contents nor phenomenal characters are required for moods' identification (see Lormand 1985). In other words, functionalism about moods vindicates both the main features of the phenomenology-free approach: moods' lack of phenomenal character and moods' lack of directedness. In the same vein, the functional description does not rely on contents. Therefore, a functionalist account of moods makes sense of the arational character of moods we addressed in section 3.1 (Lormand 1985, Sizer 2000, Grif-

[18] Pervasivity is taken to be a distinctive theory of moods by Sizer Lormand (1985), Sizer (2000), DeLancey (2006) and Chomanski (2017), Tappolet (2018), and Rossi (2019). According to these authors other mental states, especially emotions, do not have the same impact on our mental life. Pervasivity is criticized by Gallegos (2017). Chomanski (2018) offered a reply.

[19] As suggested by Price (2006), whilst she does not support functionalism.

fiths 1997). Finally, Sizer (2000) suggests that moods are best described as subpersonal states, influencing higher order states.[20] This proposal nicely fits the picture I am drawing. It explains why, strictly speaking, we encounter neither moods' phenomenal character nor moods' contents in our introspection.

These observations are enough to reach the purpose of this section: showing that there is at least *one* viable way to account for phenomenology-free and undirected moods. Remarkably, mood functionalism is compatible with Representationalism, which is the first premise of the argument for phenomenology-free moods. Functional states do not violate the rule according to which any phenomenal state must be a representational state. This is not to say that functionalism is the only game in town to account for phenomenology-free moods. I am claiming that the compatibility with Representationalism makes functionalism a good candidate to account for the nature of moods.

4.3. Representationalism and Functionalism

To recap, according to the phenomenology-free approach, moods are neither phenomenal nor directed. This conclusion is reached under the assumption of the truth of Representationalism within the argument for phenomenology-free moods. However, the phenomenology-free approach predicts that Representationalism does not apply to moods after all. This might seem a betrayal of the whole representationalist project. Standard representationalists, for example, might be motivated by a sort of theoretical unity. Not only might they believe that Representationalism is true, but also that it must be applied to any mental state (Bordini 2017).[21] So, the phenomenology-free approach may be unpalatable to those philosophers thinking that every mental state is representational. One might wonder whether it makes sense to assume Representationalism at the very beginning: the phenomenology-free approach to moods risks downplaying the force of Representationalism exactly because it accepts that some mental states are not representational.

These considerations should not be overestimated for three reasons. Firstly, theoretical unity is undoubtably a virtue of standard representationalism, but it cannot be a reason to prefer standard representationalism in this context. Indeed, our choice among two explanations can be driven by theoretical unity only when two theories have the *same performance* in respect to the *explananda*. Only if the two theories have both the same explicatory power and the same flaws, theoretical unity might be a reason to prefer one over the other. However, the discussion in section 3 has shown that standard representationalism has some problems at accounting for both introspective reports and our usage of moods in folk-psychology. These problems do not affect the phenomenology-free approach, which has all the merits of standard representationalism, with no flaws. Unless these problems are fixed, the lack of theoretical unity does not provide decisive ground against the phenomenology-free approach.

Secondly, theoretical unity may be one reason to accept Representationalism but there could also be independent ones. For example, materialistic-oriented people may agree with Dretske (1995) and claim that representations are still the

[20] As Drayson (2012) convincingly argues, the high order/subpersonal distinction and the conscious/unconscious distinction do not overlap.

[21] This is the thesis according to which directedness is the "mark of the mental" (Voltolini 2013). See footnote 4.

best shot to naturalise phenomenology, namely explaining phenomenal properties in terms of natural properties. These materialist philosophers may be less interested in theoretical unity and more prone to accept phenomenology-free moods. Such an approach would allow emotions to be naturalised, since they are representational states, and moods come to be even less problematic: they do not need to be naturalised in the first place. Thus, the phenomenology-free approach must be very attractive for materialistically-oriented philosophers.

Finally, it is possible to appeal to Sizer's suggestion of subpersonal moods to vindicate a weaker interpretation of theoretical unity behind Representationalism. It may be the case that every *higher order* state, albeit not every *mental* state, is representational. But *qua* subpersonal, moods are not higher order states.

5. Objections and Replies

In this section I will explore some objections advanced to my theory and provide some replies.

Objection1: Moods are not the only kind of affective states which seem to have a phenomenal character but not a content. There may be cases of contentless emotions which are clearly phenomenal states but do not have a content. If these states are caused by a mood, then the problem returns: you misattribute the phenomenal character of these states, which are ultimately contentless, to the mood.

Reply1: I am sympathetic to this kind of reasoning, but I think it does not affect my theory. My aim here is to provide an account for moods, under the assumption that Representationalism is true for the other mental states, including emotions. So, the working hypothesis is that there are no states like contentless emotions, exactly because Representationalism is true. If we assumed the presence of such states in our mind's architecture, then it would be a problem for the representationalist, regardless of whether my account of moods is correct. In the same vein, analyses like DeLancey's (2006) stating that moods and emotions are contentless states of the same kind are ruled out by default. Ultimately, this line of reasoning does not affect my theory of moods, which is not concerned with other contentless states.

Objection2: It is possible to pay *full attention* to a mood and thus understand that it has a genuine phenomenal character. For example, when I am elated because I read philosophy, I focus completely on my mood, and I understand that it has a phenomenal character. So, whilst it seems plausible to misattribute the phenomenal character of emotions to the mood when we dedicate peripheral attention, it is hard to maintain that there is no phenomenology in the mood when we focus on the mood only.

Reply2: This objection is based on introspection. My reply is to deny that full attention reveals anything about the mood. In this case, it reveals that I am experiencing an emotion: I am happy about philosophy. Note that according to my theory, no attention whatsoever can be directed toward the mood: it is partly directed towards emotions and mostly directed toward the environment. This is so because moods lack semantic properties. Let us assume that the functionalist proposal in section 4.2 is the right metaphysical account of phenomenology-free moods. We should notice that when we pay full attention to our mental states, we pay attention to their contents, not to their vehicles. In a functionalist framework, vehicles are inaccessible to us: we have access to contents put in a "belief box" or in a "desire box", but not to the "boxes" themselves. And if moods are

purely functional parameters, then they are "vehicles" with no contents. Therefore they cannot be targeted by our attention. Indeed, we can attend to the fact that we desire that *p* instead of believing that *p*, just because we are entertaining *p*. We do not attend to "desire" full stop.[22]

Besides the former response to the objection, which is entirely "internal" to a functionalist view, a more general reason against the idea that moods can be the target of full attention is phenomenological. The alleged phenomenal character of the mood presents itself as a sort of diffuse "affective background" connoting our actions and thoughts. Speaking metaphorically, it is something that always stays in "the back of our minds": making it the centre of our attention would make the mood lose this character.

Objection3: Amy Kind (2013) makes her case against standard representationalism by stating how standard representationalism is not able to make sense of the variation in intensity of moods. She claims that we may feel a variation of intensity in her affective states, even though the represented object does not appear different to us.

Reply3: With some adjustments, a functionalist theory of moods such as the one envisaged in section 4.2 accommodates variations in the intensity of moods. A functional description of moods may allow that moods are similar to knobs regulating *quantitatively* the amount of emotions generated: the more emotions produced, the stronger the overall phenomenal character appears to be. In other words, an intense mood is simply a mood allowing for the production of a greater amount of emotions.

Objection4: People may report moods to be directed (see Mitchell 2018a: 123, commenting on Davitz's 1969 findings). Mitchell writes:

> For example, in Joel R. Davitz (1969) study, 42% of subjects reported depression as involving a sense that "everything seems useless, absurd, meaningless" and 34% reported anxiety as involving an experience that "everything seems out of proportion." On the positive side, 66% of subjects reported cheerfulness and contentment as involving a sense that "the world seems basically good and beautiful" and 62% reported serenity as involving "peace with the world" (Mitchell 2018a: 123).

The force of these statements should not be overestimated. The fact that a relevant part of interviewed subjects reports that moods to be directed does not prevent that another relevant part of people, including philosophers interested in moods, reports that moods are undirected. Therefore, an easy way to dismiss Davitz's reports is coming back to the distinction between moods *sensu stricto* and moods *sensu lato* and claim that only moods *sensu lato* are reported to be directed (section 3.2). So, we may take both reports at *face value*, but limit our analysis to moods *sensu stricto*.

Objection5: The phenomenology-free theory is still implausible. It is problematic to accept that, appearances notwithstanding, moods are not qualitative states, for they lack a phenomenal character. According to the doctrine of the *Cartesian collapse of qualitative appearance onto reality* (Cartesian Collapse for short), if one has a certain inner sensation with a certain phenomenal character, say a pain,

[22] The only plausible exception is Deonna and Teroni's (2012) proposal according to which emotional attitudes are constituted by bodily changes. Obviously, attitude in this latter sense is completely different from attitude in the functionalist sense. See also footnote 9.

she has that sensation (Kripke 1980, but see Descartes 1641/2019). Alternatively, in a weaker formulation, if it seems to someone that she is sensing, this is enough for her to sense. So, to start with, how could she be wrong not only about the particular phenomenal character of her mood, but on the very fact that such a mood has a phenomenal character altogether?

Reply5: I am not impressed by this objection, which is question-begging in the present context. Again, it is based upon introspection. However, both the phenomenology-free approach and standard representationalism agree that some introspective reports are mistaken. The standard representationalist claims that reports about undirectedness of moods are erroneous. The endorser of the phenomenology-free approach thinks that reports about phenomenology are unreliable, instead. In other words, if any kind of Representationalism is true, we must admit that part of our introspective reports is wrong. The disagreement is about which type of reports is mistaken, and which is right. Assuming the Cartesian Collapse would set the issue in favour of standard representationalism *a priori*, by assuming that reports about phenomenology are more reliable than reports about directedness. But whether this is true is exactly the point at stake.

6. Conclusions

Anti-representationalists have elaborated the argument from moods to falsify Representationalism. A way to answer is to reverse the argument and claim that moods are both undirected and phenomenology-free. This approach is better placed than its main opponent, standard representationalism, in respect to both introspective reports and folk-psychology. So, it is the best approach to moods to adopt for those philosophers with inclinations toward Representationalism.

Moreover, let me show some little additional advantages that have arisen in the discussion, but that I have not explicitly assessed yet. My proposal is indeed able to make sense of other features commonly attributed to moods by philosophers (see Rossi 2019 for an exhaustive list). For example, why we are induced to take the phenomenal character of the emotions as similar of those of moods (see Mendelovici 2013a, 2013b) and why we take moods' "phenomenology" to be unitary and diffuse (Tappolet 2018). It accounts for why emotions "transform" into moods (Deonna and Teroni 2012). If moods functionalism is accepted, other virtues will be gained. It becomes possible to offer a reply to why we "feel" moods as varying in intensity (see Kind 2013), which is an objection to standard representationalism. Functional moods may be tailored to account for moods' arational character, their usage in our folk-psychological explanations (Lormand 1985, Griffiths 1997, Sizer 2000), and their pervasivity (Lormand 1985, Sizer 2000, Chomanski 2017, Tappolet 2018, Rossi 2019).[23]

References

Armon-Jones, C. 1991, *Varieties of Affect*, New York: Harvester Wheatsheaf.

Bordini, D. 2017, "Not in Mood for Intentionalism", *Midwest Studies of Philosophy*, 41, 1, 60-81.

[23] I follow Kriegel (2019) and claim that duration is not a reliable feature of moods. However, it can eventually be accommodated as part of the functional description of moods.

Byrne, A. 2001, "Intentionalism Defended", *Philosophical Review*, 110, 2, 199-240.

Chomanski, B. 2017, "What Makes Up a Mood Experience?", *Journal of Consciousness Studies*, 24, 5-6, 104-27.

Chomanski, B. 2018, "Moods, Colored Lenses, and Emotional Disconnection: A Comment on Gallegos", *Philosophia*, 46, 3, 625-32.

Crane, T. 2007, "Intentionalism", in Beckermann, A. and McLaughlin, B. (eds*.), The Oxford Handbook of the Philosophy of Mind*, Oxford: Oxford University Press, 474-93.

Davitz, J. 1969, *The Language of Emotions*, New York: Academic Press.

DeLancey, C. 2006, "Basic Moods", *Philosophical Psychology* 19, 4, 527-38.

Deonna, J. and Teroni, F. 2012, *The Emotions: A Philosophical Introduction*, London: Routledge.

Descartes, R. 1641/2019, *Meditazioni metafisiche*, in Landucci, S. (a cura di e trad.), Roma-Bari: Laterza.

Drayson, Z. 2012, "The Uses and the Abuses of the Personal/Subpersonal Distinction", *Philosophical Perspectives*, 26, 1, 1-18.

Dretske, F. 1995, *Naturalizing the Mind*, Cambridge, MA: The MIT press.

Eysenck, M.W. and Keane, M.T. 2010, *Cognitive Psychology: A Student's Handbook (6th edition)*, London: Psychology Press.

Frijda, N. 1994, "Varieties of Affects: Emotions and Episodes, Moods and Sentiments", in Ekman, P. and Davidson, J. (eds.), *The Nature of Emotion: Fundamental Questions*, New York: Oxford University Press, 56-67.

Gallegos, F. 2017, "Moods Are Not Colored Lenses: Perceptualism and the Phenomenology of Moods", *Philosophia*, 45, 4, 1497-1513.

Griffiths, P.E. 1997, *What Emotions Really Are*, Chicago: The University of Chicago Press.

Goldie, P. 2000, *Emotions: A Philosophical Exploration*, Oxford: Oxford University Press.

Hatzimoysis, A. 2017, "Representationalism and the Intentionality of Moods", *Philosophia*, 45, 1515-26.

Horgan, T. and Tienson, J. 2002, "The Intentionality of Phenomenology and the Phenomenology of Intentionality", in Chalmers, D. (ed.) *Philosophy of Mind, Classical and Contemporary Readings*, Oxford: Oxford University Press, 520-33.

Kind, A. 2013, "The Case Against Representationalism About Moods", in Kriegel, U. (ed.), *Current Controversies in Philosophy of Mind*. London: Routledge, 158-70.

Kostochka, T. 2020, "Why Moods Change: Their Appropriateness and Connection to Beliefs", *Synthese*, 198, 11399-420.

Kriegel, U. 2017, "Reductive Representationalism and Emotional Phenomenology", *Midwest Studies of Philosophy*, 41, 1, 41-59.

Kriegel, U. 2019, "The Intentional Structure of Moods", *Philosophers' Imprint*, 19, 49, 1-19.

Kripke, S. 1980, *Naming and Necessity*, Cambridge, MA: Harvard University Press.

Lormand, E. 1985, "Toward a Theory of Moods", *Philosophical Studies*, 47, 3, 385-407.

Lormand, E. 1996, "Nonphenomenal Consciousness", *Nous*, 30, 2, 242-61.

Mendelovici, A. 2013a, "Pure Intentionalism about Moods and Emotions", in Kriegel, U. (ed.), *Current Controversies in Philosophy of Mind*, London: Routledge, 171-93.

Mendelovici, A. 2013b, "Intentionalism about Moods", *Thought*, 2, 126-36.

Mitchell, J. 2018a, "The Intentionality and Intelligibility of Moods", *European Journal of Philosophy*, 27,1, 1-18.

Mitchell, J. 2018b, "The Varieties of Mood Intentionality", in Breidenbach, B. and Docherty, T. (eds.), *Mood: Interdisciplinary Perspectives, New Theories*, London: Routledge, 52-69.

Price, C. 2006, "Affect without Object: Moods and Objectless Emotions", *European Journal of Analytic Philosophy*, 2, 1, 49-68.

Rossi, M. 2019, "A Perceptual Theory of Moods", *Synthese*, 198, 7119-47.

Seager, W. 2002, "Emotional Introspection", *Consciousness and Cognition*, 11, 4, 666-87.

Searle, J. 1983, *Intentionality: An Essay in the Philosophy of Mind*, Cambridge: Cambridge University Press.

Sizer, L. 2000, "Toward a Computational Theory of Mood", *The British Journal for Philosophy of Science*, 51, 4, 743-69.

Siewert, C. 2017, "Consciousness and Intentionality", in Zalta, E. (ed.), *The Stanford Encyclopaedia of Philosophy*, (Spring 2017 Edition), https://plato.stanford.edu/archives/spr2017/entries/consciousness-intentionality

Solomon, R. 1993, *The Passions*, London: Hackett.

Stephan, A. 2017, "Moods in Layers", *Philosophia*, 45, 4, 1481-95.

Tappolet, C. 2018, "The Metaphysics of Moods", in Naar, H. and Teroni, F. (eds.), *The Ontology of Emotions*, Cambridge: Cambridge University Press, 169-86.

Tye, M. 1995, *Ten Problems of Consciousness*, Cambridge, MA: The MIT press.

Tye, M. 2008, "The Experience of Emotions: An Intentionalist Theory", *Revue Internationale de Philosophie*, 62, 243, 25-50.

Voltolini, A. 2013, "The Mark of the Mental", *Phenomenology and Mind*, 4, 169-86.

Social Groups and the Problem of Persistence through Change

Giulia Lasagni

Europa-Universität Flensburg

Abstract

The persistence of social groups through change is a matter of debate in social ontology. While mereological approaches contend that social groups persist if formed by the same members, other accounts leaning towards structuralism find that what ensures the persistence of social groups is instead continuity of structure. The aim of this paper is to challenge the idea that a structuralist account is bound to hold that continuity of structure is necessary and sufficient condition for persistence.

First, I consider membership changes. I argue that for structure-based metaphysics, not all changes in membership are irrelevant to persistence because, for some groups, members' continuity is made necessary by structural constraints on the node-occupiers. Then, I discuss structural changes. The main idea is that social groups can persist through structural changes that fall within the group's flexibility margins. I suggest that one way to determine the flexibility margins is to pinpoint the social factors that ground the group's structure. Finally, I raise two open questions concerning how to identify grounds and how to consider their eventual transformation.

Keywords: Social ontology, Social Groups, Persistence, Social Structures, Membership.

1. Introduction

The persistence of social groups through changes in membership and structure is a controversial issue. While mereological approaches contend that social groups persist if formed by the same members (Hawley 2017), other accounts leaning towards structuralism find that what ensures the persistence of social groups is instead continuity of structure (Ritchie 2018, Sheehy 2016). Greenwood (2019) has recently remarked that both views are of limited scope, focused exclusively and respectively on the continuity of membership and the continuity of structure as if membership/structure were necessary and sufficient condition for the persistence of every social group. The issue raises complex metaphysical questions. My

Argumenta 10, 1 (2024): 421—436
ISSN 2465-2334

DOI 10.14275/2465-2334/20220.las
First published: 7 February 2022

aim here is to challenge the idea that a structuralist account necessarily provides a one-sided approach to the persistence of social groups, bound to hold that continuity of structure is necessary and sufficient condition for persistence.

By primarily relying upon Ritchie's structure-based metaphysics (Ritchie 2013, 2015, 2018), I will argue that, if implemented with an explanation of social grounds, the structuralist view has the resources to offer more than a one-sided perspective on persistence, responsive to both membership and structural changes. This proposal aims to find in the social factors that ground the group structure limits within which the structure can change, and the group remains the same.[1]

The paper is organized as follows: Section 2 illustrates the problem of group persistence in social ontology. Then, Section 3 provides an outline of Ritchie's structure-based metaphysics of social groups, focusing on those groups that show some internal organization of the members, such as committees, teams, and music bands. I suggest that the question of persistence is particularly relevant when it comes to organized groups because they are generally recognized as group agents. Hence, asking whether a group is the same before and after a change has metaphysical and ethical implications. Thus, I examine how structured-based metaphysics handles the issue of persistence of organized social groups against changes in membership and structure. My goal is to disprove that the structuralist view is bound to hold that, for an organized social group to persist, the members can easily vary while the structure must be rigid. In Section 4, I contend that for structure-based metaphysics, not all changes in membership are irrelevant to

[1] For my purpose here is limited to discuss through what changes social groups can eventually persist, I do not consider the general metaphysical issue of how social groups persist in time. In fact, it can be assumed that it is one thing to ask through what changes social groups persist while it is another thing to ask what it means for a social group to persist through change, whether by enduring, perduring or exduring (Hawley 2001). Because for each of these theories of persistence it is possible to ask what changes can count as alterations (endurantism), what changes are variations of properties belonging to different temporal slices of the same worm (perdurantism), or variations of properties in a series of counterparts (exdurantism), we may say that each view conceptualizes change in some way (Effingham 2009). Therefore, the discussion here proposed on the kinds of changes affecting groups does not require us to favor one notion of persistence over the others. Nonetheless, for what concerns lexical choices and in order to avoid cumbersome, multiple formulations, a kind of endurantism is in the background of this paper. The first reason for this concerns the notion of change. Unlike perdurantism and exdurantism, endurantism allows for the numerical identity of persisting objects bearing incompatible properties over time (Haslanger 2003b). Changes of this kind are called alterations (cf. Haslanger 1989: 3). As I contend that structuralism can account for the persistence of social groups through (some) changes, my point here is closer to that made by endurantism. The second reason regards the anti-reductionism implied by structuralism, which conceptualizes social groups as structured wholes, materially constituted by organized members, and grounded on social factors. On this view, the group can be reduced neither to the set of members forming it at any moment in time, nor to the sole organization between the parties (social groups in fact instantiate but are not identical to social structures), nor to momentary entities that have members and/or structure essentially (it will be argued that, organized social groups survive changes in membership and structure). Noteworthy, endurantism requires non-reductive explanations treating groups neither as momentary objects in succession nor as fusions of slices, but entities that are wholly present at different moments in time. Arguments in defense of endurantism and against the four-dimensionality of objects are provided by Baker 2007 and Haslanger 1989. On the inconsistencies of endurantist-reductionist accounts, see Wahlberg 2014.

persistence because sometimes members are made necessary by structural constraints. Subsection 4.1 considers how social structures might require specific occupiers. Here I elaborate on Ritchie's assumption that organized social groups are structured wholes constitutively dependent on social factors. I consider social factors as ontological grounds and suggest that, as social factors ground the group's structure, they can also constrain nodes making continuity of membership necessary for group persistence. In Section 5, I discuss how a structuralist approach would deal with structural variations and argue that margins of structural flexibility may be provided at the grounding level. My claim is that, as long as the group's structure varies within such margins, it undergoes alterations, and the group persists. Finally, I raise two open questions concerning how to identify grounds and how to consider their eventual transformation. Despite unresolved issues, I hope to show that structuralist metaphysics, inclusive of grounding relations, can offer more than one-sided approaches to persistence.

2. The Persistence of Social Groups

Many examples show that social groups are subject to change: Succeeding Justice Ruth Bader Ginsburg, Amy Coney Barrett was appointed to the Supreme Court; after a referendum held on 20-21/9/2020, the number of members of the Italian Parliament were reduced from 945 to 600; due to the pandemic, major fashion houses converted part of their manufacturing plants to produce surgical gowns and masks. In all these cases, there is a group that exists before the change and a group that exists after the change: the question of persistence is determining whether the two groups are identical to one another.[2] Here, identity takes on a numerical property as opposed to qualitative property. Meaning that, we will be examining under what changes a social group remains one and the same.[3]

Therefore, this article is concerned with the numerical identity over time of social groups as subject to change. Admittedly, the issue of persistence has a lot to do with how we conceptualize persisting objects. Indeed, different metaphysical theories have offered different conceptions of social groups that can be classified into two factions according to whether they treat social groups as mereological compounds or as structured wholes constituted by material entities (Strohmaier 2018). The faction I consider here to be broadly akin to mereology is an inclusive group encompassing both extensional and non-extensional mereology (Hawley 2017), setism (Effingham 2010), and stage-theory (Wilhelm 2020). Differences aside, theories of this sort identify organized social groups with their members, assuming social groups have members essentially. In contrast, neo-Aristotelian approaches are prone to metaphysical structuralism and individuate social groups by both structure and matter (Fine 2020, Sheehy 2006, Ritchie 2013).[4] In general, the structuralist maintains that if we wanted to identify the group only by its material composition, we would lose sight of the function and the nature (kind) of the group. Besides, we would be faced with some metaphysical puzzles. For instance, it would be difficult to distinguish coinciding, non-identical groups

[2] As my concern is about through what changes (if any) social groups persist, I will assume there is a sense—be it epistemological or ontological—in which social groups are *real*. On realism in social ontology, cf. Laitinen and Schweikard, manuscript.

[3] 'Persistence' is thus meant as synonym of 'numerical identity over time'.

[4] In these pages I refer to the neo-Aristotelian perspective by 'structuralism' and 'structured-based metaphysics'. I use these expressions as synonyms.

or indicate the location of groups based in a certain place and have members scattered elsewhere.[5] Therefore, structure-based metaphysics assumes that the synchronic identity of a group has as necessary and sufficient conditions to have a particular structure and to have the nodes occupied by a particular set of entities, which are the members.[6]

Concerning persistence, mereology suggests that a group persists insofar as there is continuity in membership. In contrast, structuralism argues that for a group to survive "is for its parts to continue to be organized in the relevant object-making fashion, even when those parts may be subject to replacement through time" (Sheehy 2006: 139) inasmuch as "groups can vary in members across times and worlds" (Ritchie 2015: 316).[7]

An example may help visualize the diversity of these approaches: Consider the editorial board of some journal of philosophy and suppose one of the members retires and is replaced by someone else. By concentrating exclusively on the material composition of the group, namely the members, the mereologist would be inclined to say that after the change, the group is no longer the same as before. The structuralist would instead observe that insofar as the change does not affect the group's structure, the group can still be considered the same as before. Everyday experience demonstrates that cases like this are widespread, so part of the appeal of structure-based metaphysics is allowing for a conception of changes in membership that explains many ordinary events of persistence.

Despite the advantages, skepticism towards structuralism has emerged because, by bounding group identity over time to the organization of the parts, the account might have to acknowledge that (1) continuity in membership is neither necessary nor sufficient for group persistence, whereas (2) structural continuity is necessary and sufficient condition. In cases where assumptions (1) and (2) were correct, one would be justified in finding structuralism blind to the salience of membership and latched onto structural rigidity.

In structuralist literature, the case of the persistence of concrete unified wholes is indeed problematic because, in general, such objects are meant to have interchangeable parts and fixed structure. For example, Koslicki observes that

[5] On the metaphysical puzzles and shortcomings of (extensional) mereology, see Effingham 2010, Hawley 2017, Hindriks 2013, Ruben 1985, Strohmaier 2018.

[6] For a definition of organized social groups' synchronic identity, see Ritchie 2018:11.

[7] Among the accounts in support of mereology, some have proposed refined theories of persistence. Specifically, Effingham (2010) has developed a form of setism that views social groups as sets of ordered pairs of which the first member is an instant of time and the second is a set of individuals. This allows Effingham to argue that to ensure persistence, the members of the set of individuals can change across ordered pairs, while the set containing all pairs cannot change its members. Recently, Wilhelm (2020) has proposed considering groups as fusions of group-stages. Each stage is a momentary object that has its members essentially. Different stages can have different members. Persistence in this case has to do with the correlation of stages understood as counterparts. Among neo-Aristotelians, persistence has been recently discussed by Fine (2020) in a way that may not be related to all the structuralist theories at issue in these pages. In order to include form (structure) and matter (members) into the metaphysics of social groups, Fine suggests applying the notions of rigid and variable embodiment. Rigid embodiment refers to synchronic group identity by combining the component parts into a structured whole. Variable embodiment concerns persistence: The operation accounts for actual or possible change in the constitution (i.e., form and matter) of the group.

> […] unlike mereological sums, not only are these objects quite obviously capable of surviving changes with respect to their parts, while mereological sums (like sets) have their parts essentially; but, in contrast to the completely unstructured nature of mereological sums, the existence and identity of these objects is also evidently tied to the arrangement or configuration of their parts (Koslicki 2018: 2).

Turning to apply this view to social groups as concrete objects, it would seem that (1) any social group can stay numerically the same despite changes in membership whilst (2) it does not survive changes in the relations among them.

My goal here is to argue that both (1) and (2) are misleading assumptions. To do so, I will present Ritchie's structuralist metaphysics of social groups, in which the problem of persistence has not yet been investigated in depth. I will show how such a metaphysical framework is open to being integrated with explanations regarding grounding relations and then proves responsive to certain membership and structural variations.

3. Ritchie's Structure-Based Metaphysics

To determine if a social group ever survives change and when survival may occur for structure-based metaphysics, it is worth clarifying the meaning of the generic concept 'social group'. Universities, business companies, families, soccer teams, working classes, and religious communities are just a few examples of the sort of entities generally counted as social groups. In an attempt to subsume such variety within a few inclusive categories, Ritchie (2013, 2015) has proposed to divide social groups into two types: Type 1 denoting organized groups and Type 2 applicable to groups clustered around at least one attribute the members have in common.[8] Critical against the idea that a simple framework can be adequate to capture the complexity of social groups, Epstein (2017) has instead offered several criteria for establishing the most suitable metaphysical profile for each group. Importantly, given that the concept 'social group' applies to heterogeneous contexts, it is possible that if some social groups persist through change, it is not certain that they all would persist and that they would all persist in response to the same changes. Here, for the sake of simplicity, I resolve to focus only on organized social groups (Ritchie's Type 1) like committees, bands, and sports teams and investigate through what changes—according to structuralism—groups of this kind persist. In addition to being a starting point for metaphysical inquiry, the question of organized groups' persistence has ethical relevance because these are the groups that are generally accorded agency abilities (List and Pettit 2011). Thus, knowing whether a group is numerically the same before and after a change also helps us determine whether the group in the present is responsible for an action completed in the past.

Let us focus on the metaphysical question. On Richie's view (Ritchie 2018), organized social groups are structured wholes, i.e., social structures realized by sets of entities. More precisely, structures are networks of relations connecting the positions (nodes) and establishing the role of each node in the entire relational complex.[9]

[8] Further classifications for social groups can be found in French 1984, Gilbert 1989, Gruner 1976, List and Pettit 2011, Tuomela 2007, Young 1990.

[9] Social structures shape various types of social facts or objects such as the market and the transportation system. Organized social groups are special social objects because they have only individuals or groups as node occupiers.

Entities that occupy a position within the structure are members.[10] What entities may serve as members is specified by (eventual) structural restrictions defining and constraining each position. Regarding structured entities in general, Koslicki clarifies that structures make available slots for objects that meet two sorts of constraints: "(i) constraints concerning the type of object which may occupy the position in question; and (ii) constraints concerning the configuration or arrangement which must be exhibited by the occupants of the positions made available by the structure" (Koslicki 2018: 3). Meaning that, the structural relations affect both the type of object suitable to occupy some node and the overall organization of nodes and node-occupiers.

In the case of organized social groups, examples of structures are the patterns of relations fixing the players' roles in a baseball team, kinship ties in a family, and the system of offices shaping an institutional organ. Relationships can be of various types: they can be symmetrical (being married to) or asymmetrical (being mother of), hierarchical (being the leader of) or non-hierarchical (being partners), intentional (being wife of), or unintentional (being son of). The network of relations defines the role of each party in relation to the others and incorporates the function of the entire group. Importantly, social structures can be multiply realized: insofar as the requirements of the nodes are met, various sets of entities can realize the same structure.[11] This implies social structures shape but are not identical to social groups, as social groups are specific realizations of social structures. Therefore, an organized social group is normally made up of members who are organized based on some specific relational pattern.

Let us now return to the question of persistence and ask how we can approach the subject of change through the structuralist framework. The question is: Within structure-based metaphysics, what changes (if any) are organized social groups meant to survive? There are two cases that we propose to analyze: changes in membership and changes in structure. The reason for this choice is that for structuralist metaphysics, membership and structure are necessary and sufficient conditions for the synchronic identity of organized social groups. As this does not imply that such conditions play the same role for diachronic identity, my goal is to understand whether and when changes in membership and structure are relevant for persistence. In doing so, I will reject the position that deems member continuity irrelevant and structural continuity necessary and sufficient for persistence.

4. Change in Membership

First, consider membership: Experience proves that members of organized groups are replaceable in most cases, just as the editorial board's members discussed above were replaceable. Soccer teams change players; political parties change components; companies hire and fire people. Regarding such circumstances, a structuralist metaphysics observes that as long as the group's structure remains the same, the social group remains the same. Some have noted that considering structural continuity necessary and sufficient for persistence risks blinding the

[10] By virtue of being a functional status, membership is not the same as parthood. For example, Supreme Court Judge *x* is a member of the Supreme Court whereas *x*'s arm is not. The case shows that, as opposed to parthood, membership is not transitive. See Uzquiano 2004.

[11] An illustrative case is analyzed in Uzquiano 2004, in which it is argued that the Supreme Court is not the same as any specific set of Supreme Court Justices.

account to cases where membership continuity appears necessary (Greenwood 2019). Take the example of the band 'Florence and The Machine': The group consists of the vocalist, Florence Welch, and the musicians, currently keyboardist Isabella Summers, guitarist Rob Ackroyd, harpist Tom Monger. Since its formation in 2007, Florence has always been a member (specifically, the singer) of the group while the musicians have changed—for example, in 2018 drummer Christopher Hayden left the band, and new collaborations were started. Despite changes to membership, fans seem to regard the group as the same as before 2018, and this acknowledgment confirms the structuralist thesis that members of organized groups are replaceable and therefore irrelevant to group identity over time.

But what would have happened if the person leaving the group was Florence? Intuitively, as a fan, it seems reasonable to assume that the band would no longer be the same without its lead vocalist. By contrast, structuralism would not seem to validate intuition as it does not consider groups to have members necessarily.

It may be that a fan's intuition does not suffice as a philosophical argument, but it indeed urges us to delve into the topic of membership changes, for which Ritchie's structuralism suggests a solution. Although the continuity of membership in general and in itself is not a necessary and sufficient condition for persistence, it may be that some members are necessary for group identity in some cases, in conjunction with structural constraints. As mentioned already in Section 2, the possibility for understanding group persistence lies in the understanding of every position in a structure being defined by the relations between nodes and eventual restrictions on the node-occupier (cf., Koslicki 2018: 3, Ritchie 2018: 7). Because the latticework of relations fixes the characteristics required by each node, if the node requires a specific person to occupy it, then the presence of that member will be necessary to the identity of the group: "As a limiting case a node might require that it be occupied by a particular person. For instance, if bands are structured wholes, some band structures might require that specific individuals occupy particular nodes" (Ritchie 2018: 10). If the requirements on a node-occupier are instead neutral regarding the person covering the position, the occupier can change without implications for group persistence.

As structuralism provides that sometimes, and based on structural features, specific members are necessary for group persistence; the eventual rigidity of membership cannot be used as a source of counterexamples to the account. This would explain how, although the musicians of 'Florence and the Machine' have changed without affecting the group's identity over time, the replacement of the vocalist would have probably created discontinuity.

Further examples of members that are made necessary by structural constraints are highly personalistic groups in which the restricted node is often that of the leader (e.g., perhaps the political party 'Forza Italia' and its leader, Berlusconi) and by creative groups in which the originality and style of some, eventually all, members are central to group identity (The Beatles).

4.1 Constitutive Dependence

The argument presented so far makes a point in support of structuralism, remarking that sometimes groups have specific members necessarily by virtue of their ontological structure. Now, we must explain how it is that social structures eventually impose constraints on the nodes. In this regard, Ritchie's theory offers an interpretation of the metaphysical foundation of social structures that is decisive

both to understand better how group structures can make continuity of membership necessary and envision the possibility of persistence in the face of structural variations. The argument rests on the constitutive view of social reality, according to which social structures constitutively depend on social factors like practices, attitudes, and norms.[12] Following Haslanger (2003a), Ritchie holds that only structures that depend constitutively on social factors are social.

By 'constitutive dependence', Ritchie means a relation between some social structure S and more fundamental social factors:

> Structure, S, constitutively depends on social factors just in case
> (i) in defining what it is to be S reference must be made to some social factors or
> (ii) social factors are metaphysically necessary for S to exist or
> (iii) social factors ground the existence of S (or the fact that S exists)
> (Ritchie 2018: 6).[13]

Ritchie's definition of constitution can be rephrased (though not necessarily) in terms of grounding relation (iii); I will treat constitutive social factors as the metaphysical grounds of social structures.[14]

So, assuming the structure of 'Florence and The Machine' is constitutively dependent on social factors, to grasp the grounds of the group's structure (including membership constraints) is to specify what social factors ground the structure of the group. The list of social factors might encompass elements like social practices, habits, beliefs, intentions, agreements, and action patterns. According to Ritchie, social factors can be internal or external to the group. Internal social factors concern the node-occupiers just like intentions and agreements among the members; external social factors concern external facts such as norms, institutions, and non-members. In most cases, more than a single social factor contributes to the foundation of a social structure. To say that the structure of the band is partly or fully grounded on some internal social factor(s), such as the agreement between the members, means that the factor contributes to the construction of the group's structure, i.e., the group is shaped the way it is partly or fully due to the members' agreement. Similarly, to assume that the record contract partly or fully grounds the group's structure is to hold that the social factor lays the foundations for the existence of the group's structure.

Whenever the grounds set up a social structure, as with 'Florence and The Machine', any set of entities realizing that structure will form the band 'Florence and the Machine'. Moreover, if the structure is grounded in the band's vocalist

[12] In acknowledging that Ritchie does not provide any definition of what social factors could be, I will use the notion as she does, that is, in a general way. Here are two lists that Ritchie offers in different parts of the article (2018) for the purpose of illustrating some examples of social factors: "social behavior, patterns of action, habits, beliefs, intentions, processes, practices, activities, rules, laws, norms, and arrangements" (3); "social practices, patterns of interaction, agreements, beliefs, and so on" (15).

[13] Constitutive dependence is a form of non-causal dependence (Diaz-Leon 2013) and can be understood either theoretically or metaphysically. Ritchie's definition holds together in disjunction both the theoretical notion of constitutive dependence (Audi 2012, Haslanger 2003a) and the metaphysical notions of necessity and grounding relation (Griffith 2018): For constitutive dependence to occur, it is sufficient that one of the three disjuncts applies.

[14] On grounding relations and constitutive construction, see Griffith 2018.

being Florence, then Florence will have to be a necessary part of any set of entities realizing the band's structure.

5. Persistence through Structural Change

Section 4 has demonstrated that a structuralist account has the means to recognize that continuity of membership is occasionally a necessary condition for persistence. Now, the question must be asked whether, according to structuralist metaphysics, continuity in structure is a necessary and sufficient condition for the persistence of organized social groups. By establishing that continuity of membership is also required in certain cases, we have already proved that structural invariance is not always sufficient:[15] necessity must now be considered.

The issue is of paramount importance. In fact, if structural fixity were to be considered a necessary condition of persistence, many social groups that we generally regard as surviving (at least some) change in structure should instead be regarded as non-persistent. Examples of ordinary structural changes include shuffling tasks and shares among the members of a group (as is often the case among a company's shareholders), modification in the group's function (as when the responsibilities of a police department are extended) or functioning (as exemplified by eventual adjustments in the decision-making procedure of a committee), increase or decrease in the number of nodes (as happens whenever a family welcomes a new child).

Take the standard case of a committee switching from majority voting to unanimity. Undoubtedly, this is a change regarding the functional organization of the parties, and thus, the structure of the group.[16] Therefore, if structuralism assumed structural continuity was provided by group identity over time, we would have to conclude that, after the change, the committee would no longer be the same. For a group to survive this change, it would mean "for its parts to continue to be organized in the relevant object-making fashion, even when those parts may be subject to replacement through time" (Sheehy 2006: 139). Along these lines, Ritchie has taken the identity over time of organized social groups to be based on structural continuity—whereas "groups can vary in members across times and worlds" (Ritchie 2015: 316). However, in Ritchie 2018, we read that

[15] One might object that, on the account I propose, continuity of structure is indeed a sufficient condition for persistence because membership is subsumed by structural continuity. Although I take membership conditions (if any) to be specified in the structure of the group—and, in this sense, they might be regarded as structural aspects—I would not go so far as to say that structural continuity is a sufficient condition for group persistence. In fact, while retaining the same structure and thus the same membership conditions, a group can still vary its material composition. Further, only a few material components are suitable for the realization of the group structure. It can be that, at time t, group structure s is realized by a set of individuals satisfying s' membership conditions, while, at time t', s is realized by a set of individuals that fails to meet such requirements. Thus, the group at time t' is not the same as the group at time t. Cases like Florence and the Machine prove that sometimes structural continuity is not sufficient for group persistence and illustrate to what extent material continuity is also necessary, although it is made so by structural constraints.

[16] Since by 'group structure' I mean the network of relationships that define and connect the nodes, any change in the pattern of interaction is to be regarded as a structural change. So, a modification in the decision-making procedure counts as a structural change because it affects the relationships among the nodes and the operations required from each (or some) of them.

"the [structure-based] view allows for groups to persist through changes in members and through changes in structure" (Ritchie 2018: 11). Thus, one wonders to what extent *the relevant object-making fashion* is a flexible parameter and what changes in structure (if any) are compatible with the persistence of social groups. The issue is an open question as Ritchie has not offered any explanation for that.[17]

In what follows, I will analyze cases of structural change by considering the group's ontological grounds. The aim is to show that, by complementing the organization of the nodes with explanations about the grounds, it is possible to delineate margins of structural flexibility within which the structure can change, and the group can persist. In other words, the investigation is meant to demonstrate that one way to assess whether organized social groups survive at least some structural changes is to consider the grounds, i.e., the social factors, on which the group's structure constitutively depends.[18] In fact, if it is true that social structures are metaphysically dependent on social factors, then it may be that what social structures are— the way they are shaped— is not necessarily rigid but may have margins of flexibility consistent with their nature, i.e., consistent with their being structures grounded on some specific social factors. More specifically, there may be cases where certain structural variations are equally compatible with the set of social factors that ground the structure.

Let us now return to the case of a committee switching from majority voting to unanimity. The group is responsible for deciding the winner of a competition. First, consider the scenario in which the committee's structure is grounded in a charter that explicitly stipulates that the unanimity voting system is not allowed. This implies that the organizational structure of the parties cannot incorporate an unanimity-based decision-making procedure. The group members might still have the shared intention to change the voting system from majority to unanimity, but the implementation of such a change would lead to structural modifications not conceded by the group's structure. Unless the charter enables the members to change the rules, that ability to change the structure is not available to them.[19] An eventual change from majority voting to unanimous voting would therefore contradict the group's foundations and likely incur sanctions. Most importantly, the event would be a change which—based on the structuralist framework—the committee could not manifest through persistence because a structure that includes a unanimous voting mode would be contradictory to the kind of structure that has the charter as its ground. A group that votes by unanimity at time t' will therefore be numerically non-identical to the group that at time t voted by majority.

The case would be different if the charter grounding the group's structure establishes that decision-making can happen either by majority or unanimity. In this scenario, inscribed in the group's structure is the possibility of specific

[17] Rather than arguments, Ritchie offers hypotheses: "The view allows for groups to persist through changes in members and through changes in structure. Causal origin plausibly figures in the persistence conditions of organized groups. A theory of the persistence of organized groups might also involve member intentions and the intentions of authoritative non-members. Other conditions might vary widely across organized group types. The view sketched here could be developed in various ways for different sorts of organized groups and according to one's general views of persistence" (Ritchie 2018: 11).

[18] I mentioned that the study of grounding relations is one way of approaching persistence because explanations of other kinds, especially causal explanations, could direct us onto equally promising tracks.

[19] On the abilities (powers) of organized social groups and members, see Hindriks 2008.

structural changes that are established by the charter: Switching from majority to unanimity would be a change that alters the metaphysical structure of the group as designed at the grounding level. Meaning, the committee is grounded on a charter that provides the social structure. It would not matter if the committee adopted a system of majority or unanimity voting because both arrangements are consistent with the kind of structure set by the grounds. For this reason, we might argue that switching the procedure is a change that the committee can survive according to structure-based metaphysics. For a variation of this scenario, imagine that the charter leaves it up to the members to determine which decision-making mechanism is the most suitable from time to time. In this case, although the list of possible options is not explicitly provided, any variation made by the members in this area would be consistent with the grounds and would therefore represent a structural alteration through which the group persists.

Let us now consider a slightly different case: the election of a spokesperson. We will assume that the charter does not describe any node as a spokesperson and that the introduction of this role is members' initiative. As the structure is grounded in a way that makes it indifferent to have a spokesperson, its election is just an unexpected change. The issue here concerns whether structuralism could ever view organized social groups as surviving unexpected structural changes.

Presumably, groups do not survive unexpected changes when these contradict the grounds. The issue is complicated especially when, in the context of a structure with multiple grounds, some change is neutral, relative to the structural features fixed by one ground but contradictory relative to some other(s). For example, a committee could be based on a charter that sets the decision-making mechanism (ground 1) and an external authority that decides the appointments (ground 2). Having the members appoint a spokesperson could be neutral to ground 1 though at odds with ground 2. By contrast, if there were no grounding relation aimed at excluding the possibility for the members to designate a spokesperson, we might consider this structural change to be an alteration of a persisting group.

In general, the analysis illustrates that the question of persistence through structural changes must be assessed on a case-by-case basis, in consideration of the grounds of the group's structure.[20] Specifically, insofar as a change implies structural variations consistent with the spectrum of structural flexibility provided by the grounds, the group undergoing such change can legitimately be considered the same group before and after the change. If the change is an event that does not fall within the spectrum of possibilities for that structure, the group that exists

[20] Each social group has unique relationships with the social context. This causes each social group to have conditions of persistence that are specific and not entirely generalizable. We might admit that groups of the same kind have minimal conditions of identity over time related to the kind; however, it is likely that those conditions can be realized differently depending on the group. Thus, to provide an accurate explanation of group persistence, it might be worth implementing the metaphysical analysis of the kind with an empirical investigation of concrete particulars. For example, it is plausible to think that every institution is grounded on some statute. Also, we can assume that among institutions, all graduation committees in Italian Universities are based on the same bylaws. Yet, any university may present specificities or some additional, internal regulation supplementing the national one. The suggestion here is that if we want to discuss the persistence of a specific committee, we can get oriented by first considering the grounds that generally characterize the kind 'graduation committee'. A complete analysis will then require us to specify the grounds of the concrete group.

after the change is not the same as the one that existed before. Regarding unexpected changes, we can say that if such changes lead to structural arrangements consistent with the grounds, then organized social groups subjected to such changes will persist through them.

In brief, adopting a form of structuralism that upholds a constitutive view about social structures allows us to argue that organized social groups can persist through structural changes that fall within the group's flexibility margins. I have argued that one way to determine the flexibility margins is to pinpoint the foundation of the group's structure. On this basis, structuralism cannot hold that structural continuity is a necessary condition for the persistence of every social group.

This account gives us the means to assess the persistence of organized groups that commonly undergo structural changes, such as those mentioned at the beginning of this section. First, we observed that companies often undergo changes in the distribution of shares. Now, we can safely assert that, by considering the bylaws of a specific company, we may be able to tell whether the group survives such changes. Presumably, the bylaws will contain a regulation for those acts in conjunction with a description of the admissible procedures for implementing the reshuffling. Then, the study of the grounds might allow us to determine whether an expansion of the responsibilities of a city's police department to surrounding geographic areas makes the group not the same department as before the expansion. The statute that grounds the institution, along with its departments, may indeed clarify the point. In addition, we can apply the explanation of the grounds to the persistence of organized social groups subject to change in the number of nodes. Consider the Italian Parliament, which has recently undergone a 345-unit cut. Since the reduction and the enacted procedures are compatible with the Constitution, we can conclude that the change has been an alteration of a persisting group.[21] As for the addition of nodes, take the case of a nuclear family, which at time t consists of two adults—married to each other—and one child, their daughter. Assume that at time t' the family acquires a new member with the second child's birth. According to the explanation of the grounds, we may assume that the family is the same before and after the newborn, if and only if, it is contained in the definition of a nuclear family that the number of children can vary. And insofar as this is the case, the family persists despite the structural alteration.[22]

6. Conclusions and Open Problems

From the considerations made so far, we can conclude that if there are organized social groups (and we assumed there are some) and if some of them persist through change (and we assumed there are such cases), deciding what changes are compatible with the persistence of the group requires an explanation attentive to both members and structure which also considers the ontological grounds of the social structure realized by the group. I have argued that a structure-based metaphysics of social groups is fit for this purpose.[23]

[21] The compatibility of the reduction of MPs with the Constitution is enshrined in the Constitutional Law No. 1 dated October 19, 2020, which includes amendments to articles 56, 57 and 59.

[22] Further examples can be found in Fine 2020.

[23] Some might wonder why in addressing the issue of persistence in relation to social groups I have not mentioned theories of personal identity over time (Olson 1997, Parfit 1971,

As crucial as this remark is for developing a theory for the persistence of organized social groups, many aspects are in need of clarification. In this last section, I will concentrate on two issues: the identification problem and the transformation problem.

The identification problem is concerned with determining the ontological grounds of a specific social structure. The task is difficult because while some grounds are explicit and institutionalized, such as the record contract and the charter mentioned in our examples, others may be implicit and uncoded, like beliefs and habits.

The class of social factors is highly heterogeneous, but most of the time, social structures are also based on a multiplicity of social factors. For example, it can be the case that in addition to being partly grounded in the record contract, the structure of a band is built on the intentions, emotions, and behaviors of the fans who listen to and buy the band's music.

The identification problem might be approached by shifting the focus from grounding relations to what Epstein has called anchoring relations, that is, relations that determine why social structures are grounded the way they are (Epstein 2014, 2015, 2016). Although grounding and anchoring are different from each other from a metaphysical point of view, investigating the anchors of a certain structure might shed light or give us a criterion for establishing what social factors to include among the grounds. For example, suppose it is convention that social factors like charters and statutes constitute the structure of a committee. Accordingly, considering occasional individual intentions as grounds would be unfitting.[24]

In addition to the identification problem, it is worth mentioning the transformation problem, which emerges from the fact that some social factors change over time. Some cases are straightforward, having a standardized procedure for amending the grounds. An example is when the statute of a company (ground) is modified through the unanimous decision of the members (anchor): assuming consensus is what makes the statute the foundation of the group's structure, unanimous decisions might fittingly arrange the modification of the statute.[25] In other

Rovane 1998, Schroer and Schroer 2014). Although I do not intend to rule out this possibility, my concern is that the analogy with personal identity would require restricting the discussion to those groups that can be qualified as persons. In the literature on social groups, it is generally accepted that organized groups with abilities for decision-making, reasoning, and agency are good candidates to be considered as agents or performative persons (List and Pettit 2011). Analyzing the eventual similarities is beyond the scope of this article. In addition, further developments could consider how the explanation of group persistence differs from theories of persisting individuals. Indeed, the former would require (at least in some cases) an empirical investigation into the social world that might not be so decisive in adjudicating the case of individuals' persistence.

[24] Conventions are only one type of anchor for charters and statutes. According to Epstein (2014, 2015), social facts of the same type can be anchored by social factors of different types. Comments on anchoring pluralism can be found in Guala 2017.

[25] Grounding social factors changing in compliance with the anchors resemble the problem of structural changes inscribed in the grounds so that we could suggest a similar approach: As long as the change is expected or just consistent with the anchor, then the change is such that the grounds constitute a structure, which, once realized by a set of members, generates a group that is the same as the one existing before the change at the grounding level. Contrariwise, if the change is prevented by or inconsistent with the anchor, then the

cases, change is the result of slow, unintentional, and unforeseen transformative processes affecting the grounds. Although the grounds are often repeated and stabilized social factors, it may be the case that they change in exactly the way they have become stable, that is, through social interaction (cf., Griffith 2018: 395). For example, collective beliefs and practices grounding (at least in part) the structure of organized social groups like families may change over time and thus impose variations on the social structure of the respective group: Sometimes families change their structure as a result of some transformation in the relationship between the partners, who may, for example, separate but remain legally married.[26] Because interpersonal relationships are intrinsically fluid, their evolution over time is a process that could hardly be crystallized into some code.

The identification and transformation problem reveal complex issues that deserve further investigation. Concerning persistence, providing deeper explanations of grounding relations and social factors might serve to clarify the range of changes through which social groups persist. Moreover, acknowledging identification and transformation questions shows how articulate the metaphysics of social groups is and how deep-rooted social structures are in the social context. While not being full-fledged arguments in favor of structuralism, these considerations seem, at least, to call for an anti-reductionist social ontology ready to analyze organized social groups for their being complex entities tied to the social context in which they are located and rooted.

References

Audi, P. 2012, "Grounding: Toward a Theory of the 'in-virtue-of' Relation", *The Journal of Philosophy*, 109, 12, 685-711.

Baker, L.R. 2007, *The Metaphysics of Everyday Life: An Essay in Practical Realism*, New York: Cambridge University Press.

Diaz-Leon, E. 2013, "What Is Social Construction?", *European Journal of Philosophy*, 23, 4, 1137-52.

Effingham, N. 2009, "Persistence and Identity", in Le Poidevin, R. (ed.), *The Routledge Companion to Metaphysics*, London: Routledge, 296-309.

Effingham, N. 2010, "The Metaphysics of Groups", *Philosophical Studies*, 149, 21-67.

Epstein, B. 2014, "How Many Kinds of Glue Hold the Social World Together?", in Gallotti, M. and Michael, J. (eds.), *Social Ontology and Social Cognition*, Dordrecht: Springer.

Epstein, B. 2015, *The Ant Trap*, New York: Oxford University Press.

Epstein, B. 2016, "A Framework for Social Ontology", *Philosophy of the Social Sciences*, 46, 2, 147-67.

Epstein, B. 2017, "What Are Social Groups? Their Metaphysics and How to Classify Them", *Synthese*, DOI: 10.1007/s11229-017-1387-y.

Fine, K. 2020, "The Identity of Social Groups", *Metaphysics*, 3, 1, 81-91.

social structure—and so the group—will not persist through the modification of the grounds.

[26] Frequently, phenomena of transformation are not intentionally planned, they just happen. The study of grounds can serve the purpose of ameliorating social structures by operating on underlying practices and norms (Haslanger 2012).

French, P.A. 1984, *Collective and Corporate Responsibility*, New York: Columbia University Press.

Gilbert, M. 1989, *On Social Facts*, Princeton: Princeton University Press.

Greenwood, J.D. 2019, "On the Persistence of Social Groups", *Philosophy of the Social Sciences*, 50, 1, DOI: 10.1177/0048393119881098.

Griffith, M. 2018, "Social Construction and Grounding", *Philosophy and Phenomenological Research*, 97, 2, 393-409.

Gruner, R. 1976, "On the Action of Social Groups", *Inquiry*, 19, 443-54.

Guala, F. 2017, "Un'ontologia sociale pluralista: Brian Epstein su 'Anchors' e 'Grounds'", *Iride*, 80, 216-22.

Haslanger, S. 1989, "Persistence, Change, and Explanation", *Philosophical Studies*, 56, 1-28.

Haslanger, S. 2003a, "Social Construction: The "Debunking" Project", in Schmitt, F. (ed.), *Socializing Metaphysics: The Nature of Social Reality*, Lanham, MD: Rowman & Littlefield, 301-25.

Haslanger, S. 2003b, "Persistence through Time", in Loux, M.J. and Zimmerman, D.W. (eds.), *The Oxford Handbook of Metaphysics*, Oxford: Oxford University Press, 315-54.

Haslanger, S. 2012, *Resisting Reality: Social Construction and Social Critique*, New York: Oxford University Press.

Hawley, K. 2001, *How Things Persist*, Oxford: Clarendon Press.

Hawley, K. 2017, "Social Mereology", *Journal of the American Philosophical Association*, 3, 4, 395-411.

Hindriks, F. 2008, "The Status Account of Corporate Agents", in Schmid, H.B., Schulte-Hostemann, K. and Psarros, N. (eds.), *Concepts of Sharedness: Essays on Collective Intentionality*, Heusenstamm: Ontos, 119-44.

Hindriks, F. 2013, "The Location Problem in Social Ontology", *Synthese*, 190, 413-37.

Koslicki, K. 2008, *The Structure of Objects*, Oxford: Oxford University Press.

Koslicki, K. 2018, "Structure", in Burkhardt, H., Seibt, J. and Imaguire, G. (eds.), *Handbook of Mereology*, München: Philosophia Verlag.

Laitinen, A. and Schweikard, D.P., "Many Realisms in Social Ontology", manuscript.

List, C. and Pettit, P. 2011, *Group Agency: The Possibility, Design, and Status of Corporate Agents*, Oxford: Oxford University Press.

Olson, E.T. 1997, *The Human Animal: Personal Identity without Psychology*, New York: Oxford University Press.

Parfit, D. 1971, "Personal Identity", *The Philosophical Review*, 80, 1, 3-27.

Ritchie, K. 2013, "What Are Groups?", *Philosophical Studies*, 166, 2, 257-72.

Richie, K. 2015, "The Metaphysics of Social Groups", *Philosophical Compass*, 10, 310-21.

Ritchie, K. 2018, "Social Structures and the Ontology of Social Groups", *Philosophy and Phenomenological Research*, DOI: 10.1111/phpr.12555.

Rovane, C. 1998, *The Bounds of Agency: An Essay in Revisionary Metaphysics*, Princeton: Princeton University Press.

Ruben, D.-H. 1985, *The Metaphysics of the Social World*, London, Boston: Routledge & Kegan Paul.

Schroer, J.W. and Schroer, R. 2014, "Getting the Story Right: A Reductionist Narrative Account of Personal Identity", *Philosophical Studies*, 171, 445-69.

Sheehy, P. 2006, "Sharing Space: The Synchronic Identity of Social Groups", *Philosophy of the Social Sciences*, 36, 2, 131-48.

Sheehy, P. 2016, *The Reality of Social Groups*, New York: Routledge.

Strohmaier, D. 2018, "Group Membership and Parthood", *Journal of Social Ontology*, 4, 2, 121-35.

Tuomela, R. 2007, *The Philosophy of Sociality: The Shared Point of View*, Oxford: Oxford University Press.

Uzquiano, G. 2004, "The Supreme Court and the Supreme Court Justices: A Metaphysical Puzzle", *Noûs*, 38, 1, 135-53.

Wahlberg, T.H. 2014, "Institutional Objects, Reductionism and Theories of Persistence", *Dialectica*, 68, 4, 525-62.

Wilhelm, I. 2020, "The Stage Theory of Groups", *Australasian Journal of Philosophy*, DOI: 10.1080/00048402.2019.1699587.

Young, I.M. 1990, *Justice and the Politics of Difference*, Princeton: Princeton University Press.

Our Admiration for Exemplars and the Impartial Spectator Perspective: Moral Exemplarism and Adam Smith's *Theory of Moral Sentiments*

Karsten R. Stueber

College of the Holy Cross

Abstract

This essay will discuss the philosophical viability of Linda Zagzebski's refreshingly radical theory of moral exemplarism that attempts to elucidate the nature of human morality through an analysis of the structure of our admiration for morally exemplary individuals. After raising some systematic worries about exemplarism, I will turn to Adam Smith and his *Theory of Moral Sentiments*. There are indeed strands in Smith's thoughts that contain an exemplarist flavor. Nevertheless, from the Smithian perspective that I favor, our moral concepts emerge from the everyday practice of holding each other morally accountable through empathic perspective-taking. Such a practice is prior to our admiration for the exemplary person. It takes place in the domain of the "ordinary and vulgar", that is, in the domain of the butcher, the brewer, and the baker. Moreover, our normative commitment to the impartial spectator perspective can be revealed as a regulative ideal only in light of an analysis of such practices. Ultimately, what is truly admirable is tied to our commitment to the impartial spectator perspective, whose normative authority should be established independently of our urge to admire, or at least so I am inclined to argue.

Keywords: Exemplarism, Empathy, Admiration, Adam Smith, Impartial spectator.

> Moreover, worse service cannot be rendered morality than an attempt be made to derive it from examples. For every example present to me must itself first be judged according to principles of morality in order to see whether it is fit to serve as an original example, i.e., as a model. But in no way can it authoritatively furnish the concept of morality. Even the Holy One of the gospel must first be compared with our ideal of moral perfection before he is recognized as such.
>
> (Kant 1981: 4, 408)

Argumenta 10, 1 (2024): 437—451
ISSN 2465-2334

DOI 10.14275/2465-2334/20230.stu
First published: 20 April 2023

1. Introduction

Throughout history we have admired the works of exceptionally talented people in the arts, the sciences, the humanities, and even in sports. We are also in awe of the extraordinary deeds by ordinary people—such as a policeman sacrificing his life in trying to save a drowning child—the lifestyle of the rich and powerful, or the perceived accomplishments of our political leaders. We admire these individuals because aspects of their lives exemplify features that we hold dear, that we value, and that are part of our ideals in light of which we orient and regulate our own lives. To recognize the socially and morally beneficial nature of such admiration we only need to think about our reverence for inspirational figures such as Mahatma Gandhi, Nelson Mandela, Martin Luther King, and Abraham Lincoln, whose morally exemplary behavior in the fight for justice led to a variety of mass movements and bent the arch of the moral universe towards justice.

Linda Zagzebski has used these intuitions to develop a refreshingly radical theory of moral exemplarism. She claims that it is best to elucidate the nature of human morality by focusing our philosophical attention on the structure of our admiration for morally exemplary persons. Zagzebski also poses a direct challenge to the above Kantian epigraph by turning it on its head. It is not through prior familiarity with moral concepts that we recognize the moral worth of exemplars. Rather it is by being admiringly attuned to them that our moral concepts get content and gain a motivational and normative hold on our agency. As it is well known, Kant's account of morality is often regarded to fall short of answering the question of why it is that moral commands possess a special normative authority and why our recognition of such authority motivates us to act. Kant himself seems to answer these questions by appealing to a mysterious noumenal realm. He thereby violates the widely accepted framework of naturalism according to which the philosophical explication of basic metaphysical, epistemic, and moral features of our lives and the world must be compatible with what the sciences tell us about human nature and the natural world. In emphasizing the emotion of admiration, Zagzebski is more aligned with the ethical and meta-ethical framework proposed by moral sentimentalists who emphasize that moral concepts are in some sense anchored in our emotional reactivity to each other and to the world rather than being grounded in pure reason (Debes and Stueber 2017: Introduction). Like the moral sentimentalists, Zagzebski is open to insights from the empirical sciences and welcomes an empirical investigation of moral agency.

Yet, regardless of how one thinks about the plausibility of Kant's moral philosophy, the epigraph raises a serious question any exemplarist position must answer, that is, how can our admiration for exemplars ground our moral practices if we can identify exemplars only because of a prior understanding of moral concepts. Moreover, the emotion of admiration is a rather double-edged sword since it also has its dark sides morally speaking. We admire persons for all kinds of reasons ranging from rather mundane traits, such as physical prowess, fame, money to intellectually inspiring and morally elevating features such as amazing historical knowledge, oratory skills, or unexpected generosity, integrity, or courage. As Adam Smith already pointed out, admiration is certainly an emotion necessary for the cohesion of society helping us to "maintain the distinction of ranks and the order of society". Yet he also was wary of admiration for the "rich and powerful" since it constitutes "the great and most universal cause of the corruption of our moral sentiments" (Smith T.M.S. 1976: 61; Irwin 2015). Here one need only think of the

contribution that the admiration for Hitler made in bringing about the catastrophe of World War II and the Holocaust. Currently, the admiration for people like Putin and Erdogan prop up autocracies all over the world. Closer to home, one could argue that the admiration for somebody like Trump constitutes a serious threat endangering the very foundation of American democracy.

In the following, I will critically discuss Zagzebski's exemplarism and investigate whether she can meet the above challenges. In the first section, I will briefly outline the structure of her exemplarist position. In the second section, I will raise three systematic worries about exemplarism before turning my attention to Adam Smith and his *Theory of Moral Sentiments* in the final section. There are indeed strands in Smith's thoughts that contain an exemplarist flavor and raise the same systematic worries as Zagzebski's position. Nevertheless, from the Smithian perspective that I favor, our moral concepts emerge from the everyday practice of holding each other morally accountable through empathic perspective-taking. Such a practice is prior to our admiration for the exemplary person. It takes place in the domain of the "ordinary and vulgar", that is, in the domain of the butcher, the brewer, and the baker. Moreover, it is within the context of an analysis of such practices that our normative commitment to the impartial spectator perspective can be revealed as a regulative ideal. All of this is not to deny that thinking about moral saints is important for our moral life since it reveals that moral action is humanly possible even in extraordinarily challenging circumstances. Yet what is truly admirable is conceptually tied to our commitment to the impartial spectator perspective, whose normative authority should be established independently of our urge to admire, or at least so I am inclined to argue.

2. Zagzebski's Exemplarism: Admiration, the Admirable, and Moral Concepts

Zagzebski weaves an intricate philosophical web consisting of three elementary threads: The notion of exemplars, the analysis of the emotion of admiration, and an externalist and direct theory of reference a la Hilary Putnam and Saul Kripke. In this manner, Zagzebski intends to delineate the complex conceptual landscape of our moral perspective on the world without presupposing a prior conceptual grasp or a normative acknowledgment of moral terminology (117).[1] Most importantly, she wants to ground moral concepts based on "something non-moral" (169).

Exemplars are understood as persons that are at least in some respect "supremely excellent" and therefore "supremely admirable". Additionally, Zagzebski concentrates only on individuals that are exemplary because of acquired excellences rather than natural talents since within the moral realm we have to do with things that are under our control or that could have been otherwise, as Aristotle might express it. While we certainly admire extraordinary natural talents and properties such as perfect teeth, good hair, and a certain height such admiration seems to be of a different type than the admiration for talents that involve some effort in attempting to acquire them. Zagzebski talks specifically about the categories of the hero, who like the Holocaust rescuer is exemplary in showing courage in achieving a moral end; the saint, who shows extraordinary amounts of charity and benevolence; and the sage who, like Confucius, exemplifies the virtue of wisdom.

[1] All page numbers, unless otherwise indicated refer to Zagzebski 2017.

Most significantly, Zagzebski claims that our admiration for individuals tracks their exemplarity without us being able to conceptually articulate why it is that they are exemplary. We are so to speak more certain of their exemplarity than that they are excellent in regard to courage, wisdom, prudence, benevolence, or kindness. From a semantic perspective our access to kinds of moral exemplarity is on par with our access to other natural kinds as suggested by theories of direct reference. We refer to water not because our descriptions of water are necessarily true. Rather our access to water proceeds indexically. It is the stuff to which we are causally exposed in our environment and to which we can demonstratively point as that type of liquid around here. Similarly, the emotion of admiration points us to instances of moral exemplarity and it is through further empirical exploration that we can find out more about its exact nature. Zagzebski mentions specifically narratives, personal experience, but also controlled empirical research as the relevant modes of examination (65ff). Accordingly, when Zagzebski proposes to define value terms such as the notions of virtue, good motive, good end, or good life, and deontic concepts of right, wrong, or duty by referring to exemplars she does not mean to provide us with necessary and sufficient criteria for applying these concepts. In defining virtue as a "trait that makes an exemplar admirable in a certain respect" (113) or a right act as the act that a "person with phronesis […] would characteristically take to be most favored by the balance of reasons for A in circumstances C" (201), she is quite adamant that such definitions contain an irreducible indexical element ("a person like that"). These definitions presuppose further knowledge gained through the empirical investigation of the lives of exemplars. For this very reason, the moral domain could turn out to be broader than traditionally conceived of since our investigation might make us recognize that intellectual virtues such as epistemic humility or open-mindedness are also traits essential for realizing human exemplarity.

Accordingly, Zagzebski circumvents the Kantian challenge against exemplarism in claiming that we have prior non-conceptual access to exemplars through the emotion of admiration. Emotions for Zagzebski are constituted by an irreducible amalgam of affective, cognitive, motivational, and normative components (Zagzebski 2003, 2015, and 2017: Chpt. 2). Admiring is an appreciative emotion in which we are affectively attuned to somebody, whom we sense to be superior to ourselves, whom we are motivated to be close to, and whose activities we are motivated to imitate. Emotions also have their own unique standards of fittingness and our feeling an emotion makes its object appear to satisfy those standards. In admiring a specific person, we see him or her as being admirable, whereby such seeing cannot be understood as a separate cognitive state that is independent of our admiration. We do not feel admiration because we first judge or perceive another person as admirable in contrast to our seeing ice cream causing a desire to eat it. Rather we only see somebody as being admirable in feeling admiration. A fortiori, our admiration can misfire or can be criticized as being inappropriate because of its inherent appeal to a normative fittingness standard of admirability. It also can be regulated by our reflective capacities. In becoming doubtful about the admirability of the persons whom we admire, our admiration for them diminishes in the same manner that our compassion for the distress of another person might diminish when finding out that the person himself was very much responsible for causing his distress by driving under the influence.

The exemplars that Zagzebski has in mind are thus not only people whom we admire but people who are objectively admirable. Moreover, exemplars are

objectively admirable only if our admiration for them survives a process of continuous and conscientious reflection considering additional information. For instance, if we find out that our trusted companions do not admire them, we become more skeptical about our own emotion and might infer that the individuals whom we admire are objectively not admirable (64). Unfortunately, Zagzebski does not say much about what exact type of information might lead us to withdraw our admiration. She seems to think that we uncover it by further investigating our admiring attitudes towards the world. We know that our admiration of a person has to do with the deep structure of a person's character since we admire a person more deeply if we determine that his action is due to an underlying character trait. Our admiration, on the other hand, diminishes if we realize that a person has been mainly motivated by selfish interests (63ff and 107) since a mere selfish motivation would not distinguish him or her from us ordinary folks. Admiration surviving conscientious reflection should therefore be seen as a reliable standard for judging other people as being admirable. Those judgments provide us with good reasons for imitating and emulating the actions and judgments of our chosen exemplars; an emulation that involves taking up their perspective. In simulating their perspective, Zagzebski suggests, we also acquire the motives and reasons for acting that characterize the exemplar (139-40). To make a long story short, exemplarism promises an elucidation of the moral realm that is naturalistically based, that seems to be able to account for the motivational aspects of our moral judgments, and that, in addition, could provide us with means for improving moral education.

3. Systematic Worries about Exemplarism and its Naturalist Credentials

Zagzebski's exemplarism raises, however, a variety of systematic worries that I fear undermine the very foundation of her position. I will focus here on three of them, which are particularly concerning. First, Zagzebski is rather optimistic that different cultures can find common ground by focusing their attention on exemplars (4). After all, human nature is sufficiently similar so that our emotional capacities are very unlikely to track very different kinds of moral exemplars across cultures (17). At the same time, Zagzebski is suggesting that her proposal is a revisionary and a countercultural one since within modernity we not only admire but also vehemently resent extraordinary accomplishments. Not a day seems to go by in recent years without the saintly status of traditional exemplars being challenged, including the "founding fathers" and even Mother Teresa (see, for instance, Michelle Goldberg, New York Times, May 21, 2021).

Zagzebski might respond by arguing that this is just part of our ordinary practice of reassessing the admirability of people in order to determine whether the people who we admire are also genuinely admirable. To be honest I tend to be more skeptical than Zagzebski about the power of reflection to separate the truly admirable from the merely admired. Zagzebski points to how most people view Hitler as a moral monster to suggest that we can distinguish the admirable from the admired in light of the emotional reactions of trusted others. Yet given our evolutionary history, as social creatures we are psychologically predisposed to trust our ingroup more than members who we perceive to belong to the outgroup. Accordingly, we are not naturally committed to what I refer to as the moral stance from within which we treat each other as having equal worth and dignity and as

being morally equidistant of each other. Rather we are profoundly moralizing creatures who endow certain of our norms (including norms of loyalty and purity, see Haidt 2012) with an exalted moral status and as such creatures we favor members of our own group. We also tend to conform in our judgments and our emotional attunement to the social world with members of the ingroup. From this perspective, that most people in the world find Hitler to be a monster might be completely irrelevant for Nazis who do not regard most people as members of their trusted ingroup. It is for this very reason that followers of Trump still admire him and find him admirable, despite acknowledging his many moral failings. What they admire about him is that he projects the resentment of their group and that he wants to "stick" it to the liberal elite.

Even if one is less skeptical about the power of reflection to regulate our admiration in light of a conception of the admirable, the constant reevalutions of our former heroes in contemporary times points in my opinion to a central feature of our practices of admiration and of assessing admirability, that is, its essential cultural and historical relativity. While we certainly should admire Mahatma Gandhi, Nelson Mandela, the Dalai Lama and so on as exemplary human beings in their times and within their cultural traditions, it is not so clear that such admiration carries over easily to contemporary times. Yet, why should they then be regarded as the standards for judging what is morally right and wrong? Moreover, Zagzebski distinguishes among different kinds of exemplars each exemplifying a very specific virtue, that is, courage, charity, and wisdom. If this is so, why should our moral judgments be guided in all domains by how exemplars think about these issues? Does exemplarism really commit us into thinking that Mother Teresa would necessarily have any specific authority to make moral judgments about abortion, the death penalty, or our moral obligations to animals? Alternatively, Zagzebski might appeal to the idea of a perfectly wise and virtuous person or an "idea of exact propriety and perfection" that Adam Smith at times talks about (Smith 1982: 248). Nevertheless, it is ultimately doubtful whether such perfectly wise and virtuous person is an embodied one, a person in flesh and blood whom we would be capable of meeting and admiring or whether such person exist merely in our imagination dependent on a prior grasp of the ideal of moral perfection. Relatedly, we do not merely admire moral exemplars but extraordinary achievements in a wide range of domains of human activity. Admiration then does not naturally limit its scope to the morally admirable. To distinguish the scope of the merely admirable from the morally admirable we could, of course, appeal to our idea of moral perfection. It would, however, imply that Kant's dictum against exemplarism still stands.

Second, Zagzebski claims to ground morality naturalistically since the emotion of admiration is a natural rather than a supernatural phenomenon. The framework of naturalism certainly discourages a philosophical theory to appeal to supernatural properties. Equally important, however, it encourages philosophers to consider what the sciences tell us about human nature (see also De Caro and Macarthur 2010). A fortiori, a naturalist account of human morality would need to look more closely at whether admiration is a phenomenon that from the perspective of evolutionary and ontogenetic accounts can be seen as the basis of human morality. I am more than skeptical in this respect. Scientifically, admiration is regarded to be a "uniquely human emotion". It is particularly an emotion of societies where rank differences are based on so-called prestige hierarchies rather than dominance hierarchies, which one finds among chimpanzees and which

are imposed by threat, aggression, and mediated by fear (Onu et. al. 2016: 217 and 223; Seetman et. al. 2013). Admiration thus presupposes the foundation of a special form of human social cooperation dependent on the enforcement of social and moral norms (Tomasello 2019, Boehm 2012, Wrangham 2019) as it is only within this somewhat more egalitarian spirit that social differentiations among humans are formed. Ontogenetically, children from the age of three years old are already able to distinguish between conventional and moral norms (Smetana et. al. 2014). They know that not hitting another person is a norm that does not depend on social agreement or social authority and recognize that it would not be ok to hit another person, even if their teacher tells them otherwise. Yet only later do children develop an understanding of the category of the supererogatory, which is the basis for the emotion of moral admiration. In the beginning of their moral awakening, they are fully focused on what is obligatory rather than what is admirable or supererogatory (Dahl et. al. 2020).[2]

Zagzebski acknowledges as much when she says that a human society could not exist without a shared sense of what constitutes an intolerable act (192ff). We would also have to assume that the sense for the intolerable is enforced among members of a society and that it would be backed up by humans being emotionally very sensitive to the violation of the norms of the intolerable. Zagzebski refers to such moral sensibility as "morality light". I am a bit perplexed why one would call the basis for our social existence derogatively "morality light". Moreover, I assume that American society would be in a much better shape if people would at least abide by the norms of the intolerable (and refrain from constantly shooting each other). Even more puzzling is the fact that Zagzebski insists that the category of the morally intolerable, of the morally wrong and of moral duty, is determined in respect to what "exemplars cannot tolerate". Given the forgoing considerations, the reference to exemplars seems to be rather superfluous. Even without the existence of any saints we seem to know perfectly well what is intolerable. Moreover, we would be in no position to be sensitive to what is truly extraordinary and admirable without first having acquired knowledge of what is morally intolerable.

Philosophers are however not merely interested in providing a causal explanation of why it is that humans are normative animals and distinguish between moral and conventional norms. Philosophers are ultimately interested in explicating why we ought to be moral. They want to explain why it is that moral commands have a unique normative authority over us even though their validity does not depend on the particular social practices that we are part of. Exhortations and judgments such as don't be cruel, or slavery is wrong are understood as having universal validity. They do not address us in our particularity as Americans, Germans, or Chinese. They speak to us as human beings from the perspective of the moral stance where we possess equal dignity and value and leave behind the framework of mere personal relations. On behalf of Zagzebski one might argue that in focusing on admiration and admirability, she primarily wants to address the above normative question that is central for a philosophical explication of the moral realm. One could then admit that a conception of the intolerable is causally basic for the functioning of a society without admitting that such conception is also normatively foundational. From this perspective, admiration is motivating

[2] I was made aware of this research through a talk by Christina Starmans at the 2021 conference of the Society for Philosophy and Psychology, where she presented new and yet unpublished results of experiments that supports this developmental picture.

us to imitate the person whom we admire, and our judgment of admirability (as the result of such admiration surviving a process of conscientious reflection) provides us with normative reasons for imitating such persons.

Yet one wonders whether admirability in its most general form can adequately ground the normative authority of morality. When I was growing up in Germany in the 1970ties, every boy admired the soccer star Gerd Müller. He was an amazingly effective striker (a classic number 9), who was a member of the German national team that brought home the soccer world cup for a second time in 1974 when the championship was played in Germany. Trust me, a lot of boys at the time tried to be Gerd Müller and they had good reasons to do so. After all, he was a truly admirable striker. Nevertheless, it also tended to be perfectly clear to us that while we had all the reasons in the world to imaginatively enact being Gerd Müller, these were not sufficient reasons to become Gerd Müller in real life. There are indeed more important things to do than playing soccer. Such merely optional reasons however are not the reasons that we are after in trying to normatively ground our commitment to morality. Morality does not seem something that is merely optional for us, like becoming a soccer player. Yet why should admiration and admirability of moral exemplars be different than my childish admiration for Gerd Müller? Why does it mean that I have to take the judgments of moral exemplars more seriously in real life? Pointing out that in this case we encounter the moral kind of admirability seems to beg the very question that we are asking of how admirability normatively grounds morality.

To some extent, Zagzebski acknowledges the above points (see 169-70) but dismisses them in that she asserts that the emulation of moral exemplars proceeds via taking up their point of view and simulating their reasons for acting (136-39 and 170). It is exactly in this respect, one could argue, that admiration for the character of a person differs from our admiration of her skills or accomplishments. Yet even if we grant that the admirability for whole persons provides us with reasons for taking up a person's point of view (rather than merely trying to imitate their external behavior)—and Zagzebski never fully explains why this is so—it is not clear why such simulation changes the above equation. Properly understood, imaginatively taking up another person's point of view does not mean that I become the other person, that is indeed a conceptual impossibility (see Goldie 2011). Empathically taking up another person's point of view means that I am at the same time aware of the fact that it is not my perspective that I am simulating. This is particularly true in situations in which I and the other person are otherwise quite different such as is the case with every normal person and the exemplar they admire. Why then should the reasons or motivations of the admired person automatically become my reasons or motivations for acting? Certainly, taking up the perspective of Gerd Müller and reenacting his reasons for becoming a soccer player does not automatically imply that those reasons should be my reasons for acting, even if I admire him as a soccer player. The reasons of the exemplars must therefore be of a very different kind. We might be tempted to say that this is so because they are moral reasons. But such an answer is very much question begging. Accordingly, it is high noon to turn, as promised, to a discussion of Adam Smith. As I interpret him, we should not conceive of Smith as an exemplarist, even if some of his arguments for the impartial spectator perspective at times contain an exemplarist flavor. Most importantly, Smith allows us to weave the various elements that Zagzebski so rightly appeals to in her exploration of the moral

realm—that is, simulation, admiration for exemplars, the basic sense of the intolerable—into a more plausible map of our moral life.

4. Adam Smith, Empathy, and the Impartial Spectator: How to Acknowledge Exemplars without Being Committed to Exemplarism

Let me start my brief exploration of Adam Smith's conception of the moral realm by acknowledging that his conception of virtue is an ambivalent one and that the centrality of it for his moral philosophy has also been disputed interpretive territory. As Smith is one of the preeminent philosophers thinking about human morality within the context of modern commercial society, this fact is not that surprising since the ancient notion of virtue was a controversial one in the modern context. As it is well known, some political and moral philosophers, like Machiavelli and Mandeville, took a decidedly negative even if nuanced view in this respect (see for example Messina 2017). And they were at times quite happy to let moral hypocrisy rule and allow the "invisible hand" take care of the rest, supposedly creating a buzzling, creative, and rich society from which all of us could benefit.

Smith clearly does not belong to this category of thinkers, despite some of his interpreters being puzzled by the so-called Adam Smith Problem, that is, of how to reconcile his *Theory of Moral Sentiments* (TMS) with the *Wealth of Nations*. He understood modern commercial society from a moral point of view both as an opportunity to expand our moral horizon and as a challenge for the education of our moral sentiments. Scholars have, however, been divided in their judgment about how central Smith takes the notion of virtue to be for his account of our moral life and the foundation of our moral judgments. Generally, it has been regarded to play a secondary even if important role since part VI of TMS, "Of the Character of Virtue", was only added to the sixth edition. Only recently has it been suggested that we should read Smith as being closely aligned with ancient and Christian virtue theory even if adjusted for the modern commercial society (Hanley 2009). Moreover, while Smith talks about virtues throughout the book, he uses the notion of virtue in TMS in a decidedly ambiguous manner. On the one hand, Smith seems to allow for the fact that virtue is achievable for most human beings in ordinary circumstances, what Smith also calls the "middling and inferior stations of life" (Smith 1982: 263). Accordingly, Charles Griswold (1999: 13) views Smith mainly as a philosopher defending the "middling human virtue". On the other hand, Smith at times favors a notion of virtue understood as extraordinary human excellence—something that we admire in that our sense of approbation is "heightened by wonder and surprise" (Smith 1982: 20). The paragon of such virtue is the "wise and virtuous man" whose conduct and judgment are not only oriented at the "idea of exact propriety and perfection" but who also fully comprehends that human nature allows at most for an approximation to such an ideal (Smith 1982: 247-48). The wise and perfect man (whose virtues include both ethical and intellectual virtues) in Smith is best seen as a person whose perspective embodies the ideal of the impartial spectator and who therefore also possesses sufficient humility. He recognizes that, even if he is superior in virtue to individuals in the middling and ordinary stations of life, he is ultimately "but one of the

multitude in no respect better than any other in it" (Smith 1982: 83 and 137).[3] Insofar as we admire such virtuous person we not only agree with his judgments, but those judgments also "lead and direct" our own (Smith 1982: 20).[4]

However, the exploration of the psychological mechanisms with the help of which we hold each other normatively accountable is at the very heart of Smith's elucidation of the moral realm.[5] As he expresses it (Smith 1982: 111), "a moral being is an accountable being", that is, "a being that must give account of its action to some other". In holding each other accountable we do not judge an action to be right and wrong independent of an agent's reasons for acting. Smith strongly objects to Hume who regards "utility or hurtfulness" (Smith 1982: 188) as the primary principle of judging the appropriateness of an action. Indeed, we still blame a person if he has done the right thing for the wrong reasons. Think in this context about an agent who pulls the lever in the famous Trolley case (saving 4 people and letting one other person die in the process) but only because he wanted to get rid of a serious competitor for a job or an award. From a utilitarian perspective we still could judge the action to be the right one. Yet it certainly does not possess any moral worth and the agent is morally blameworthy.

Most importantly, Smith is relevant to contemporary metaethics because he views our practice of holding each other morally accountable as being based on psychological capacities necessary for the constitution of the social realm within which humans live and cooperate. Normative distinctions and normative judgments emerge as effects of our ability to mutually empathize with each other's thoughts and sentiments. Humans as social creatures are constituted so that they cherish being empathized with. Smith understands such empathy—or what one called sympathy in the 18th century—as imaginative perspective-taking, as putting oneself in another person's point of view and simulating the manner in which that person thinks about the situation that he has to respond to. While Smith certainly differs from Hume in his conception of the concrete mechanisms of empathy (Stueber 2015), he agrees with him that empathy allows the "minds of men" to be "mirrors to one another" (Hume 1978: 365). We mirror the other person's thoughts and

[3] In this respect Smith's ideal of the "wise and virtuous man", even if very much inspired by ancient and Christian philosophers is very much a creature of the modern commercial and cosmopolitan world. I am not so sure how I would classify Zagzebski's notion of exemplars in this respect as she points to Confucius as the paradigmatic sage.

[4] My remarks in these two paragraphs have greatly benefitted from the insightful interpretations of Fleischacker 2013, Hanley 2013, and Schliesser 2017 (particularly chpt. 9).

[5] In Part VII Smith claims that moral philosophy generally addresses two questions: "First wherein does virtue consists in? [...] And secondly, by what power or faculty of the mind is it, that this character, whatever it is, is recommended to us? Or in other words, how and by what means does it come to pass, that the mind prefers one tenour of conduct to another, denominates the one right and the other wrong; considers the one as the object of approbation, honour and reward, and the other of blame censure and punishment" (Smith: 265). As I read Smith, the first question is the one that he addresses in parts VI and VII in situating himself within traditional virtue theory. Within the context of modernity and its skepticism about the normative domain the second question is, nevertheless, the philosophically foundational one. Accordingly, Smith addresses it in the first sections of the *Theory of Moral Sentiments.* Independent of the question of interpretive accuracy, only in this manner can we understand Smith as providing us with a plausible foundation of morality within the contemporary metaethical context committed to the naturalist framework. See also Stueber 2017. I further elaborate on how to use Smith within the contemporary context in my book manuscript *The Moralizing Animal* (under contract with MIT Press).

sentiments in taking another person's perspective by bringing the other's thoughts and sentiments "home to ourselves", as Smith is fond of expressing it. Equally important though, in resonating with the other person we also hold up a mirror that allows that person to become aware of his thoughts and sentiments as something for which he can be held normatively accountable. Only in a social context is a human being provided with a "mirror which he wanted before" and which allows him to think of "his own character, of the propriety or demerit of his own sentiments and conduct, of the beauty or deformity of his own mind" (Smith 1982: 110).

More specifically, Smith suggests that our ability to reenact another person's thoughts and sentiments is directly tied to judging the propriety and the merit of his actions or even the propriety of his sentiments themselves. Simplistically expressed, our ability to reenact a person's sentiments by taking her perspective leads us to approve of them and to judge her actions to be appropriate or to possess merit. Our inability to do so leads us to disapprove of their actions. For our purposes, Smith's description of the exact and intricate part of the mechanisms leading to such approval is of secondary importance. For reasons I have explicated elsewhere (Stueber 2017), it is best to understand reenactment of another person's sentiments as grasping their thoughts as their reasons for acting (see also Stueber 2006). Such reenactment might lead to our approval since in understanding another person's thoughts as reasons for actions I view them as considerations that from her perspective speak for her actions. If I can indeed bring such thoughts "home to myself" in recreating them in my mind, I can then also understand them as considerations that could be my reasons for acting. A fortiori it seems that I myself would then approve of such actions or sentiments.

Yet such approval seems to be rather subjective or at most an approval that reflects the social norms of a particular group, culture, or society. We generally tend to listen to these merely subjective judgments because we all like to be liked by the people we live with. Nevertheless, this fact cannot explain why they have the authority of the moral stance from which we make demands that are normatively binding to all human beings regardless of what group or culture they belong to. That is, it is not at all clear why we ought to take a person's approval and disapproval seriously based on his ability or inability to reenact our thoughts. Ultimately such ability and inability might merely reflect certain limits in a spectator's empathic capacity rather than a moral defect in our agency. As it is well known, Smith attempts to address these concerns by referring to the perspective of the impartial spectator. For him, any evaluative judgment based on the ability of an impartial spectator to empathize with an agent's sentiments provides that agent with a normative and moral reason for taking that judgment seriously.

Here I do not want to spend much time discussing how exactly we should characterize the perspective of the impartial spectator. As it is commonly understood, Smith impartial spectator is not an omniscient one nor is he a person devoid of normal human emotions. Rather he is a spectator who is removed from the immediate heat of the action, knows all the relevant facts of the circumstances (as they might be accessible to an agent who is diligent enough to pay attention), and has no selfish interest in the outcome of the action. If I am allowed one more soccer analogy: The impartial spectator could be compared to a soccer fan who watches a game on TV between teams whom he normally does not cheer for, just for the enjoyment of the game. It is a person who is emotionally attuned to watching soccer, who knows the game and the emotions it can elicit, but who is one

step removed from being really interested in any of the teams winning.[6] More importantly for my purposes, however, is the question of why it is that the judgments from such a stance have a special normative authority to make demands on us, why it is that we should take them seriously, or why our actions can be justified only if they can gain approval from that perspective.

One can find two strategies within the Theory of the Moral Sentiments to answer this question. The first is the more obvious and official one (see in this respect also Griswold 1999: 129ff). In light of our discussion of Zagzebski, one could also call it the exemplarist strategy. To motivate the need for the impartial spectator perspective Smith appeals to our experience of being judged wrongly by our peers because they do not fully understand all the relevant factors (including my own mental states) of the situation. The experience of such discordance makes us aware of the fact that we do not merely desire to be praised but that we want such praise to be accorded to us because we are praiseworthy. Besides a desire for praise, human beings are also motivated by a desire of praiseworthiness, a desire that Smith regards to be "by no means derived altogether from the love of praise" (114). It is exactly in this context that Smith refers to our admiration of virtuous exemplars since he sees the

> love and admiration which we naturally conceive for those whose character and conduct we approve of, necessarily dispose us to desire to become ourselves the objects of like agreeable sentiments, and to be as amiable and as admirable as those whom we love and admire the most. Emulation, the anxious desire that we ourselves should excel, is originally founded in our admiration of the excellence of others. Neither can we be satisfied with being merely admired for what other people are admired. We must at least believe ourselves to be admirable for what they are admirable. But, in order to attain this satisfaction, we must become the impartial spectator of own character and conduct (Smith 1982: 114).

Without doubt Smith's account of our desire for praiseworthiness (that is our desire to be praised from the perspective of the impartial spectator) has a very exemplarist flavor. For that very reason, it also encounters all of the philosophical worries that we talked about in the last section. Ultimately, Smith regards persons to be virtuous to the highest degree because they are embodying the impartial spectator perspective. They will thus also be praised from that perspective. Accordingly, our admiration for the "wise and virtuous" enables us to causally explain the desire for praiseworthiness. The central philosophical question that we try to answer is, however, not a causal one but a normative one. We want to know why we should have the desire for praiseworthiness and why we should accept the perspective of the impartial spectator perspective as having normative authority for the evaluation of our character and actions. Or to ask the question differently, if truly excellent people embody the impartial spectator perspective, why does that fact make them admirable? Pointing to our admiration to these exemplars as an answer appears to be begging the question.

Smith's text, however, allows us to reconstruct a philosophically more promising strategy for answering the normative question. For that purpose, I take my departure from Smith's conception of the impartial spectator perspective as reason or principle (Smith 1982: 137). It is the highest tribunal to which we implicitly have

[6] I think this is a better analogy than a comparison to the referee of the game as that person is still too close to the "action" on the field.

to appeal in order to negotiate the comparative strength of our reasons for actions (Smith 1982: 128ff) within our practice of mutual empathic perspective-taking.[7] Central to my argument is the fact that for us to properly simulate another person's perspective we have to take into account differences between us and the target of our empathy. We have to imaginatively adopt the attitudes that we do not share with the other person and quarantine our own attitudes that the other person does not share with us for our reenactment to provide us with reliable insights into the other person's mind. Yet, and here we have to be a bit more careful than Smith (and also Zagzebski), bringing another person's case home to myself in this manner does not automatically constitute approval of his actions since simulating his reasons does not automatically mean that they would be reasons I would act on in his situation. In recreating his perspective, I am at the same time aware of the fact that our perspectives on the world differ in relevant respects. I recognize his thoughts as potential reasons that I would act on only if my own perspective would also undergo relevant changes. It is exactly in this situation, however, that our reenactment of another person's reasons addresses us as a critical, reflective, and therefore self-critical reasoner. Reenacting another person's perspective and his reasons makes a demand on us that requires a rational response. It demands an answer to the question of why it is that we do not make his perspective our perspective, given the fact that his reasons are perfectly intelligible to us. And it is exactly in this context that we implicitly appeal to the normative authority of the perspective of the impartial spectator within which we conceive of ourselves as equal reasoners, or so I would like to argue. The impartial spectator perspective as the "highest tribunal" within which we adjudicate between our reasons for acting, is therefore neither "our own place nor yet his". It is a stance where we look at our reasons "neither with our own eyes, nor yet with his, but from the place and with the eyes of a third person, who has no particular connection with either and who judges with impartiality between us" (Smith 1982: 135).

To fully understand the demand that the reenactment of another person's reasons makes on us, it is important to grasp that in reenacting another person's perspective we reenact a holistic web of attitudes within which a person's thoughts constitutes a reason. Moreover, as already Aristotle understood, our reasons tend to be hierarchical organized. Not only do we have first-order reasons we also have reasons for having those reasons. We not only recognize that somebody likes a neat office. Such recognition would indeed not put much pressure on us to change our messy ways of "taking care" of our office. Additionally, we can recognize that the other person has a reason for keeping his office neat such as that cleanliness is next to godliness. In imaginatively taking up another person's point of view, we ultimately reenact a differently structured framework of reasons. I would suggest that in this manner we enlarge own possibilities of conceiving of rational agency and of considerations that could count as reasons for acting. In reenacting them in our own mind, we imagine them as reasons we can "live" by, that we might feel at home with. Such reenactment ultimately sharpens our sensitivity to our common humanity as rational agents in our local distinctiveness. It is a sensitivity that does not yet constitute full approval. It constitutes a somewhat appreciative engagement with the "vitality" and "life potentiality" (Lipps 1903) that lies in the reenacted perspective.

[7] More specifically, the passages that I find in this context most interesting were taking out by Smith for the sixth edition or were parts of a draft. They were added by the editor for the Glasgow editions of his works. See Smith 1982: 128-30.

Appreciating another person's perspective in this manner has a very positive valence when we try to reenact a Buddhist perspective with its emphasis on sympathy. It might however also resonate with us in a negative and almost scary manner such as when we try to reenact the perspective of a Holocaust perpetrator.

To conclude my discussion of contemporary exemplarism within the context of Smith's Theory of Moral Sentiments, I admit that our recognition and normative acknowledgment of the impartial spectator perspective causally involves a quasi-aesthetic dimension, an appreciative sensitivity, and an appreciative grasp of the intricacies of another person's point of view. Such appreciative component allows us to grasp the strength of another person's reasons requiring us to call for a normative judgment from the perspective of the impartial spectator. I would also acknowledge that appreciative emotions like admiration, but also awe and reverence, for persons with extraordinary achievements at times facilitate our understanding of another person's reasons. Admiration can prime us to think highly of another person's point of view even before we fully engage in simulating that person's perspectives. That probably is a good thing if we admire moral exemplar. Yet as already Smith pointed out, it can also contribute to moral corruption if the wrong person is admired. Admiration for moral exemplars thus should be thought of as being able to play a role in moral education if properly constrained by the impartial spectator perspective. But Kant still seems to have a point in claiming that reference, even an admiring one, to exemplars cannot ground the conceptual framework for our moral life. For that purpose, it is best to follow Smith's analysis—or my favorite reading thereof (Stueber 2017)—of how we hold each other accountable among rather ordinary folks as such practices implicitly commit us to the ideal of the impartial spectator perspective.

References

Boehm, Chr. 2012, *Moral Origins: The Evolution of Virtue, Altruism, and Shame*, New York: Basic Books.

Chalik, L. and Rhodes, M. 2020, "Groups as Moral Boundaries: A Developmental Perspective", in Benson, J.B. (ed.), *Advances in Child Development and Behavior*, vol. 58, Cambridge, MA: Academic Press, 63-93.

Dahl, A., Gross R.L. and Siefert, C. 2020, "Young Children's Judgment and Reasoning about Prosocial Acts: Impermissible, Suberogatory, Obligatory, or Supererogatory?", *Cognitive Development*, 55, 3, https://doi.org/10.1016/j.cogdev.2020. 100908

Debes, R. and Stueber, K. 2017, *Moral Sentimentalism*, Cambridge: Cambridge University Press.

De Caro, M. and Macarthur, D. 2010, *Naturalism and Normativity*, New York: Columbia University Press.

Fleischacker, S. 2013, "Adam Smith and Equality", in Berry, Chr., Paganelli, M.P. and Smith, C. (eds.), *The Oxford Handbook of Adam Smith*, Oxford: Oxford University Press, 484-500.

Goldie, P. 2011, "Anti-Empathy", in Coplan, A. and Goldie, P. (eds.), *Empathy: Philosophical and Psychological Perspectives*, Oxford: Oxford University Press, 302-17.

Griswold, C. 1999, *Adam Smith and the Virtues of Enlightenment*, New York: Cambridge University Press.

Haidt, J. 2012, *The Righteous Mind: Why Good People are Divided by Politics and Religion*, New York: Vintage Books.

Hanley, R.P. 2009, *Adam Smith and the Character of Virtue*, Cambridge: Cambridge University Press.

Hanley, R.P. 2013, "Adam Smith and Virtue", in Berry, C., Paganelli, M.P., and Smith, C. (eds.), *The Oxford Handbook of Adam Smith*, Oxford: Oxford University Press, 219-40.

Hume, D. 1978 [1739], *A Treatise of Human Nature*, Oxford: Clarendon Press.

Irwin, T.H. 2015, "Nil Admirari: Uses and Abuses of Admiration", *Proceedings of the Aristotelian Society, Supplementary Vol.*, 89, 223-48.

Kant, I. 1981, *Grounding for the Metaphysics of Morals*, Indianapolis: Hackett.

Lipps, Th. 1979 [1903], "Empathy, Inner Imitation and Sense-Feelings", in Rader, M. (ed.), *A Modern Book of Esthetics*, New York: Holt, Rinehart and Winston, 374-82.

Messina, J.P. 2017, "Kant, Smith, and the Place of Virtue in Political and Economic Organization", in Robinson, E. and Surprenant, C.W., *Kant and the Scottish Enlightenment*, London: Routledge, 267-85.

Onu, D., Kessler, T., and Smith, J.R. 2016, "Admiration: A Conceptual Review", *Emotion Review*, 8, 218-30.

Schliesser, E. 2017, *Adam Smith: Systematic Philosopher and Public Thinker*, Oxford: Oxford University Press.

Seetman, J., Spears, R., Livingstone, A.C., and Manstead, A.S.R. 2013, "Admiration Regulates Social Hierarchy, Dispositions, and Effects on Intergroup Behavior", *Journal of Experimental Social Psychology*, 49, 534-42.

Smetana, J.G., Jambon, M., and Bell, C. 2014, "The Social Domain Approach to Children's Moral and Social Judgments", in Killen, M. and Smetana, J.G. (eds.), *Handbook of Moral Development* (2nd edition), New York: Psychology Press, 23-45.

Smith, A. 1982 [1759], *The Theory of Moral Sentiments* (Raphael D.D. and Macfie A.L. eds., The Glasgow Edition of the Works, and Correspondence of Adam Smith), Indianapolis: Liberty Fund Press.

Stueber, K. 2006, *Rediscovering Empathy: Agency, Folk-Psychology and the Human Sciences*, Cambridge, MA: MIT Press.

Stueber, K. 2015, "Naturalism and the Normative Domain: Accounting for Normativity with the Help of 18th Century Empathy-Sentimentalism", *Rivista Internazionale di Filosofia e Psicologia*, 6, 24-36.

Stueber, K. 2017, "Smithian Constructivism: Elucidating the Reality of the Normative Domain", in Debes and Stueber 2017, 192-209.

Tomasello, M. 2019, *Becoming Human: A Theory of Ontogeny*, Cambridge, MA: Harvard University Press.

Wrangham, R. 2019, *The Goodness Paradox: The Strange Relation between Virtue and Violence in Human Evolution*, New York: Vintage Books.

Zagzebski, L.T. 2003, "Emotion and Moral Judgment", *Philosophy and Phenomenological Research*, 66, 104-24.

Zagzebski, L.T. 2010, "Exemplarist Virtue Theory", *Metaphilosophy*, 41, 41-57.

Zagzebski, L.T. 2015, "Admiration and the Admirable", *Proceedings of the Aristotelian Society, Supplementary Vol.*, 89, 205-21.

Zagzebski, L.T. 2017, *Exemplarist Moral Theory*, Oxford: Oxford University Press.

Human Enhancement and Reproductive Ethics on Generation Ships

Steven Umbrello and Maurizio Balistreri***

* *Delft University of Technology*

** *University of Turin*

Abstract

The past few years have seen a resurgence in the public interest in space flight and travel. Spurred mainly by the likes of technology billionaires like Elon Musk and Jeff Bezos, the topic poses both unique scientific as well as ethical challenges. This paper looks at the concept of generation ships, conceptual behemoth ships whose goal is to bring a group of human settlers to distant exoplanets. These ships are designed to host multiple generations of people who will be born, live, and die on these ships long before they reach their destination. This paper takes reproductive ethics as its lens to look at how genetic enhancement interventions can and should be used not only to ensure that future generations of offspring on the ships, and eventual exoplanet colonies, live a minimally good life but that their births are contingent on them living genuinely good and fulfilling lives. The paper further claims that if such a thesis holds, it also does so for human enhancement on Earth.

Keywords: Space ethics, Human enhancement, Reproductive ethics, Generation ships.

1. Introduction

Over the past several decades, both scholarly and popular literature has actively attempted to highlight and explore the various existential risks that might jeopardise continued life on planet Earth. Ranging from nuclear winter (Baum 2015) and climate change (Butler 2018) to runaway nanotechnology (Umbrello and Baum 2018) and artificial superintelligence (Bostrom 2016). Despite some of these existential risks being more plausible, what has concerned scholars is how to avoid, ameliorate, and mitigate some of the threats. More recently, one of the proposed solutions made quite popular by science fiction in shows like The 100 (2014) and movies such as Passengers (2016) and Interstellar (2014) is to have humans leave Earth and colonise other planets. Recently, Tesla and SpaceX CEO Elon Musk stated that to ensure the species' long-term survival, humans must become a multi-planet species (Sheetz 2021).

Argumenta 10, 1 (2024): 453—467
ISSN 2465-2334

DOI 10.14275/2465-2334/20230.umb
First published: 16 June 2023

In lieu of the ability to travel beyond the speed of light or harness the power of a theoretical gravity propulsion system (i.e., Tajmar and Bertolami 2005), habitable planets within the perennial circumstellar habitable zone remain beyond reach for us currently (Schulze-Makuch 2020). Mars has been the subject of much recent attention as the most likely candidate for initial extraterrestrial colonisation. However, it will require significant geoengineering efforts for the planet to be able to sustain large and growing populations, a considerable engineering challenge (McInnes 2009). Exchanging one engineering challenge for another, should habitable planets present themselves in the perennial circumstellar habitable zone yet lie beyond reach within the average human lifespan (i.e., outside our solar system), what will be required is an interstellar ark starship or simply a generation ship. These are hypothetical spacecraft meant to travel between star systems at sublight speeds. This means that the original crews of the ships, and in many cases multiple generations following them, would not live long enough to arrive at their destination planets. They would be born, live, and die on the ships with the goal of becoming the carriers of the genetic heritage of future generations that would populate their destination planet(s) and (Szocik 2021); beyond that, the shepherds of cryopreserved human and animal embryos that can be used to seed new planets (Edwards 2021). The motivations underlying the need for such vessels could be (1) life on Earth may remain habitable, at least for a finite number of people; hence, using these ships, current inhabitants of Earth can leave for a new habitable homeworld and/or (2) generation ships can be a means of last resort for the survival of the species, i.e., perhaps as a consequence of global climate change (i.e., as depicted in *Interstellar* 2014).

Although such ships pose a gargantuan engineering obstacle, they have nonetheless drawn scholarly attention from projects like Project Hyperion, which looked at the ideal population sizes to man these ships as well as what current and future technologies would be required to ensure the success of such enterprises (Smith 2014; Hein et al. 2012). Not only this, but various ethical issues emerge as a consequence of such a venture. Ethicist Niel Levy (2016) noted several ethical considerations to take into account when considering generation ships, primarily that despite the original crews will almost certainly have a better quality of life on the vessel in terms of access and quality of health care, education, and nutrition, they will almost certainly have little if any control over personal, career, or reproductive choices given the need to tightly control and ensure the long-term success of the mission. Hence, such individuals will have their freedoms almost entirely curtailed despite having the best versions of the things that are required to meet the minimum threshold for well-being (Lester 2013).

In this paper, we argue that it is not *a priori* morally responsible[1] to have children on a generation ship despite their ensured and perhaps abundant access to the minimum necessary conditions for survival. We argue that in order for the choice to have a child on a generation ship to be morally responsible, parent crew members must ensure to the best extent possible that they give their children a good life, a life worth living[2] beyond that of a means to some extremely distant

[1] Moral 'responsibility' here is best read as moral 'permissibility'.

[2] In this sense, a 'life worth living' is best understood as internal, that is life is of sufficient value for the individual concerned to be worthwhile; not unlike that of that delineated by McMahan (1998: 226-28).

end (i.e., genetic carriers for future planet colonisers).[3] Not only this, but we use this scenario to demonstrate that this principle not only obtains to generation ships but on Earth as well.

2. Born, Living, and Dying on a Starship

Given the extreme distances of other exoplanets (i.e., planets beyond our solar system) that have currently been discovered and even if we could overcome the monumental engineering challenges of building a generation ship large and sophisticated enough to ensure the long-term survival of the humans aboard so that their descendants could reach their new home world, this would necessarily take many generations, feasibly more generations than there has been up until this point on earth (Szocik 2021). Still, such a multi-generational journey aboard a craft that is fundamentally different in almost every way then the environment on Earth, one that poses existential challenges to the biology of those on board, would almost certainly require genome editing interventions so that the crew members could more safely survive such a long journey with the greater risk of being exposed to stellar radiation and potential changes in gravity conditions. Aside from technical requirements, such a ship would need to ensure that such radiation and gravitational anomalies would be at a minimum, so it makes the most sense to intervene at the crews' genetic levels to make them as impervious as possible.

Intervening at the individual (human) level rather than at the environmental (starship) level would naturally be the least costly of the two options; although it would be reasonable to hypothesise that both strategies should be taken in unison to a degree to ensure that redundancies increase the likelihood of mission success. Still, this latter suggestion may be the most technically feasible in the interim, given current trends in genetic biology (e.g., Daly 200; Singh et al. 2011). This would initially mean intervening at the genetic level on the pioneer crew who board the ship before takeoff, and perhaps, should the ship be carrying human embryos on those embryos on Earth before takeoff. Even if one assumes that modifying an eventual exoplanet via geoengineering techniques raises no morally relevant concerns, we take the position that the more straightforward approach of intervening on the individual genetic level poses the least, if any, moral problems.

Some would indeed argue that genetic interventions of any kind are immoral given that our genetic heritage is sacred or is held in common. Thus, intervention at the individual level to change this heritage would be fundamentally immoral (e.g., Sandel 2007; 2009; Kass 2003). Despite many issues with this position (e.g., see Kudlek 2021), biologically speaking, such a position is simply without grounds. Sexual reproduction (or assisted reproduction) *de facto* modifies the

[3] Dominic Wilkinson (2011) distinguishes between various ways of understanding a life worth living (see also, Parfit 1984: 493-502; DeGrazia 1995; Griffin 1986: 7-74. There is an internal sense of a life worth living (life is of sufficient value for the individual concerned to be worthwhile) and an external sense of a life worth living, and its value to others (Buchanan and Brock 1986: 74). In addition, some authors make a distinction between the level of a life worth starting (for an individual who does not yet exist) and the level of a life worth continuing (for an existing individual) (Benatar 2006: 22-23). Some authors also argue that it is possible to distinguish whether life is above or below the zero point (Buchanan et al. 2000: 224; Wilkinson 2011; Glover 2006: 57; Garrard and Wilkinson 2006: 486; Wyatt 2005).

genetic heritage of each offspring it produces. Hence, each time a child is born, its genetic makeup is necessarily diverse from that of its progenitors. To support a position where the genetic heritage of humans is monolithic is simply incorrect; rather, what is an immutable feature of human nature is that such heritage is dynamic and changes from birth to birth. Even in the case of cloning, where we produced an embryo using the nuclear DNA of the somatic cells of an adult, such genetic heritage would nonetheless be diverse from that person (Ayala 2015). Likewise, the argument for a monolith genetic heritage via cloning fails even more given that only females can receive both the mitochondrial and nuclear DNA of the same person, meaning that cloned males will necessarily have diverse outcomes, lest we condemn that sex to die out, which, as a consequence again, would render the genetic heritage of humanity to change (Balistreri and Umbrello 2022b).

Furthermore, it would be hard to sustain the position that genetically engineering our offspring is morally egregious when such modification produces outcomes that positively impact the quality of those offspring's lives. A simple hypothetical example would be the use of such genetic engineering techniques to intervene in our offspring's genetic code to ensure that, when born, they are more resistant, if not entirely immune, to certain diseases (even presently terminal ones) as well as physical and cognitive enhancements that can make them and their descendants better apt at coping with the rigours of their lives and environments (i.e., Hofmann 2017). The moral challenges often levied against these types of techniques are those raised by making the distinction between therapeutic interventions and those that are for enhancement purposes. Still, these arguments make a distinction without a moral difference and have yet to provide watertight arguments (i.e., see Kudlek 2021, who challenges these positions; see also Balistreri and Umbrello 2022a). On such arguments is that those born with such enhancements would have had such enhancements chosen for them, and, as a consequence, would no longer be the master of their own lives, but mere passengers in the driver's seat given that those who were not subject to such interventions (the unenhanced) would not recognise them as part of the same species and thus not *de facto* attribute them the same degree (if any) of human dignity and all those rights/benefits as a consequence. Although this latter suggestion is not necessarily true given the marked rise in the suggestion and application of the attribution of such dignity and subsequent legal rights to nonhuman animals and other entities like AI systems, hence the attribution of such would not be far-fetched for humans who have received enhancements (e.g., Vink 2020; Pagallo 2018). This application of rights and dignity has even been proposed for (sufficiently anthropomorphic) potential extraterrestrial life, something that would be used as a designation after millions of years of speciation pressures on a generation ship and eventual exoplanet colony (Frietas 1977).

Still, beyond this, the argument that the freedom of enhanced individuals is *de facto* curtailed does not hold water. Such individuals would still have the freedom to use those enhancements in the ways that they desire, as well as to further modify/remove such enhancements or to augment themselves further. Even further, such enhancements do not expropriate the needs for skill and effort to be exerted in order to take advantage of their benefits, like the skill that current humans possess now, they are best understood as propensities and dispositions that require work and training in order to benefit from their use. Finally, certain moral enhancements can feasibly augment the enhanced person ability to empathise,

disposing them to greater sensations of gratitude towards their progenitors for their currently enhanced dispositions and make them better apt at putting the interests of their community members ahead of their own (i.e., see Rakić 2017; Ahlskog 2017). Taking these arguments into account, the genetic enhancements that potential generational ship members would undertake should not be considered elective or vanity medical procedures, but therapeutic, as they would permit the astronauts to have a greater probability of success in both surviving the many generations that such a ship would need to make its journey and the survival on certainly diverse (in comparison to Earth) exoplanets.

Here the reader would surely raise the notion that such interventions would be best undertaken only *after* the child is born rather than in anticipation. This ex-post intervention would be described as somatic line enhancement (Balistreri 2020), where the person's cells are directly intervened on while leaving the oocytes and/or spermatozoa untouched, thus, such enhancements would not be passed down through reproduction into the next generation. Such methods would require each born generation to undertake the interventions. Naturally, this would permit more research and innovation to take place, thus increasing the potential safety of the interventions prior to their application, if, hypothetically, in such a future scenario of intergeneration ships such a technology has not already been perfected. Still, adopting the somatic line enhancement approach would potentially risk the lives of newborns to the environmental hazards (i.e., potential celestial radiation, gravitational anomalies, etc.) that they would otherwise not be exposed to if they were born with the enhancements. Germline enhancement interventions then pose themselves as the more ideal solution. This approach would take place by intervening on the level of embryos or gametes prior to their fertilisation. Theorists who have explored extraterrestrial colonisation argue that the transportation of large quantities of embryos and gametes serves as one of the best methods for large-scale colonisation endeavours. Such could even be fertilised and gestated in artificial wombs via ectogenesis (Edwards 2021). Regardless, germline enhancements would remove the need for somatic line interventions post-birth since the enhancements of any given individual would be passed down to subsequent offspring. This latter (germline) approach could, and perhaps should, take place prior to the departure of such a ship, and, would therefore take place on Earth. This latter point is not insignificant, given that an important thesis in this paper is that the place in which these types of enhancement interventions take place do not post any *per se* moral quandaries. Similarly, given that germline enhancement approaches take place prior to birth, this means that the beneficiaries of such enhancements could not have *a priori* consented to such interventions, however, like the previous point, we are that this too is not *mala in se* as long as the interventions are proven safe and does not expose the offspring to any unwarranted risks.

The latter point, concerning consent, is particularly important to address head-on. One would think that somatic line enhancement approaches would be more ethical. However, despite the safety concerns raised above on why they may be best avoided, it does not explain how children, enhanced or otherwise, are not capable at a young age at making autonomous choices. As such, parents have the moral responsibility to make choices in their place, as their *de facto* representative, all while *not* being considered unduly paternalistic nor in violation of the child's autonomy or right to consent (i.e., Scanlon 2000; Orfali 2004). Should we oppose such a position, which runs contrary to the accepted positions in bioethics

concerning parental roles in neonatal medical decision-making, then we would have to accept the position that permits preventable risks to newborn offspring on generation ships and future exoplanets to take place. Here, the reader may induce that the position we are arguing for can be boiled down solely to that of a principle of minimal well-being where genetic enhancement interventions are morally permissible, if not morally obligated, in order to, but not beyond, ensuring that those born have a minimally sufficient capacity to meet the demands and challenges of prolonged space flight and exoplanet habitation. We, instead, take the position that this principle of minimal well-being via genetic enhancement is not a sufficient condition for making the morally responsible choice of having a child on a generation ship. We argue that, although the initial (adult) pioneers made the informed choice to face significant challenges and make arduous sacrifices, these challenges and sacrifices should not automatically be subsumed onto subsequent generations that will *necessarily* be born on a generation ship without first being able to ensure that they can be given a sufficiently good life beyond that of mere survival. We argue that genetic enhancements are one of the means by which this can be achieved.

3. Morality of Birth on a Generation Ship

As we mentioned, our goal is to show that a principle of minimal well-being is not a sufficient condition to be considered responsible when deciding to give birth to a child on a generation ship. Naturally, one can make the argument, and they would probably be correct in doing so, that as time progresses, and thus scientific research and innovation, such germline genetic enhancement interventions will continually advance, bringing with them not only novel and more efficacious outcomes but all this in a more safe way. We can, therefore, say that in some hypothetical future in which the technological readiness level of Earth is sufficient enough to permit or necessitate the creation and manning of a generation ship, then we can say that such a readiness level would allow a sufficiently advanced form of genetic enhancement that would make those who are born, live, and die on a generation ship relatively safe. This means that we can safely assume that those who are the beneficiaries of these enhancements on those ships would be quite resilient against the environmental hazards native to the hostile environments of such a journey. Still, despite the efficaciousness and safety of such interventions, simply ensuring the minimum well-being (i.e., not exposing offspring to preventable harms), and thus, is only a necessary but not sufficient condition for being considered moral in the decision to have offspring on a generation ship.

However, despite the 'technofix' proposed (i.e., germline genetic enhancements), these environmental hazards do not necessarily account for the psychosocial issues that such individuals will face on a generation ship and on the initial settlements on the destination exoplanets. Conceptually, such ships will be limited in size; thus, the crew will necessarily be constrained by the space provided to them within the internal space of the ship. Given that such a journey is necessarily life-long, confined proximity with a finite number of individuals poses unique social and psychological pressures on crew members. Although scholars have proposed that should the sufficient technological readiness levels that permit generation ships actually arrive, that readiness level would similarly permit a large enough ship so vast to ameliorate or negate this issue entirely (Levy 2016). Still, this remains to be seen. If we take the issue of lifelong close proximity on a ship

seriously, as our current technological readiness level allows us to explore, then we can already begin to investigate means to ameliorate these challenges. Szocik et al. (2020: 7), for example, imagine a panopticon-style internal ship to permit more open spaces so that members can be continually exposed to novel stimuli. However, this raises privacy concerns which further raise other psychosocial challenges. Others, however, have proposed the use of virtual, augments, and mixed reality technologies (not dissimilar to the *Holodeck* in the Star Trek TV and film series) to permit crew members not only to be exposed to novel stimuli but to integrate themselves into more familiar natural environments that stimulate the evolutionary propensities innate in human development (Salamon et al. 2018; Joshi and Mardon 2021; Del Mastro et al. 2021).

Still, even if such technologies present themselves as a potential solution, it remains more probable than nought that crew members, if/when they arrive at their destination exoplanet, will remain, live, and work in relatively close proximity for more of their waking time to promote the cause and support the success of their mission to ensure a working and sustainable colony. This does not mean that these pioneers will not have any individual time, which would be a difficult position to hold; however, it does make sense to say that such time would be relatively limited, and all the time they are not alone would be dedicated to the coordinating work of the mission., not unlike we see currently and historically with space exploration endeavours (Struster 2010). This constraining feature that would most likely be necessary for such enterprises will undoubtedly affect the quality of life of those who would unquestionably see their preprogrammed lives quite constrained as means towards so future end, a future which they will almost certainly not live long enough to experience. VR/AR/XR techniques would be helpful here to permit the most diverse access to experience possible. However, this *Matrix*-like solution would certainly not resolve the more substantial issue of lack of freedom in the choice of the crew members to self-determine their own goals and desired outcomes. The success of such a mission may be determined by limiting these very freedoms, dedicating all efforts and cultivating skills towards the mission's goals. Levy is clear in this thesis, saying that

> A generation ship can work only if most of the children born aboard can be trained to become the next generation of the crew. They will have little or no choice over what kind of project they pursue (Levy 2016).

Hence, despite the access to the best healthcare, nutrition, and safety on board a generation ship (such would be necessary to ensure success), it is certainly offset by the psychosocial constraints likewise necessarily imposed on those who have such access to likewise assuring mission success. So, we see a context of minimally sufficient well-being offered, perhaps much more than many currently living on Earth have access to, yet this is hardly a sufficient condition to have a "good" or "fulfilling " life, regardless of the definition used to conceptualise those arguably abstract adjectives. The source of the issues, fundamentally, is an issue of timespan. Here, we can hardly argue with the moral responsibility assumed via the sacrifices of the original crew members. These pioneers decided to undertake the mission and accept the challenges and consequences. However, by doing so, they also assume the explicit assumption that such a mission necessitates future generations to be born on board, who could not make the same choice to make those sacrifices towards the mission's objectives.

It would be hard to argue that those born on board should not be able to self-determine their interests, goals, career, and lives. One could feasibly imagine that those born aboard the generation ship could, once reaching adulthood, or the age in which their ability to make fully autonomous choices can be made (i.e., Leisman et al. 2012), could choose not to sustain the mission's goals, to abandon the enterprise, and to return to Earth. However, this would certainly be impossible, or at least existentially challenging, given that generation ships are predicated on the fact that faster-than-light speed travel is not discovered or possible. Hence, the vast distances such a ship is designed to traverse exclude the necessity and possibility of return journeys. Consequently, those born on board are condemned to remain on board. Likewise, the genetic modification interventions that will almost certainly be required to ensure the survival of the people who arrive on the destination exoplanet will certainly not permit, at least not without further modifications, the seamless return to Earth, which will have a non-native environment for those exoplanet colonisers. These more material challenges aside, there remains the apparent issue of biological and cultural speciation, which would occur at an evolutionary rate in missions that last thousands and millions of years. The differences, despite the potential choice of the crew members to return, may make the similar cultural and biological speciation that will, in the meantime, happen on Earth an obstacle for integration by the crew members. Simply put, the culture on both Earth and the generation ship will necessarily evolve, with natural evolutions divergences which will, over long periods, create fundamental differences making reintegration between the two groups difficult, if not impossible (i.e., see issues of speciation in Avise and Walker 1998).

4. Surviving on a Generation Ship is not Living on a Generation Ship

The lives of those who will board generation ships, and certainly more of those born on those ships, will almost certainly be different from those of most people born, live, and die on Earth. Many of the unique environmental, social, and psychological challenges that emerge as a consequence of such an endeavour require a substantial investment in ensuring that those who populate such ships have access to the necessities to ensure that their existence, their survival, is not jeopardised by any possible or emergent threats. In many ways, those who will live on such generation ships will have, whether they know it or not, access to many fundamental necessities to survival those currently living on Earth are not privy to. Access to optimal healthcare (both psychological and physiological), nutrition, entertainment, and knowledge (i.e., access to Earth's repositories (locally stored or via quantum connections to Earth, e.g., see Sidhu et al. 2021)).[4]

Of course, critics may argue that access to these unprecedented resources and being part of an unprecedented and monumental endeavour such as exoplanet colonisation via a generation ship will ameliorate or provide the fundamental meaning to sustain those born on board despite the constraints on their individual freedom. Likewise, an argument could be proposed that life, even that of mere

[4] The latter, arguably, would permit cultural co-evolution by a constant and lag-free exchange of knowledge development and dissemination. Of course, that would be contingent on the time constraints put on the crew members to engage in scientific and cultural developments given their potentially constrained conditions.

survival, is sufficient to deem it worth living (e.g., Magni 2021). Although this choice may be adopted by the individual decisions of the original pioneers of the generation ship, it cannot be *a priori* abdicated to subsequent generations. As such, the minimum threshold for well-being cannot serve as the exhaustive condition for determining the moral acceptability of reproductive decision-making on generation ships (Glover 2006).

To begin, if we take the minimum threshold for well-being as the criterion for determining the morality of reproductive choices, then the vagueness of what would be considered such well-being would mean that it would be difficult, if not impossible, to determine cases of irresponsible reproductive choices clearly. More precisely, the threshold is not delineated, consequently permitting violations. Of course, this threshold's philosophical and pragmatic benefit is that it is partial to the difference between well-being and a life full of suffering. This means that giving birth to someone who cannot be birthed into or beyond this threshold is absolutely immoral; likewise, it does not morally obligate progenitors to birth children into lives beyond that threshold (even though it is naturally preferable than nought). To a certain degree, the use of this principle exclusively can obtain on Earth, with highly dynamic and unfixed variables that impact the contexts of birth. However, in the highly fixed contexts of generation ships, birth, particularly those selected and directed via embryonic fertilisation and subject to genetic modification, should be gestated if and only if their lives can not only meet the threshold but are allowed to achieve a full and good life.

We thus shift the threshold above that of the classical understanding of the minimum threshold of well-being. Given these available choices (of which embryos and which modifications) we have access to, if we cannot guarantee that the offspring can have a fulfilling life, the crew members shouldn't reproduce. This, of course, undermines the underlying principle of generation ships entirely. Hence, the philosophical principle of this higher threshold for a fulfilling life either morally jeopardises the generation ship project or, more optimally, provides the philosophical norms for ensuring responsible reproductive practices for the future of such ships and eventual exoplanet colonies.

More fundamentally, however, the classical minimum threshold is that it does not make sufficient nuance between the variety in the lives of the offspring that could feasibly be birthed. For example, as long as the offspring has access to the minimally necessary resources for well-being, then it would be considered responsible in this principle to *knowingly* give birth to offspring with physical disabilities such as blindness, anhidrosis, and/or congenital insensitivity to pain, among others even if it were possible to give birth to the offspring without such issues (e.g., see Savulescu 2001: 417; see also Schon et al. 2020). If knowing that the outcome could be directed in a different, better direction, it would be difficult to sustain the position that the minimum threshold of well-being is a sufficient criterion to evaluate the moral acceptability of reproductive choices. Likewise, there is an inherent vagueness in adjudicating when the threshold is traversed. If we consider the same offspring with further illnesses or disabilities, we can reasonably imagine that, for the child, their existence is so consumed by suffering that such a life does not meet the minimum threshold for well-being. This, of course, is a non-subjective perspective. The child itself may be driven to such suffering that they subjectively determine that their life is no longer worth living; however, they may, despite all this suffering, still determine for themselves that

their life is still worth living. However, the principle does not make such distinctions *a priori* and thus undermines itself.

These cases, however, are not necessarily relegated to space *per se*; in fact, the above examples are fairly common on Earth today. Nonetheless, the philosophical underpinnings of such cases can likewise be extended to contexts that may be found on generation ships, i.e., imagining cases in which life on a generation ship would no longer be worth living. Let us imagine a relatively large-sized generation ship that can hold a few thousand crew members. As we mentioned, the relative success of such a mission would more than likely constrain the individual freedom of any given crewmate. To this end, even if a return to Earth mid-journey were technically feasible, it is reasonable to assume that such would not be permitted given that each member would necessarily need to be trained and consequentially contribute labour and a particular set of skills and expertise that are mission-critical (i.e., Pellerin 2009; Galarza and Holland 1999). Hence, given the necessarily multi-generational nature of the mission (aside from the last generation on the ship before arrival), the direct benefits of the work done to ensure mission success cannot be derived by those who are born, live and die on board. Their lives, of course, would be quite good (materially speaking), so it is unlikely they would live lives of great physical suffering. Likewise, the natural periods in which crew members will suffer from psychological issues concerning their constraints of freedom, such as depression and boredom can be feasibly ameliorated via pharmaceuticals or entertainment systems. Even in such cases, where one's life is not their own, but functionally a vehicle for the success of future generations yet to be born does not entail that those living those lives in the present are lives not worthy of being lived.

One may then argue that we can modify such an example by inserting progressively degenerative conditions. Life on the generation ship is necessarily adaptive to the minimum crew necessary for mission success. The ship necessarily functions as a closed system to a degree. However, upon arriving at the new exoplanet, settlement and expansion can begin, necessarily increasing resource demand, something that would not have occurred within the closed system and controlled system of the generation ship. Life within this new settlement, particularly for those born on the ship and settling on the new world, will arguably be more complex and challenging to adapt to. Still, even in this worsening case, the argument cannot be sustained that their lives have ceased to be worth living given the worsening conditions. We would continue to add to this degeneration of states without *a priori* arriving at some logical conclusion where we can determine that the lives of these settlers are no longer worth living. Such conditions can be imagined to be increasable, isolating, constrained, and psychophysically strenuous without logically being able to determine the unworthiness of that life. To remind the reader, these cases are the logical conclusions of the minimum threshold of well-being. The principle makes no distinction between these cases, even where degeneration of conditions *ad absurdum* is present. To remind the reader, we take the position that this principle is flawed for this reason, i.e., that the principle of the minimum threshold of wellbeing is flawed given that it makes no distinction on the wrongness of a child born into a life that is barely worth living (i.e., directly on the threshold).

Beyond this fundamental issue, an issue fundamentally predicated on the vagueness of thresholds is the difficulty, if not impossibility, of objectively setting limits that determine moral responsibility. Likewise, concentrating overly on

establishing such a precise threshold, the principle also misses the mark in its ability to properly characterise the problems and challenges that characterise those lives it aims to evaluate as worthy or not to be lived. Adults could be said to be autonomous in the capacity that they can decide that a sacrifice that necessarily diminishes their well-being is worth it, and thus, their life as a consequence, remains worth living. However, this is different for a child who we may overestimate to be worth living in harsh conditions like those we purport will be most plausible on a generation ship and eventual exoplanet colony. This overestimation, even if made with the best intentions for the child, does not mean our choice is unquestionably morally acceptable. On the contrary, even with the noblest intentions for those offspring, we may nonetheless condemn them to a challenging life that is not worth living. Or, more clearly, it is never acceptable to be born into a life that is barely worth living. The principle of the minimum threshold of well-being would argue that the preceding sentence is morally unacceptable. This is because, as we explained, the principle does not make moral evaluations on progenitors as long as they are above that threshold, even if living at that threshold is one of extreme suffering. (i.e., the principle defends the notion that it is better to be born into a life barely worth living and full of suffering than no life at all). We contend that this is morally irresponsible, given that the slightest change in any person's life at the threshold can instantly make their life no longer worth living.

5. Moral Obligations for Progenitors on Generation Ships

If we then take the stand against the minimum threshold of well-being principle, where does that leave us concerning our moral responsibility and obligations concerning reproduction on generation ships? We argue, similar to that of Julian Savulescu and Guy Kahane, that we not only must *not* have offspring whose lives are barely worth living but, more radically, that we have a moral obligation to give birth to the *best* offspring. Savulescu and Kahane take the position that:

> If reproducers have decided to have a child, and selection is possible, they have a relevant moral reason to [should] select the child, of the possible children they could have, whose life can be expected, in light of the relevant available information, to go best or at least not worse than any of the others (Savulescu and Kahane 2009: 274).

A closer look at this position reveals that logically speaking, it does not present itself as a negation of the principle of the minimum threshold for well-being, which we argued is fundamentally flawed (cf. McMahan 2002: 170; 2009). Consequently, we contend, at least *prima facie,* that Savulescu and Kahane's position does not add any (problematic) moral elements to the issue of reproduction on generation ships. Regardless, we find some problems with their conception of this moral reason they stipulate. Firstly, they argue that progenitors have moral reasons but that there is no clear moral obligation to choose, among the open options of potential offspring they may be presented with, is the best option. Secondly, they relegate their decision for making the best choice to the level of genes. This can be expanded to the potential embryo selection created via reproductive enhancement techniques. Finally, they argue that using the adjective 'best' in reference to the child chosen is never objective but relative concerning the possibilities and *not* the potentials available (Savlulescu and Kahane 2017).

Hence, we can imagine that on the generation ships, the best offspring can be conceived via genetically modified embryos and potentially enhanced to ensure that the child is more resistant to the risks persistent in long-term space journeys. Savulescu and Kahane would argue that there would be nothing morally dubious about having offspring, even with genetic enhancements that were not the best *per se*. We argue that this point is morally criticisable. Even if we choose the best genetic modifications for our offspring does not mean that we automatically put them into a position to have a good life, given the various ethical issues delineated above. This position aligns better with the principle of parental responsibility forwarded by Bonnie Steinbock and Ron McClamrock (1994), which provides progenitors with a more stable condition concerning the selection of how they should support the kind of life their offspring will have and ensure that their life is worthy of living. Parents, hence, should actively conceive of what the best life for their child looks like, rather than a minimally worthy one, and actively endeavour to promote and support such a good life. Logically speaking, should all the arguments in this paper obtain, then the context of generation ships serves as a helpful context that demonstrates that such principles obtain regardless of their loci of application. Hence, what obtains on a generation ship or exoplanet obtain also on our home world, Earth.

6. Conclusions

In this paper, we explored the unique ethical issues that emerge when we consider the concept of generational ships designed for multiple generations to be born, live, and die to fulfil the mission of making humans a multi-planet species. We explored how it may not be *a priori* ethical to give birth on such ships, as is their innate function, simply if we guarantee the offspring a minimally sufficient life worth living, i.e., *de facto* abdicating to them the challenges and sacrifices that their original progenitors assumed when accepting their mission. We argue in this paper that such a position is not morally responsible, and that, before giving birth on such ships, and perhaps in the initial settlements on the destination exoplanets, the progenitors must ensure that their offspring live not only a minimally sufficient life worth living, but also a good life. We argue that human enhancement techniques are a suitable means for achieving both a minimally sufficient life and a good life for offspring on generational ships. Likewise, and philosophically important, the arguments used to support this thesis, if they obtain, obtain also for those currently living on Earth!

References

Avise, J.C. and Walker, D.E. 1998, "Pleistocene Phylogeographic Effects on Avian Populations and the Speciation Process", *Proceedings of the Royal Society of London, Series B: Biological Sciences*, 265, 1395, 457-63.

Ayala, F.J. 2015, "Cloning Humans? Biological, Ethical, and Social Considerations", *Proceedings of the National Academy of Sciences*, 112, 29, 8879-86.

Balistreri, M. and Umbrello, S. 2022a, "Space Travel Does Not Constitute a Condition of Moral Exceptionality: That which Obtains in Space Obtains also on Earth!", *Medicina e Morale,* 71, 3.

Balistreri, M. and Umbrello, S. 2022b, "Should the Colonisation of Space Be Based on Reproduction? Critical Considerations on the Choice of Having a Child in Space", *Journal of Responsible Technology*, 11, 100040.

Baum, S.D. 2015, "Confronting the Threat of Nuclear Winter", *Futures*, 72, 69-79.

Benatar, D. 2006, *Better Never to Have Been: The Harm of Coming into Existence*, Oxford: Clarendon Press.

Bostrom, N. 2016, *Superintelligence*, Oxford: Oxford University Press.

Buchanan, A. and Brock, D.W. 1986, "Deciding for Others", *Milbank Quarterly*, 64 (Suppl. 2), 17-94.

Butler, C.D. 2018, "Climate Change, Health and Existential Risks to Civilization: A Comprehensive Review (1989-2013)", *International Journal of Environmental Research and Public Health*, 15, 10, 2266.

Daly, M.J. 2000, "Engineering Radiation-Resistant Bacteria for Environmental Biotechnology", *Current Opinion in Biotechnology*, 11, 3, 280-85.

DeGrazia, D. 1995, "Value Theory and the Best Interests Standard", *Bioethics*, 9, 1, 50-61.

Del Mastro, A., Monaco, F. and Benyoucef, Y. 2021, "A Multi-User Virtual Reality Experience for Space Missions", *Journal of Space Safety Engineering*, 8, 2, 134-37.

Edwards, M. 2021 "Android Noahs and Embryo Arks: Ectogenesis in Global Catastrophe Survival and Space Colonization", *International Journal of Astrobiology*, 20, 2, 150-58.

Freitas, R. 1977, "Metalaw and Interstellar Relations", *Mercury,* 6, 3, 15-17.

Galarza, L. and Holland, A.W. 1999, "Critical Astronaut Proficiencies Required for Long-Duration Space Flight", (No. 1999-01-2096), *SAE Technical Paper*.

Garrard, E. and Wilkinson, S. 2006 "Selecting Disability and the Welfare of the Child", *Monist*, 89, 4, 482-504.

Glover, J. 2006, *Choosing Children: Genes, Disability, and Design*, Oxford: Clarendon Press.

Griffin, J. 1986, *Well-being: Its Meaning, Measurement and Moral Importance*, Oxford: Clarendon Press.

Hein, A.M., Pak, M., Pütz, D., Bühler, C., and Reiss, P. 2012, "World Ships: Architectures & Feasibility Revisited", *Journal of the British Interplanetary Society*, 65, 4, 119.

Hofmann, B. 2017, "Limits to Human Enhancement: Nature, Disease, Therapy or Betterment?", *BMC Medical Ethics*, 18, 1, 56.

Joshi, Y. and Mardon, A. 2021, "Using Virtual Reality for Long-Duration Space Missions", *Technium Soc. Sci. J.*, 20, 627.

Kass, L.R. 2003, "Ageless Bodies, Happy Souls: Biotechnology and the Pursuit of Perfection", *The New Atlantis*, 1, 9-28.

Kudlek, K. 2021, "Is Human Enhancement Intrinsically Bad?", *Medicine, Health Care and Philosophy*, 24, 2, 269-79.

Leisman, G., Machado, C., Melillo, R., and Mualem, R. 2012, "Intentionality and 'Free-Will' from a Neurodevelopmental Perspective", *Frontiers in Integrative Neuroscience*, 6, 36.

Lester, D. 2013, "Measuring Maslow's Hierarchy of Needs", *Psychological reports*, 113, 1, 15-17.

Levy, N. 2016, "Would it Be Immoral to Send out a Generation Starship?", *Aeon*, retrieved 9 June 2022, from https://aeon.co/ideas/would-it-be-immoral-to-send-out-a-generation-starship.

Magni, F. 2021, "In Defence of Person-Affecting Procreative Beneficence", *Bioethics*, 35, 5, 473-79.

McInnes, C.R. 2009, "Mars Climate Engineering Using Orbiting Solar Reflectors", in Badescu, V. (ed.), *Mars*, Berlin/Heidelberg: Springer.

McMahan, J. 1998, "Wrongful Life: Paradoxes in the Morality of Causing People to Exist", in Coleman, J. and Morris, C. (eds.), *Rational Commitment and Social Justice: Essays for Gregory Kavka*, Cambridge: Cambridge University Press, 208-47.

McMahan, J. 2002, *The Ethics of Killing: Problems at the Margins of Life*, New York: Oxford University Press.

McMahan, J. 2009, "Asymmetries in the Morality of Causing People to Exist", in Roberts, M. and Wasserman, D., *Harming Future Persons: Ethics, Genetics and the Non-Identity Problem*, New York: Springer, 49-70.

Orfali, K. 2004, "Parental Role in Medical Decision-Making: Fact or Fiction? A Comparative Study of Ethical Dilemmas in French and American Neonatal Intensive Care Units", *Social Science & Medicine*, 58, 10, 2009-22.

Pagallo, U. 2018, "Apples, Oranges, Robots: Four Misunderstandings in Today's Debate on the Legal Status of AI Systems", *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376, 2133, 20180168.

Pellerin, C.J. 2009, *How NASA Builds Teams: Mission Critical Soft Skills for Scientists, Engineers, and Project Teams*, Hoboken: John Wiley & Sons.

Rakić, V. 2017, "Compulsory Administration of Oxytocin Does not Result in Genuine Moral Enhancement", *Medicine, Health Care and Philosophy*, 20, 3, 291-97.

Salamon, N., Grimm, J.M., Horack, J.M., and Newton, E.K. 2018, "Application of Virtual Reality for Crew Mental Health in Extended-Duration Space Missions", *Acta Astronautica*, 146, 117-22.

Sandel, M.J. 2007, *The Case against Perfection: Ethics in the Age of Genetic Engineering*, Cambridge, MA: Harvard University Press.

Sandel, M. 2009, "The Case Against Perfection: Children, Bionic Athletes, and Genetic Engineering", in Savulescu, J. and Bostrom, N. (eds.), *Human Enhancement*, Oxford: Oxford University Press, 71-89.

Savulescu, J. 2001, "Procreative Beneficence: Why We Should Select the Best Children", Bioethics, 15, 5-6, 413-26.

Savulescu, J. and Kahane, G. 2009, "The Moral Obligation to Create Children with the Best Chance of the Best Life", *Bioethics*, *23*, 5, 274-90.

Savulescu, J. and Kahane, G. 2017, "Understanding Procreative Beneficence", in Francis, L. (ed.), *The Oxford Handbook of Reproductive Ethics*, Oxford: Oxford University Press, 592-622.

Scanlon, T. 2000, *What We Owe to Each Other*, Cambridge, MA: Belknap Press.

Schon, K.R., Parker, A.P.J., and Woods, C.G. 2020, "Congenital Insensitivity to Pain Overview", in Adam, M.P, Everman, D.B, Mirzaa, G.M et al. (eds.), *GeneReviews®[Internet]*, Seattle: University of Washington, available from https://www.ncbi.nlm.nih.gov/books/NBK481553/.

Schulze-Makuch, D., Heller, R., and Guinan, E. 2020, "In Search for a Planet Better than Earth: Top Contenders for a Superhabitable World", *Astrobiology*, 20, 12, 1394-1404.

Sheetz, M. 2021, "Elon Musk Wants SpaceX to Reach Mars so Humanity Is Not a 'Single-Planet Species'", *CNBC*, retrieved 9 June 2022.

Sidhu, J.S., Joshi, S.K., Gündoğan, M., Brougham, T. et al. 2021, "Advances in Space Quantum Communications", *IET Quantum Communication*, 2, 4, 182-217.

Singh, J.S., Abhilash, P.C., Singh, H.B., Singh, R.P., and Singh, D.P. 2011, "Genetically Engineered Bacteria: an Emerging Tool for Environmental Remediation and Future Research Perspectives", *Gene*, 480, 1-2, 1-9.

Smith, C.M. 2014, "Estimation of a Genetically Viable Population for Multigenerational Interstellar Voyaging: Review and Data for Project Hyperion", *Acta Astronautica*, 97, 16-29.

Steinbock, B. and McClamrock, R. 1994, "When Is Birth Unfair to the Child?", *Hastings Center Report*, 24, 6, 15-21.

Struster, J. 2010, *Behavioral Issues Associated With Long Duration Space Expeditions: Review and Analysis of Astronaut Journals*, Santa Barbara, CA: NASA, retrieved from https://ntrs.nasa.gov/citations/20100026549

Szocik, K. 2021, "Humanity Should Colonize Space in order to Survive but not with Embryo Space Colonization", *International Journal of Astrobiology*, 20, 4, 319-22.

Tajmar, M. and Bertolami, O. 2005, "Hyopthetical Gravity Control and Possible Influence on Space Propulsion", *Journal of Propulsion and Power*, 21, 4, 692-96.

Umbrello, S. and Baum, S.D. 2018, "Evaluating Future Nanotechnology: The Net Societal Impacts of Atomically Precise Manufacturing", *Futures*, 100, 63-73.

Vink, J. 2020, "Enfranchising Animals in Legal Institutions: Fundamental Legal Rights", *The Open Society and Its Animals*, 263-335.

Wilkinson, D.J., 2011, "A Life Worth Giving? The Threshold for Permissible Withdrawal of Life Support from Disabled Newborn Infants", *The American Journal of Bioethics*, 11, 2, 20-32.

Wyatt, J. 2005, *Quality of Life*, available at http://www.cmf.org.uk/literature/content.asp?context=article&id=1702, accessed October 24, 2022.

Book Reviews

Jacobsen, Michael Hviid, *The Anthem Companion to Zygmunt Bauman*. London: Anthem Press, 2023, pp. 220.

Zygmunt Bauman's legacy as a leading thinker on modernity and its discontents is explored with new insights in Michael Hviid Jacobsen's edited volume, *The Anthem Companion to Zygmunt Bauman*. This review delves into how well the book captures the essence of Bauman's thought across different stages of his career.

Undoubtedly, Bauman represented one of the most influential representatives of contemporary sociological thought. Born in Poznań on 19 November 1925 to Jewish parents, in 1939, following the Nazi invasion of Poland, he was forced to flee to the Soviet occupation zone where he enlisted in a Soviet military unit. From the end of the conflict to 1948, he took part in some operational tasks for Soviet military espionage. After the war, he began studying and eventually teaching sociology at the University of Warsaw. In 1968, the incessant rise of anti-Semitism among the various levels of Polish society, pushed many Polish Jews to emigrate abroad, including Bauman. He therefore emigrated first to Israel, where he taught at Tel Aviv University, and later accepted a professorship of sociology at the University of Leeds, where he taught from 1971 to 1990. A resident of Leeds ever since, Bauman passed away on 9 January 2017 at the age of 91.

Scholarly rooted in the great European intellectual tradition of the second half of the 20th century, Bauman's thought can be contextualized in the frame of the theorizations of Karl Marx, Friedrich Engels, Antonio Gramsci, Émile Durkheim, Max Weber, Georg Lukács, and Georg Simmel. Generally, Bauman belonged to the tradition of the Frankfurt School's Critical Theory. He gained international fame thanks to his studies regarding the connection between the culture of modernity and totalitarianism, particularly in the case of Nazism and the Holocaust. Also, among his innumerable research interests, he focused on the transition from modernity to postmodernity and the related ethical issues. Notably, he compared the concept of modernity and postmodernity to the solid and liquid state of society respectively, underscoring that while in the modern age everything was given as a solid construction, the post-modern (or late-modern) "liquid" society was characterized by no clear outlines and certainties, giving space to insecurity and fear.

In this frame, *The Anthem Companion to Zygmunt Bauman* represents a valuable edited book comprising most of the key facets of Bauman's sociological thought, which are introduced by the authors in an all-encompassing, wide-ranging, and well-nuanced perspective. Specifically, ten chapters organize the book, covering different themes of Bauman's thinking and work, each focusing on topics and ideas that were characteristic of Bauman's way of doing and writing sociology. In its thorough analysis, the book retraces the four key phases of Bauman's sociological thought, namely the Marxist phase (1960s-1970s), the critique of modernity phase (late 1980s), the post-modern phase (1990s), and the liquid-modern phase (early 2000s)—with its focus on the conception of "liquid modernity". As highlighted in the introductory section of the book, all the main themes of Bauman's sociological endeavours are scrutinized throughout the work, including the concepts of modernity, post-modernity, liquid modernity, morality, ethics, culture, the Holocaust, Jewish identity, freedom, religion, poverty, inequality, utopia, "retrotopia", nostalgia, "adiaphorization", consumerism, identity, globalization, love, fear, security, ambivalence, suffering, the working class, the stranger,

Argumenta 10, 1 (2024): 469—480
ISSN 2465-2334
DOI 10.14275/2465-2334/202419.boo
First published: 30 November 2024

the "other", the migrant, and death. Given the vastity of Bauman's scholarship and the variety of its themes of research, in addition to highlighting Bauman's central themes, the volume also examines the more neglected areas of his work. From a methodological perspective, the book reconstructs Bauman's thinking using the sociologist's primary literature, the classical sociological and philosophical literature that had influenced his theorizations, and the coeval and successive secondary scholarly literature and debates built around his works.

In the first chapter "Zygmunt Bauman: Weberian Marxist?", Peter Beilharz analyses the influence of Marx and Weber in Bauman. The chapter first discusses the notion of "Weberian Marxism" and then, in separate sections, examines in-depth the influence of respectively Marx and Weber in Bauman's sociological theory. Specifically, the chapter deepens Bauman's analysis of social classes and elites through scrutinizing his works *Between Class and Elite: The Evolution of the British Labour Movement. A Sociological Study* (1972), *Officialdom, and Class: Bases of Inequality in Socialist Society* (1974), and *Memories of Class: The Pre-History and After-Life of Class* (1982). Here, Marx and Weber appear as two paramount references in Bauman's theoretical posture, contributing to building some of the sociologist's paradigmatic assumptions linked to class identity, inequality, consumerism, and bureaucratization. Special attention is given to the concept of "modern" rationalization, a *Meistermotif* of Bauman's understanding of "solid" modernity. As modern incarnations of Prometheus and Sisyphus respectively, Marx and Weber embody the principles of revolution and repetition, change and stasis—which Bauman reconnects to the unfolding of modern society and its transition to post-modernity.

The second chapter "A Freudian without Psychology: The Influence of Sigmund Freud on Zygmunt Bauman's Sociology" by Matt Dawson highlights the sociological elements in Freud's works and how they affected Bauman. The author suggests that Bauman makes use of Freud's insights in five specific areas: the Freudian conflictual pendulum between freedom and security, in which Bauman highlights society's trade-off between uncertain freedom and restricting security; the relation between reality and the pleasure principle, that is core to the transition from solid to liquid modernity; the concepts of ambivalence and death, the former seen as the element that rational modernity has sought to remove from its well-ordered "social garden" also through expunging the latter; the construction of the identity of the stranger and the idea of community based on exclusivist, anti-Kantian understandings of the "others"; and the shifting forms of narcissism, considered as a widespread tendency in the hyper-individualistic liquid modern life.

Turning to the third chapter "Modernity and the Holocaust: Exploring Zygmunt Bauman's Contribution to the Sociology of the Holocaust", Adele Valeria Messina deals with the fundamental Baumanian topic of the uncanny relationship between modernity and the Jewish *Shoah*. Naturally, the key reference for the chapter's analysis is Bauman's eminent masterpiece *Modernity and the Holocaust* (1989), in which the sociologist provocatively asks whether the massive slaughter of Europe's Jews represented a return to a barbaric past or a nasty aspect of modernity. Bauman's core argument vis-à-vis the Holocaust rested on the idea that the extermination of the Jews was an outcome of the concepts of modern rationality and bureaucracy. In this respect, the chapter's author reports an exhaustive review of both historical and sociological scholarly literature on Bauman's fundamental book. Unlike the sociological secondary literature, the historical one is affected by the novelties following the opening of the archives after the demise of

the socialist bloc. Generally, this literature review unveils a core critique towards Bauman's *Modernity and the Holocaust*, i.e., the underestimation of the role of anti-Semitism in Germany (and elsewhere in Europe) because of the belief in the genocidal potential of any modern nation-state. Echoing Hannah Arendt's ideas of "desk murderer" (*Schreibtischtäter*)—an impersonal bureaucrat performing administrative functions in rational mode without regard to moral consequences—and "banality of evil"—a concept, as highlighted by the literature,[1] originally conceived by Everett C. Hughes[2]—Bauman believed that the Holocaust had been possible because of the process of rationality in modern bureaucracy. Thus, Auschwitz was conceived as an example of "murderous Fordism", that, while producing largescale death, resembled the mass production of goods typical of modern industrial society.

Instead, in chapter four "Zygmunt Bauman and the Continental Divide in Social Theory" Stjepan G. Meštrović, Michael Ohsfeldt, and Jacob Hardy explore how Bauman's sociology is deeply rooted in European rather than American sociological tradition in terms of attitudes, origins, values, and even prejudices. Unlike American sociology, which is markedly optimistic, pragmatic, and empirical, Bauman's social theory—in line with the writings of Marx, the critical theorists, European existentialists, and philosophers—is pessimistic, unempirical, idealistic, and sceptic towards progress.

Consistent with the previous, chapter five "Zygmunt Bauman on the West: Re-Treading Some Forking Paths of Bauman's Sociology" by Jack Palmer deals with the issue of sociology's "Eurocentrism". Here, Bauman's reflections on colonialism and decolonization, the Jewish question, the interpretation of modernity, and the communist project in central-eastern Europe are clarified, showing how Bauman understood the contemporary discussion about Eurocentrism in sociology and underscored the importance of "decolonizing" its canons and operative concepts. Also, the chapter highlights Bauman's so-called "cultural turn", i.e., his humanist revision of Marxism and the elaboration of a cultural sociology hinging on semiotics and hermeneutics.

Furthermore, in chapter six "Death as a Social Construct: Zygmunt Bauman and the Changing Meanings of Mortality" Michael Hviid Jacobsen and Nicklas Runge evaluate Bauman's main work on the theme of death and its meanings to humans and societies *Mortality, Immortality and Other Life Strategies* (1992). Bauman applied the specific notion of "deconstruction" to describe how society in different ways seek to turn death and immortality into manageable or acceptable concepts that can keep people occupied and engaged as a distraction from real death ("death proper"), which remains an unsolvable mystery. In this vein, Bauman analytically distinguished between modern society's "deconstruction of mortality" and postmodern society's "deconstruction of immortality", both serving the purpose of making life meaningful despite the inevitability of death.

Then, chapter seven "Zygmunt Bauman and the 'Nostalgic Turn'" by Dariusz Brzeziński focuses on Bauman's vision of "retrotropia", which is described as a multidimensional process of turning to the past as a reaction to the increasing uncertainty and unpredictability of contemporary conjunctures. In this sense, the author scrutinizes Bauman's late work *Retrotopia* (2017), underlining the main

[1] Messina, A.V. 2020, "New Perspectives on Everett C. Hughes's Sociological Works about the Holocaust, 1930s–1980s", *Journal of Modern Jewish Studies*, 19 (3), 337-361.

[2] Hughes, E.C. 1962., "Good People and Dirty Work", *Social Problems*, 10 (1), 3-11.

ideas behind the concept in the context of liquid modernity. "Retrotopia" is considered a consequence of nostalgia, which tends to reappear as a defence mechanism in times of accelerated rhythms of life and historical upheavals. Recent examples of nostalgic turns comprise Donald Trump's presidential campaign (with the slogan "Make America Great Again"), Brexit, Russia's neo-imperial posture vis-à-vis the war in Ukraine, and post-Covid-19 "defensive" social reactions. Per Bauman, post-modern nostalgic society would be characterized by a resurgence of violence due to distrust towards individuals and institutions ("back to Hobbes"), tribal forms of solidarity ("back to tribes"), rampant socioeconomic inequalities ("back to inequality"), and self-centered reaffirmation that places security above freedom.

Moving to chapter eight "Bauman on Borders: The Role of *Our* Door in the Construction of the Stranger", Shaun Best describes Bauman's complex understanding of how the stranger as a distinctive analytical, social, and cultural category comes into being in contemporary sociological discourse. According to the sociologist, in current liquid modernity the stranger appears in various forms, including the poor, the flawed consumer, the unwanted foreigner, the forced refugee, and the reluctant migrant. In this context, the concept of border is paramount by "constructing" and "deconstructing" the stranger. Moreover, while in solid modernity the stranger was seen predominantly as an element that spoiled the harmony of the "social garden" or "garden state", compelling the authorities to correct, repair, assimilate or ultimately exterminate him, in liquid modernity the stranger ignites fear and insecurity, making individuals adopt a defensive mechanism. Crucially, due to the hyper-individualistic, post-liberal tendencies and the lack of forms of communitarianism, in liquid modernity any other person beyond the individual is a potential stranger.

The next contribution in chapter nine "Seeking Windows in a World of Mirrors: Zygmunt Bauman's Difficult Art of Conversation" by Mark Davis and Elena Álvarez-Álvarez introduces Bauman's last books, which are structured in the form of conversations. Here, some fundamental aspects of liquid society are assessed, including the concepts of "liquid evil" and "adiaphorization" (i.e., the cancellation or denial of moral impulse through mounting social indifference and the loss of collective solidarity). Special attention is dedicated to the analysis of the features of contemporary social medias, which enhance a "world of noise", discarding real communication. The central importance given by liquid modernity to social media and virtual networks validates the need to rediscover the importance of conversations and dialogues in the frame of mutual respect and understanding.

Finally, chapter ten "Ambivalence (Not Love) is All Around: Zygmunt Bauman and the (Ineradicable) Ambivalence of Being" by Michael Hviid Jacobsen highlights Bauman's reflexions on ambivalence—expressed chiefly in the work *Modernity and Ambivalence* (1991)—which is considered a fundamental condition of human existence and social life. Originally conceived by the psychological and psychiatric literature, the notion of ambivalence in post-modernity indicates that individuals are increasingly confronted with an unprecedented number of choices and an equally unprecedented range of contradictions, leading to chaos, uncertainty, and insecurity. Historically, ambivalence and ambiguity have been overcome either through incorporation and assimilation ("anthropophagic" strategy) or expulsion and destruction ("anthropoemic" strategy) of the deviant, the

strange, the alien, and the ambiguous. Still, paradoxically, order could not exist without ambivalence since it manifests as a reaction to it.

In conclusion, given its multifaceted nature and variegated perspectives, the volume represents a thorough and clear compendium of Zygmunt Bauman's sociological thought. The book's main merit lies in the analysis of the more overlooked concepts of Bauman's sociology, while also including its mainstream themes. Generally, the book is clear, refined, and well-written, arousing interest and curiosity in the reader. In terms of its scope and readership, the study can be considered a precious—albeit auxiliary—tool for researchers and scholars whose fields of research embrace sociology, political science, political theory, and philosophy, as well as for a broader audience willing to engage in key elements of Bauman's sociology.

Jagiellonian University of Kraków PAOLO PIZZOLO

[This book review has been developed in the frame of the project "Promoting Order at the Edge of Turbulence (POET)" that is conducted in the Center for International Studies and Development (CISAD) at the Jagiellonian University in Krakow (Poland). The project is co-financed by the Polish National Agency for Academic Exchange under the NAWA Guest Professorship program and the Polish National Agency for Academic Exchange within the NAWA Chair program. The author wishes to acknowledge the financial assistance of the NAWA Grant (PPN/PRO/2020/1/00003/DEC/1) from the Polish Academic Exchange Council and NCN grant (ZARZADZENIE NCN 94/2020) from the Polish National Science Council.]

Kitcher, Philip, *Moral Progress*.
New York: Oxford University Press, 2021, pp. xix + 200

What criteria can we appeal to for qualifying a change in what we believe and do as an instance of moral progress? Do these criteria necessarily presuppose a reference to a universal and objective moral truth? And how can we promote progressive moral changes? These are the fundamental questions that Philip Kitcher's latest book, *Moral Progress*, tackles.

The book presents, in written form, the text of the first *Munich Lectures in Ethics* that Kitcher delivered at LMU in 2019. As often happens with this type of publication, the organization of the content is less than optimal, the argumentation is sometimes a bit rough, and the comparison with the literature on the subject limited. But the text, on the other hand, maintains some of the pleasant intellectual agility usually associated with lectures of this sort and level. Moreover, it is accompanied and complemented by three sets of excellent replies from three outstanding philosophers, namely Amia Srinivasan, Susan Neiman, and Rahel Jaeggi.

The first chapter of Kitcher's text provides an overview of his pragmatist and anti-realist theory of moral progress. The two following chapters deal with specific issues related to this theory although, in doing so, they add much more than just a few finishing touches. The second part, dedicated to the problem that the phenomenon of false consciousness represents for Kitcher's theory, actually does much more than proposing a solution it, as we will see. The third and final part is dedicated to clarifying the limited and quite specific ways in which this pragmatist theory allows us to frame the notion of progress in terms of "truth" and

"moral knowledge". The readers with little interest or sympathy for the pragmatist tradition—within which the conceptualization of truth is notoriously a long-standing issue—will be pleased to discover that they can skip this part without missing out on much.

For reasons that will be clear in a minute, a good place to start outlining the contours of Kitcher's theory of moral progress is his evolutionary account of morality itself, which he offers in part II. According to Kitcher, morality represents a bio-cultural innovation specific to the species Homo sapiens and it emerged in the late Paleolithic (49). According to Kitcher, "the best available picture of pre-moral hominin—and human—life portrays our predecessors as possessing a capacity for identifying the desires and intentions of their fellow band members and for adjusting their behavior so as to engage in joint projects with others" (50). For social creatures whose survival depends on the group to which they belong, this ability, which Kitcher refers to as "responsiveness" (50) is somewhat necessary to ensure some degree of cooperation within the group, and thus the survival of the group itself.

This limited responsiveness, for Kitcher, was likely shared by the first sapiens, who spent the vast majority of their stay on planet Earth (which began around 300,000 years ago) organized in small bands of hunter-gatherers. This limited responsiveness constituted a limit to intra-group cooperation and, thus, to the maximum size a group could hope to reach (51-52). Morality, against this background, functionally presents itself as a social technology that allowed us to overcome this impasse and increase the responsiveness of our species' members, enabling the formation of larger and more cohesive groups. What mechanisms allowed its emergence? Kitcher provides only a few details on this matter, and the reader who wants to know more will have to return to the first four chapters of *The Ethical Project* to which Kitcher's current account remains substantially faithful.[1]

How does the theme of moral progress fit into these views of our evolutionary past? Just as in *The Ethical Project* (2011, chap. 6), Kitcher establishes the continuity between the two themes through a functionalist perspective. On such a perspective, the evolutionary understanding of the original function of morality allows us to define what moral progress consists of. More specifically, if the original function of the moral device is to compensate for the limits of human responsiveness, i.e., to correct and amplify their limited ability to adopt others' perspectives, needs, interests, and desires, then moral progress is primarily "a matter, if you like, of improving this device, the responsiveness amplifier" (148). Historical cases such as the abolition of slavery, the emancipation of women, and the acceptance of homosexual relationships are interpreted by Kitcher in these terms.

As anticipated, Kitcher characterizes this conception of moral progress as essentially pragmatic and anti-teleological, contrasting it from the outset with the realist conception that sees moral progress as an approximation to moral truth, a progressive activity of discovering previously ignored bits of moral knowledge (15). Instead of seeing moral progress as an alignment of our beliefs with reality based on epistemic standards, we should see it as the solution to practical problems afflicting the moral architecture of society: not progress *towards* truth or correct moral beliefs, but progress *from*, based on overcoming limitations and problematic situations (25).

[1] Kitcher, P., 2011, *The Ethical Project*, Cambridge, MA: Harvard University Press.

Conceiving moral progress in these terms, Kitcher argues, has several advantages. A very important advantage is that, starting from this pragmatic conception, we can have a better understanding of what happens when a society progresses morally, and use this understanding to outline a method that helps us in identifying morally problematic situations and ways to resolve them for the best.

The development of this method is the fundamental contribution of the volume. It is articulated in a long series of steps that occupy much of the first and second chapter. Simplifying, we can summarize it as follows. First, if an individual or a group complains about a situation despite the current moral code allowing it, this situation is to be considered prima facie problematic and is to be further examined to evaluate the actual justification of the initial complaint (34-36). How should this examination be conducted? Kitcher appeals here to the regulatory model of an "ideal conversation"—an ideal that leads him to label his view as "democratic contractualism" (57-58). According to this model, problematic situations are those that a society would see as such if representatives of all involved viewpoints, having to deliberate together based on justified factual beliefs and in conditions of deep mutual respect and sympathy, would agree on their problematic character (37). The same model then comes into play in defining the standard that makes a change a progress. A proposal is a justified resolution of a problematic situation only if the transition from the problematic situation to the proposed one would be accepted in an ideal conversation where the perspectives of all stakeholders are represented (38).

What should be done in cases where a situation is objectively problematic but no one complains about it, perhaps because they have internalized the prejudices of a given culture despite being victims of it? In the second chapter of the book (aptly titled "Problems of False Consciousness"), the proposed method is integrated to address these cases. Even in the absence of actual challenges, Kitcher clarifies, "societies should periodically check whether the restrictions they impose on the range of appropriate self-models for a certain subgroup can be justified" (67). The kind of social experimentation proposed by John Stuart Mill and Harriet Taylor in their time to question the validity of Victorian prejudices about gender remains for Kitcher the principal tool for this purpose (68).

This proposal will not sound extremely original to those who have been following the debate for some years. Peter Railton and, more recently, Elizabeth Anderson have advanced similar and influential ideas, and it is a pity that Kitcher does not spend more resources clarifying how his position differs from theirs, especially from Anderson's, who share with Kitcher a broadly pragmatist view.[2]

Additionally, there are several problems that Kitcher's text leaves open or does not address entirely satisfactorily. For example, one might wonder if the theoretical framework offered by Kitcher truly does away with notions such as "moral truth". In fact, the appeal to an ideal deliberation procedure characterized by sympathy and mutual respect seems to presuppose and embody, in some way, the idea that at least the judgment "everyone has an equal right to participate in

[2] See Railton, P., 1986, "Moral realism", *Philosophical Review*, 95 (2), 163-207; Anderson, E., *Social movements, experiments in living, and moral progress: Case studies from Britain's abolition of slavery.* The Lindley Lecture, University of Kansas, https://kuscholarworks.ku.edu/handle/1808/14787; Anderson, E., 2015, "Moral bias and corrective practices: A pragmatist perspective", *Proceedings and Addresses of the American Philosophical Association,* 89, 21-47.

this conversation" is true in a strong and non-pragmatic sense. And what is this if not a moral judgment? Furthermore, one cannot but wonder whether his methodological proposal for fostering progress presupposes an overly rationalist view of the phenomenon, underestimating the importance of volitional obstacles, rather than cognitive ones, that it must overcome. After all, many people in many circumstances know what would be morally right to do, but this is often insufficient to motivate them to do it. How can the ideal conversation (or some institutional embodiment of it) address this problem? Kitcher, as I have said, leaves these and other questions unanswered.

Nevertheless, for the clarity and the degree of detail with which it is articulated, his contribution remains a highly recommended read for anyone interested in the theme of moral progress.

University of Milan FRANCESCO TESTINI

[This book review was developed in the frame of the project No. 2021/43/P/HS1/02247 co-funded by the Narodowym Centrum Nauki and the HORIZON EUROPE Marie Sklodowska-Curie Actions [grant agreement no. 945339]. For the purpose of Open Access, the author has applied a CC-BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission.]

McKenzie, Kerry, *Fundamentality and Grounding*.
Cambridge: Cambridge University Press, 2022, pp. 74.

Fundamentality and Grounding is an academic publication that stands out in the landscape of contemporary metaphysics. Its general intent is to assess some of the central issues that arise around the widely debated notion of "grounding", according to a naturalistic methodological viewpoint proper to the metaphysics of science. Such methodology aims at understanding what is possible to "import" from science to "update" or "inform" metaphysics and how to implement this task. Specifically, three issues are considered:

- What are the relationships between the notions of fundamentality and grounding?
- Is the notion of grounding used in the various philosophical discussions ambiguous? In other words, are there substantially different types of grounding?
- Should we exclude the possibility of infinite regress in the order of grounding?

McKenzie is clear from the outset in stating that the concepts of fundamentality and grounding are intimately linked. As it shall be clear, she regards "grounding" as a "level connecting explanation" (8) among facts or entities belonging to different metaphysical categories. Grounding bears interesting relationships to the notion of ontological priority, which is undoubtedly the most common way of thinking about fundamentality: *x* is fundamental if there is no *y* ontologically prioritized over *x*. The interest in grounding is motivated by its close connection with the concept of fundamentality, so conceived. The reason for this interest, McKenzie explains, arises from the fact that fundamentality plays a key role in the way metaphysics is often understood, namely, as the study of the fundamental.

In what follows, I critically review Chapters 2, 3, and 4 of *Fundamentality and Grounding*, the stated purpose of which is to naturalize the metaphysics of grounding, grounding being a relation often relegated to a priori metaphysical analysis

only. By naturalization, in this case, McKenzie means the reevaluation of some important features commonly attributed to the notion of grounding in light of what science, in the present case, physics, says. Two positions characterize McKenzie's philosophical stance. They emerge clearly in the third and fourth chapters:

- grounding is not a single relation, but various relations of grounding must be recognized;
- in science, infinite explanatory regressions, often deemed vicious by metaphysics, are permitted. Consequently, grounding relations, closely tied to the concept of metaphysical explanation, can be involved in such regressions without concern (as metaphysics must heed the insights from science.).

The second chapter is aimed at identifying how grounding should be understood. This task is particularly challenging due to the high complexity and multitude of positions expressed regarding this notion. Philosophers have tried to make sense of the following ideas:

- the world possesses a gradually *stratified* structure;
- such stratification obtains in virtue of the *explanatory determination* of one level over another;
- there exists a fundamental, i.e., ontologically prioritized level, which *explanatory determines* the others.

Capturing the specifics of such a determination required the introduction of a new notion, that of grounding, and the reasons behind this necessity are the following:

- causation is not the relationship of determination sought. Indeed, the concept of causation connects different temporal moments, while the notion of explanatory determination must be capable of establishing a hierarchy between levels (e.g., Schaffer 2012)[1];
- modal notions are inadequate to capture explanatory notions, such as that of explanatory determination (e.g., Sider 2020)[2];
- the notion of determination has quite different characteristics from those of ontological dependence, not the least of which is that it entertains a different relation to the notion of priority: to say that *x* depends, at least in part, ontologically on *y* implies that y has priority over *x*. If *x* depends on *y*, however, the existence of *x* also implies in a metaphysically necessary way that of *y*. From a standpoint of determination, therefore, *x* is prioritized over *y*.

The notion of grounding often appeals to the notion of *metaphysical explanation*. Remarkably, the "grounding school" divides into two main families, the unionist and the separatist. Unionists claim that the grounding relation coincides exactly with the metaphysical explanation, while separatists do not. The separatists claim that grounding relations are what *justify* or what *underly* explanations. There appears to be a good reason to avoid treating the notion of metaphysical explanation according to a single notion. In fact, a unifying approach runs the risk of slipping into unclear theoretical involutions. Among them, for example, one can find such questions as "what is the grounding of the notion of grounding?" According to

[1] Schaffer, J. 2012, "Grounding, Transitivity, and Contrastivity", in F. Correia and B. Schnieder (ed.), *Metaphysical Grounding: Understanding the Structure of Reality*, Cambridge University Press, 122-138.
[2] Sider, T. 2020, "Ground Grounded", *Philosophical* Studies, 177 (3), 747-767.

Wilson (2014)[3] and Koslicki (2015),[4] these envelopments of the notion of grounding have been dictated by an abuse of the a priori metaphysics approach, which seems to self-generate problems for itself, to the detriment of their relevance. The author's view looks favorably on the vision of a separatist grounding approach and argues that there are theoretical resources in elementary physics that push for such an approach, which she sets out to defend in the book.

In the third chapter McKenzie specifies how grounding can be understood as a *connector of levels*. After all, the author argues, there are two ways of connecting levels (and I believe this constitutes this book's major contribution to the existing literature on grounding). The first way connects levels belonging to the same category, which can be, for example, the category of physical objects, physical properties, physical laws, and so on. The second way, on the other hand, is to understand grounding as a connector between transcategorical levels, that is, as a connector of different categories. For McKenzie, the distinction between these two kinds of "connection between levels" is well founded in that it refers to two different kinds of metaphysical explanations. Levels that are connected by remaining within the same category are called "levels of nature" by McKenzie. In contrast, levels of the second kind, that is, levels between different categories, are called "levels of metaphysics".

As an example, within the category of "objects", it is possible to recognize the level of ordinary objects and the level of subatomic objects such as protons or electrons. Following McKenzie's analysis, these two levels are levels of science. The distinction between these two levels within the same category is attributed, according to McKenzie, to the recognition of a priority status of subatomic entities over ordinary ones. Such recognition pertains to the science. The category of "objects" is just one of the categories that one can introduce. Alongside it, it is possible to admit the existence of the categories of properties or even physical laws. Now, these different categories represent the various levels of metaphysics, and the priority relations among them belong to metaphysics and are obtained through the grounding relations between the different categories.

The distinction McKenzie outlines thus raises the following question: what relationship exists between the levels of science and the levels of metaphysics? Given the different relationships in each hierarchy, these questions have no obvious answers. Nonetheless, if one thing becomes clear from McKenzie's analysis, it is that to speak of "stratified" metaphysics acquires a specific meaning, since, as it turns out, one is faced with two different hierarchies, on the one hand that of the levels of nature and on the other that of the levels of metaphysics. By appealing to the Humean mosaic, McKenzie contends it is not possible to examine the levels of nature based on those of metaphysics and vice versa. The moral to be drawn from this, according to McKenzie, is that there are two notions of fundamentality, and thus priority, that are not inter-reducible. One is faced with a pluralist thesis about priority that favors a very specific insight: the levels of nature and those of metaphysics establish two different dimensions of priority. The hierarchical direction of the levels of nature is thus essentially different from the hierarchical direction of the levels of metaphysics. This "multi-dimensionality" aspect has, in the

[3] Wilson, J.M. 2014, "No Work for a Theory of Grounding", *Inquiry: An Interdisciplinary Journal of Philosophy*, 57 (5-6), 535-579.
[4] Koslicki, K. 2015, "The Coarse-Grainedness of Grounding", *Oxford Studies in Metaphysics*, 9, 306-344.

author's view, been seldom the subject of philosophical debate and, indeed, often overlooked. Indeed, a considerable number of philosophers have often argued that the levels of metaphysics go deeper than the levels of physics as "metaphysics 'takes things a level deeper' than physics" (33). However, such a comparison implies a certain degree of commensurability between the two types of levels, which McKenzie excludes on the strength of her analysis. Ultimately, through the plurality of priority relations, one must recognize a plurality of relations of metaphysical explanation. Since grounding and metaphysical explanation are closely related (and often even identified), McKenzie's argument thus far reveals direct implications for the supposed "unity" of grounding.

In chapter four, McKenzie addresses the following question: is the grounding relationship well-founded? That is, must every grounding sequence (or chain) end at some point, a thesis known as foundationalism? If so the existence of every non-foundational entity is grounded in a set of foundational entities. McKenzie believes that discussing the foundationalism of grounding is important, if only to understand whether the definition of metaphysics as the study of the fundamental is, for all intents and purposes, acceptable. How should we characterize metaphysics in case a fundamental level doesn't exist? McKenzie argues that foundationalism is a thesis assumed almost at the axiomatic level, or at the level of metaphysical law, supported often more by mere intuition than by actual philosophical justification. McKenzie asks the following questions:

- what are the criteria for determining that a regression to infinity is vicious?
- do regressions to infinity of a sequence of grounding relations satisfy such criteria?
- does satisfying such criteria mean incurring some kind of metaphysical contradiction?

There are two theses that McKenzie proposes about the last questions:

1. first, there is no reason to think that an infinite sequence of grounding relations must necessarily be vicious;
2. second, it is argued that a form of "viciousness" is present in every regress to infinity known by means of scientific methods.

To justify thesis 1, McKenzie argues that regressions to infinity are not necessarily vicious for grounding. For them to be so, "what explains" (*explanans*) and "what is explained" (*explanandum*) must share the same "form" at each stage of the regress. For McKenzie, the viciousness of an infinite regress emerges as a "function of the explanatory interests" (54) we have along with the degree of abstraction of the *explanandum*. Since the degree of detail in science is highly refined and its aspirations are less abstract, there is no *a priori* reason to argue that infinite regressions don't arise in science. To justify thesis 2, McKenzie argues that even though there is not necessarily form invariance for the metaphysical explanations proposed by science, those involved in infinite chains nevertheless exhibit such uniformity. This is sufficient to label them as vicious. A case-study offered by a physical theory proves that infinite regressions exist in science, but this doesn't imply any form of contradiction. The theory in question is the "S-matrix", popular in the 1960s in high-energy physics. The aspect of interest here is that this theory posits a gunky world, that is, a world in which each object has a proper part. In fact, the S-matrix theory accepts the existence of hadrons and also claims that

each hadron in turn contains hadrons of each type, including additional specimens of its own type. The example offered by the S-matrix theory is illustrative, therefore, of the fact that science presents infinite regressions in which each successive step of the regression is characterized by the same form as the previous step, thus making the regression itself *homogeneous* in form. The case study examined here, McKenzie argues, is only a special case of a phenomenon that occurs within scientific theories: infinite regressions are always vicious. The reason for this derives from the fact that the form scientific explanations take is inevitably constrained by the basic postulates of the relevant theory, containing a certain number of predicates. In the case of infinite explanatory regression, therefore, the general framework and its stock of predicates remain the same even though the structure of determination never ends. Therefore, McKenzie argues, the resulting regressions are flawed in some substantive sense. Ultimately, McKenzie asserts that her analysis points in a very specific direction: foundationalism is false and should be consequently abandoned.

In the last instance, I would like to focus on McKenzie's analysis on foundationalism. Certainly, there are those, such as Schaffer (2010),[5] who have argued that *every grounding chain terminates*. However, this characterization of foundationalism, which McKenzie assumes, doesn't consider the theoretical developments that have taken place in recent years to make foundationalism more precise. There are those who, like Dixon (2016)[6] or Rabin and Rabern (2016),[7] have proposed to characterize foundationalism in terms of maximal grounding chains by requiring that "every maximal grounding chain terminates" (Pearson 2022: 1544),[8] whereby maximality of a grounding chain requires that there is no entity that is not a member of the chain and that partially grounds every member of the chain. But there are also those, such as Pearson 2022, who have proposed to capture the idea of foundationalism by appealing to the notion of inclusive grounding chain: "an inclusive grounding chain is a chain of grounding such that it is not the case that each member of the chain is grounded by a fact or facts that are not members of the chain" (Pearson 2022: 1542). Pearson redefines foundationalism so that "every grounded entity is a member of at least one inclusive full grounding chain and that every inclusive full grounding chain terminates" (Pearson 2022, 1546). It wouldn't be surprising if some of the objections in the naturalistic vein proposed by McKenzie could be resolved by adjusting the adopted definition of foundationalism, which has not been thoroughly investigated and remains formulated only in its most basic definition. If you aim to demonstrate that foundationalism is to be discarded, you must first show that every effort has been made to salvage it, and yet, despite these efforts, the sciences are indicating a wholly different direction. Consequently, the last word has not yet been said about grounding foundationalism, which I believe still enjoys a good reputation amongst philosophers.

University of Padua JACOPO ROSINO GIRALDO

[5] Schaffer, J. 2010, "Monism the Priority of the Whole", *Philosophical Review*, 119, 31-76.
[6] Dixon, S. 2016, "What is the Well-Foundedness of Grounding?", *Mind*, 125, 439-468.
[7] Rabin, G., & Rabern, B. 2016, "Well Founding Grounding Grounding", *Journal of Philosophical Logic*, 45, 349-375.
[8] Pearson, O. 2022, "Grounding, Well-Foundedness, and Terminating Chains", *Philosophia*, 51 (3), 1539-1554.

Advisory Board