UNIVERSITY OF BIRMINGHAM

Putting Explanation First:

Progress in Science and Philosophy

By

Nicholas James Charles Emmerson

A thesis submitted to the University of Birmingham for the degree of

DOCTOR OF PHILOSOPHY

Department of Philosophy | School of Philosophy, Theology, and Religion |

College of Arts and Law | University of Birmingham | July 2023

Contents

Preface

While the claim that science makes progress is uncontroversial, the question of "*how?*" such progress is made has long been contested. Until the turn of the 21st century, this debate had largely been dictated by meta-scientific considerations. On the one hand, realists like Popper (1959) and Niiniluoto (1980) argued that progress is made when we increase our stock of *verisimilar*, or truthlike, beliefs. On the other, antirealists like Kuhn (1962) and Laudan (1977) argued that progress is made by solving "puzzles" or "problems", success in which can be judged only from within a "paradigm" or "research tradition".

However, in 2007 new life was breathed into this debate with the emergence of a second realist account. According to Bird (2007), science progresses not through the increase of truthlike beliefs, but rather through the accumulation of knowledge; true belief, *justified* by reliable scientific methodology. Since 2007 the literature on scientific progress has seen rapid expansion and a third realist account has now been proposed.[1] On Dellsén's (2016) view, science progresses when *understanding* increases, that is when scientists grasp how to explain, or predict, more aspects of the world than they could before.

In stark contrast, debate concerning the nature and scope of philosophical progress has focused upon the question of "*whether?*" philosophers can claim to have made any significant progress. What's more, responses to this question have largely been pessimistic. This pessimism is typically motivated either by the apparent lack of consensus on paradigm topics, or the observation that philosophical theories are rarely superseded.[2] Nonetheless, recent literature suggests that the "*how?"* question concerning scientific progress, and the "*whether?*" questions concerning philosophical progress are more closely connected than one might expect.

---

[1] See e.g., Niiniluoto (2014, 2017); Bird (2007, 2008, 2019, 2022); Rowbottom (2008, 2015); Mizrahi (2012, 2013, 2017, 2021, 2022); Cevolani & Tambolo (2013); Dellsén (2016, 2017, 2018, 2021, 2022); Park (2017; 2020); Sterpetti (2018); Shan (2019, 2022a); Dellsén, Lawler & Norton (2022, *forthcoming*); Emmerson (2022a); Lawler (2022).

[2] For the former see e.g., Horwich (2012); Bourget & Chalmers (2014); Rescher (2014); Chalmers (2015); Shand (2017); and for the latter see e.g, Sterba (2004); Dietrich (2011); Jones (2017); Slezak (2018). As if to prove the point, however, there is not universal consensus on this topic. Some take a more optimistic view, typically arguing either that there is more consensus in philosophy that pessimists admit, or that such collective agreement simply isn't necessary for progress. See e.g., Rapaport (1982); Stoljar (2017); and Frances (2017), in the first instance, and Brock (2017); Cappelen (2017); Bergson *et al* (2019); and McKenzie (2020) in the second.

Dellsén, Lawler & Norton, for example, argue that although we currently lack a general definition 'on which to stand as we evaluate whether (and the extent to which) philosophy has made progress', mature debate concerning the nature of scientific progress is the natural place to search for such a vantage point (2022:2). McKenzie (2020) advocates for a similar methodology but goes further still, arguing that science has a monopoly on progress and, consequently, that philosophy could only make progress if it does so in the same *way* as science.[3] Despite growing interest in this topic, attempts to *unify* analyses of scientific and philosophical progress remain remarkably thin on the ground.

The papers contained in this thesis aim, in part, to rectify this situation; to show how we might define a notion of progress which can be applied across science and philosophy. For the sake of candidness, however, I must make two important admissions regarding the scope of this project. First, the analysis of progress put forward here is by no means entirely novel. Rather, it builds upon the recent *epistemic* turn within the scientific progress literature, which has seen an increased focus upon the distinction between knowledge and understanding. Where my analysis diverges from the recent literature is in the adoption of a novel "explanation-first" methodology.

The idea that science aims to provide understanding is remarkably widespread, as is the idea that understanding is a form of knowledge; knowledge which results from grasping *explanations*.[4] It follows that a thorough analysis of the nature of explanation is a necessary step in presenting a viable account of scientific progress. However, the literatures on scientific explanation and scientific progress have remained almost entirely disconnected. The core motivation for an explanation-first approach lies in recognizing that different theories of explanation provide for markedly different interpretations of progress. On the *interventionist* analysis which I shall defend, scientific progress occurs when scientists providing increasingly invariant explanations of target phenomenon.

The second admission which I must make, concerns the scope of philosophical progress under discussion. Historical interest in the apparent contiguousness of science and philosophy has largely been motivated by *metaphysics.* As is well documented, analytic philosophy and metaphysics have had a tempestuous relationship over the last century.[5] While the Vienna Circle followed Wittgenstein (1921) in rejecting metaphysical claims as

[3] Although McKenzie (2020) targets *metaphysics* specifically, her arguments generalize.
[4] See e.g., Hempel (1965); Friedman (1974); Salmon (1984); Kitcher (1981); Achinstein (1989); Kitcher (2002); Lipton (2003); Bird (2007); Stevens (2008); Grimm (2014); Khalifa (2017).
[5] See Gillies (1993) for an excellent analysis of this history.

meaningless, Popper's (1959) critique of the principle of verifiability helped to repair metaphysics' reputation.[6] Despite this, the dawn of the 21st century brought with it a new wave of anti-metaphysics sentiment.[7]

As a result, metaphysics is the natural place to start in constructing a unified analysis of progress across science and philosophy. If an analysis of scientific progress cannot also be applied to metaphysics, the philosophical discipline which is most closely associated with science, then there seems little prospect of extending such an account to other, less closely associated, areas. By expanding my explanation-first methodology and showing that progress in metaphysics can similarly be understood in terms of the production of increasingly invariant explanations, my hope is that we can achieve a more positive outlook with respect to the prospects of progress in other areas of philosophy. With these admissions out of the way, we can now turn to the structure of this thesis.

The first three chapters which follow have all been published elsewhere, in roughly the same format seen here. In chapter 1, I defend Bird's (2007) knowledge-first analysis of scientific progress, against the claim that science progresses, instead, through increasing understanding.[8] Crucially, I argue that scientific understanding cannot be distinguished from knowledge on the basis that only the latter requires justification, as Dellsén (2016, 2017, 2018) has argued. The justification provided by a reliable scientific methodology is, I maintain, indispensable for gaining such understanding.

One initial hurdle, in applying a specifically *interventionist* explanation-first analysis of progress to philosophy, is the prevalence of a position that I have labelled "intervention puritanism" (*I*-puritanism, for short), which suggests that interventionism itself is an essentially *causal* theory of explanation.[9] Given that explanations provided by philosophy are clearly *non*causal in nature, it follows from *I*-puritanism that a unifying, interventionist account of explanation across science and philosophy is impossible, and thus, so too is a unifying, interventionist, explanation-first account of progress. However, in chapter 2, I argue

[6] Popper maintains (*contra* Quine [1951]) that although metaphysics and science *can* be demarcated, the former can have both a positive and negative influence on the latter.
[7] See e.g., Ladyman & Ross (2007); Callender (2011); French & McKenzie (2015); Bryant (2020); Mckenzie (2020).
[8] This chapter was published in *Synthese* prior to the submission of this thesis under the title "Understanding and scientific progress: Lessons from epistemology" See Emmerson (2022a).
[9] See e.g., Woodward (2003, 2015, 2018); Bokulich (2011); Leuridan (2012); Saatsi and Pexton (2013); Harinen (2014); Pexton (2014); Baumgartner & Gebharter (2015); Rice (2015); Romero (2015); Reutlinger (2016, 2017, 2018); Baumgartner & Casini (2017); French and Saatsi (2018); Khalifa *et al* (2018, 2020); Saatsi (2018); Jansson & Saatsi (2019); Lange (2019, 2021).

that *I*-puritanism is false. In its place, I motivate "intervention liberalism" (*I*-liberalism, for short).[10] According to *I*-liberalism, interventions do not carve nature at its causal joints; noncausal explanations *can* be accurately characterized using interventionist counterfactuals and structural equations models. By showing *I*-puritanism to be false, this chapter paves the way for an interventionist explanation-first account of scientific progress to be extended into philosophy.

Building on this idea, in chapter 3, I not only argue that interventionism can be used to characterize explanations within metaphysics, but that the interventionist notion of *explanatory depth* plays a particularly interesting role in such contexts.[11] Explanatory depth, on the interventionist model, is measured with respect to the range of interventions under which an explanatory generalization remains invariant. By applying this notion to explanations in metaphysics we are able to show, among other things, that interventionism itself provides deeper explanations of the connection between explanans and explanandum than rival accounts of explanation.

In chapter 4, I turn my attention back to the topic of understanding and its connection to explanation. While there has been significant interest in this topic in recent years, those involved have typically adopted a methodological stance which *privileges* the former over the latter.[12] These "understanding-first" approaches attempt to provide analyses of understanding which are silent upon the nature of explanation. I argue, however, that the explanation-first methodology, introduced but later abandoned by Kim (1994), represents a viable alternative to the *status quo*. The prevalence of understanding-first approaches has meant that *theories* of explanation are rarely evaluated with respect to how successfully they account for the connection between explanation and understanding. By adopting an explanation-first approach, I argue that interventionism emerges as our best hope for accurately characterizing understanding in terms of explanatory knowledge.

Chapter 5 brings together much of the work of the previous four chapters and represents the culmination of the central argument of this thesis. I begin by motivating the explanation-first approach to scientific progress already mentioned. I argue that differing theories of explanation generate remarkably divergent accounts of scientific progress, and

---

[10] This chapter was published in *Erkenntnis* prior to the submission of this thesis under the title "A Defence of Manipulationist Noncausal Explanation: The Case for Intervention Liberalism". See Emmerson (2021).
[11] This chapter was published in *Philosophical Studies* prior to the submission of this thesis under the title "Plumbing Metaphysical Explanatory Depth". See Emmerson (2022b).
[12] See e.g., Khalifa (2013, 2017); Dellsén (2016, 2017, 2018, 2021, *forthcoming*); de Regt (2015, 2017, 2020, 2022),

that an ill-informed choice in this respect can render epistemically-driven accounts of progress untenable. By adopting an *interventionist* explanation-first approach, however, I argue that we can provide a viable analysis of scientific progress.

On this analysis, progress occurs when scientists provide increasingly invariant explanations of target phenomena, and *correspondence* is achieved where the range of invariance displayed by a superseded theory is *strictly contained* with the range of invariance displayed by a progressive theory. I then utilize this novel analysis of progress in science, to characterize an analogous notion of progress in metaphysics. I apply this interventionist explanation-first account to a case study concerning the identity and distinctness of concrete objects and argue that progress in this example can similarly be characterized in terms of metaphysicians providing increasingly invariant explanations of target phenomena.

The final chapter of this thesis represents an attempt to *begin* the process of expanding the unifying, interventionist, explanation-first analysis of progress, developed in chapter 5, into other areas of philosophy. I argue that debate between moral generalists and moral particularists is founded on a false equivocation and that morality can be principled (*contra* particularism), even if there are no exceptionless generalizations governing moral inquiry (*contra* generalism). By adopting an interventionist methodology with respect to normative explanation, I argue that moral principles are merely those explanatory generalizations which are *most invariant.* I call the resulting position "moral invariantism".

I believe that this interventionist account of noncausal explanation in ethics, and the accompanying invariantist analysis of moral principles, suggests that my interventionist explanation-first approach to progress has applications beyond the confines of science and metaphysics. This novel analysis of normative explanation provides us with all the tools necessary in order to define an analogous notion of *ethical* progress. Progress, in ethics, can thus similarly be understood in terms of ethicists providing increasingly invariant explanations of target phenomenon. However, there is not the space to mount a sustained defence of this application in this thesis. While I believe that this final chapter suffices for a proof-of-concept at least, any further pursuit of this idea will have to wait for future research.

Acknowledgements

First and foremost, my thanks go to Prof Alastair Wilson, who gave me the opportunity to join the FraMEPhys team at a time when I had all but given up hope of being able to complete my PhD. As a supervisor, Al has been incredibly generous with his time and his ever-insightful advice has been invaluable. He has never been too busy to answer an inane question or provide reassurance in a period of perceived crisis. Most importantly, however, Al has always taken me, and my ideas, seriously. This is no small thing, and for the confidence which I have gained over the last three years I owe Al a huge debt of gratitude.

My time spent as a part of FraMEPhys has been a very happy one and I want to thank everyone who has been involved. However, a special thanks is due to Dr Joaquim Giannotti, Dr Michael Hicks, Dr Noelia Iranzo Ribera and Dr Katie Robertson, for the warmth with which I was welcomed and for collectively fostering such a friendly and supportive collegiate environment. For making the project possible, I am thankful to the European Research Council for the funding received under grant agreement no. 757295.

Work on this thesis began in 2018 at King's College London, where I was an MPhil/PhD student under the supervision of Prof Alexander Bird. The two years I spent at KCL were tumultuous, with the death of my partner's father, Prof John Richardson, being followed by the COVID-19 pandemic. During this difficult period, and in the years since, Alexander has helped me in innumerable ways, going above and beyond the call of duty. I could never have made it this far without his generous support and for this I am immensely grateful.

There are countless others who have had a hand in shaping this thesis, including the graduate community at the University of Birmingham, as well as conference and workshop audiences at the University of Bristol; University of Exeter; University of Hamburg; University of Kent; University of Konstanz; University of Lisbon; and IHPST Paris 1 Panthéon-Sorbonne. For providing detailed comments on draft chapters, however, Samuel Andrews, Prof Kerry McKenzie and Dr James Norton deserve a special thanks.

I would likely not have embarked upon a PhD at all, had it not been for the support of Prof Lubomira Radoilska, with whom I worked closely during my undergraduate and master's study at the University of Kent. My thanks go to Lubomira, not only for having been

a dedicated and inspiring supervisor, but also for having encouraged me to apply for the MPhil programme at the University of Cambridge; something I would otherwise never have had the nerve to do.

Outside of the academy, I must thank my partner of 11 years, Barbara Richardson. Having met as undergraduate students in 2011, Barbara has been an incredible source of support ever since. Her patience and positivity have kept me going when seemingly insurmountable challenges would otherwise have seen me throw in the towel. I can only hope that she is as proud of me as I am of her. I also wish to thank my mother Elizabeth Stones, my father Leslie Emmerson and my brother Thomas Emmerson without whom, for more reasons than I could possibly hope to mention, this thesis would have been impossible. Having sparked and nurtured my intellectual curiosity, their faith in me and my decision to continue studying has never wavered. My successes, such as they are, are owed unequivocally to them.

Chapter 1

Understanding, Justification and Scientific Progress

Until recently, debate surrounding the nature of scientific progress had focused upon the precise role played by justification, with two realist accounts having dominated proceedings. However, a third realist account has now been put forward, one which offers no role for justification at all. According to Finnur Dellsén's (2016) noetic account, science progresses when understanding increases, that is, when scientists grasp how to correctly explain or predict more aspects of the world that they could before. In this paper, I argue that the noetic account is severely undermotivated. Dellsén provides three examples intended to show that understanding can increase absent the justification required for true belief to constitute knowledge. However, I demonstrate that a lack of clarity in each case allows for two contrasting interpretations, neither of which serves its intended purpose. On the first, the agent involved lacks both knowledge and understanding; and, on the second, the agent involved successfully gains both knowledge and understanding. While neither interpretation supports Dellsén's claim that understanding can be prised apart from knowledge, I argue that, in general, agents in such cases ought to be attributed neither knowledge nor understanding. Given that the separability of knowledge and understanding is a necessary component of the noetic account, I conclude that there is little support for the idea that science progresses through increasing understanding rather than the accumulation of knowledge.

I. Introduction

In his 2006 paper, "Is Understanding a Species of Knowledge?", Grimm notes a striking division between philosophers of science and epistemologists regarding their respective characterisations of *understanding*. On the one hand, there had been a relative consensus among philosophers of science that understanding is merely a species of knowledge, knowledge which stands in some privileged relation to explanation.[13] On the other, according

[13] See e.g., Achinstein (1983); Salmon (1989); Kitcher (2002); Woodward (2003); Lipton (2003); Bird (2007).

to Grimm: 'virtually every major epistemologist… has come to the conclusion that understanding is *not* a species of knowledge' (2006: 516).[14]

Since 2006, however, the landscape of debate surrounding understanding has changed dramatically, with several leading philosophers of science having broken ranks.[15] One area where this shift has had significant impact is in contemporary debate upon the nature and scope of *scientific progress* (SP). In recent years, discussion surrounding SP has focused upon the precise role played by *justification*, with two contrary realist accounts dominating proceedings.[16]

According to the *semantic* account, most closely associated with Popper (1962, 1979) and Niiniluoto (1984, 1987, 1999, 2014), progress is made when scientific theories become closer to the truth, that is, when they become more verisimilar or approximately true.[17] According to Bird's (2007, 2008, 2019, 2022) *epistemic* account, on the other hand, science progresses through the accumulation of knowledge; true belief, justified by reliable scientific methodology.[18] Where these perspectives diverge is in the extent to which justification is considered *constitutive* of progress.[19] Both sides appear to agree, however, that 'beliefs without any justification simply do not belong to the scope of scientific progress' (Niiniluoto, 2014:76).

---

[14] Here, Grimm cites: Elgin (1996, 2004); Zagzebski (2001); and Kvanvig (2003). We could now expand this list to include (among others): Elgin (2007); Pritchard (2008); Kvanvig (2009); Gardiner (2012); and Mizrahi (2012).
[15] For examples not covered in detail here, see de Regt (2015 & 2017) and Wilkenfeld (2017, 2019).
[16] While antirealist notions of SP are available, most notably the *functionalist-internalist* accounts of Kuhn (1962, 1991) and Laudan (1977, 1981, 1984), interest in them has waned in recent years; Shan (2019) is a notable exception.
[17] The connection between truth and progress is most famously highlighted in Putnam's (1975) 'no miracles' argument, a more contemporary interpretation of which can be found in Lipton (2003). However, not all realists agree that arguments of this form are convincing; see Bird (2022) for example.
[18] Also see Mizrahi (2013, 2017); and Mizrahi & Buckwalter (2017), Park (2017), and Sterpetti (2018).
[19] Darrell Rowbottom (2008), for example, argues that justification is merely *instrumental* to progress; and Gustavo Cevolani & Luca Tambolo (2013) argue that the semantic conception can handle issues of justification by means of a distinction between "real" and "estimated" progress without needing to concede that SP is the accumulation of scientific knowledge.

It ought to come as some surprise then, that in several recent papers, Dellsén (2016, 2017, 2018) has put forward a third realist account of SP, which does away with the notion of justification all together. According to this *noetic* account, science progresses when understanding increases, that is, when scientists grasp how to correctly explain or predict more aspects of the world that they could before.

In arguing that science progresses through understanding *rather than* knowledge, Dellsén is committed to a rejection of the orthodox view among philosophers of science, that 'all (genuine as opposed to apparent) understanding is also knowledge' (Bird, 2007:84). Rather, Dellsén claims that understanding is distinct from knowledge in not requiring justification to be attained. It thus follows from Dellsén's position that beliefs without any justification *do* belong in the scope of scientific progress.

The principal aim of this paper is to show that this controversial position of Dellsén's is undermotivated, if not outright untenable. Dellsén (2016, 2017) provides three apparent examples of understanding without justification. The first two are hypothetical cases concerning Alice, a struggling student, and Bernie, a retired mechanic. The third example comes from the history of science itself: Albert Einstein's (1905/1956) explanation of Brownian motion.

What links these scenarios is that the agents involved supposedly come to understand a given phenomenon despite being in possession of evidence which 'undermines the justification for their respective beliefs' (Dellsén, 2017:9).[20] It is my belief, however, that

[20] In "Understanding without justification or belief" Dellsén notes that it is common to distinguish between different kinds of understanding 'some of which are not relevant to epistemology and thus clearly would not replace knowledge as the primary focus of the field' (2017:2, fn.1). *Objectual understanding* is understanding that is grammatically followed by an object 'as in 'she understands economic depression' (Dellsén, 2017:2, fn.1). It is this form of understanding which is of central interest to Dellsén and as a result, to *me*. However, Dellsén notes that 'what I have to say will apply *mutatis mutanda* to understanding-why as well' (2017:2, fn.1). For a detailed account of the distinction between understanding-why and objectual understanding see e.g., Khalifa (2013).

these cases are in fact analogues of the Comanche-style cases first introduced by Kvanvig (2003). While this might not seem immediately problematic, I shall argue that concerns ought to arise once we consider that Comanche-style cases are rendered wholly unconvincing by an analysis put forward by Grimm (2006).

In the next section I provide a more detailed account of Dellsén's position, and how his three examples are supposed to support the thesis that understanding is separable from knowledge and that SP tracks the former rather than the latter. In section III I examine Kvanvig's (2003) Comanche case as well as Grimm's (2006) criticism of it. Such cases suffer from a lack of clarity and, according to Grimm, once they are properly fleshed out, two contrary interpretations of these scenarios emerge.

The problem for Dellsén is that neither interpretation suggests that understanding and knowledge are separable: 'on any way of filling out the details knowledge and understanding seem to sway together' (Grimm, 2006:520). In sections IV and V, I then show how this same argument can be levelled against Bernie's case and Alice's case, respectively. In the final two sections of this paper, I turn to Dellsén's third example.

In section VI, I argue that Einstein's explanation of Brownian motion, like that of Alice and Bernie, is an analogue of Kvanvig's Comanche case and thus fails to draw any distinction between understanding and knowledge. However, I also show that a comparison between Einstein's case and recent research on SARS-COV-2, by José Lourenço *et al* (*ms*), provides novel insight into Comanche-style cases more generally. In section VII I highlight two fundamental problems with the noetic account of SP, raised by this comparison.

These problems suggest that it is Grimm's first interpretation which ought to be adopted with respect to Comanche-style cases in general, and that neither understanding nor knowledge can be attributed to agents in such cases. Given that the separability of knowledge

and understanding is a necessary component of Dellsén's view, I conclude that there is little support for the idea that science progress through increasing understanding *rather than* the accumulation of knowledge.

II. Understanding Without Knowing

It is interesting to note that, unlike knowledge, understanding is not an intrinsically realist notion. In recent work, for example, de Regt (2015, 2017) has argued for an *anti*realist notion of understanding which is not even moderately factive.[21] In contrast, Dellsén is keen to maintain his realist convictions and thus takes understanding to be *quasi*-factive, suggesting that 'the explanatorily/predictively essential elements of a theory must be true in order for the theory to provide grounds for understanding' (2016:73, fn6).[22]

Dellsén argues that, being based upon *this* notion of understanding, his noetic account of SP can easily accommodate the realist mantra that progress is made by getting closer to the truth, since 'one's degree of understanding would simply be determined, at least in part, by how close to a fully correct representation of something is used for explanatory and predictive purposes' (2018:10).[23] Consequently, Dellsén argues that it is *justification*, rather

[21] According to De Regt, understanding is a matter of *intelligibility*; being able to use and manipulate a model 'in order to make inferences about the system, to predict and control its behaviour' (2015: 3791). Since this notion of understanding does not track truth, an account of SP based upon it would provide a natural alternative to traditional, antirealist, functionalist-internalist accounts of SP, although de Regt does not pursue this line of enquiry (see fn.1 for more detail). More recently still, Wilkenfeld (2017, 2019) has advocated a similar position.
[22] While some epistemologists have argued for the stronger position, that understanding requires *full*-factivity (e.g. Grimm 2006; Pritchard 2010; and Hills 2009), most follow Dellsén in taking the notion to be quasi-factive (e.g. Kvanvig 2003, 2009; Mizrahi 2012; and Wilkenfeld 2017).
[23] In this sense, Dellsén admits that his noetic account of SP is remarkably similar to the semantic account (2018:10). The principal difference being 'whether proposing new explanations or making new predictions could itself constitute progress, even when there is no change in the theories with which one would explain and predict' (2018:10). While Rowbottom (2015) has argued that the semantic view cannot accommodate such progress, Dellsén (2016, 2018) believes that his own account can.

than truth, which stands to distinguish understanding from knowledge.[24] In defending this idea, Dellsén (2017) introduces us to Alice, Bernie and Einstein.

Alice is a struggling student who has failed every assignment she has attempted this year. Despite this, Alice turns out to have an innate knack for geometry and successfully derives the Pythagorean theorem from a version of the original proof (without any help from a teacher or textbook). In this case, Dellsén argues, Alice does not have justification for believing that her proof is accurate, because her previous failures provide 'good reasons to believe that *this* attempt at understanding a new subject matter in school is a failure as well' (2017:6). It is nonetheless clear, according to Dellsén, that Alice *understands* the Pythagorean theorem.

Bernie, on the other hand, is an automobile mechanic who reads in a newspaper that a convicted confidence trickster is coming to town. The newspaper article includes a picture of the man and a warning that not a single word he says can be trusted. Unfortunately for Bernie, the following day the man in question rings his doorbell and reports that his car has broken down. From the description of the car's behaviour prior to the breakdown, provided by the trickster, Bernie *understands* that the issue is a broken timing belt. Regardless, Bernie is not justified in believing the issue to be a broken timing belt, since 'he should know better than to trust a convicted con man' (Dellsén, 2017:8).

What is apparently peculiar about these cases, is that Alice and Bernie possess evidence, which is not explanatorily relevant to the object of understanding, but nevertheless undermines any justification they might have for their respective relevant beliefs. Dellsén

[24] Severing the truth or justification connections are not the only options available in distancing understanding from standard accounts of knowledge. There is, of course, a third notion involved: belief. Dellsén (2017) himself puts forward the idea that a notion of understanding based upon *acceptance* rather than belief would also suffice to distinguish knowledge from understanding. However, since the distinction between belief and acceptance plays little role in Dellsén's (2016) account of scientific progress, I shall not discuss it any further here.

does not limit his examples to hypotheticals, however. Indeed, being pulled from the annals of the history of science itself, his third example is perhaps his most convincing.

This case concerns Einstein's (1905/1956) proposed explanation of Brownian motion in terms of the kinetic theory of heat, which was highly speculative at the time. As Einstein himself admits: '[i]t is possible that the movements to be discussed here are identical with the so-called "Brownian molecular motion"; however, the information available to me regarding the latter is so lacking in precision, that I can form no judgement in the matter' (1905/1956:1). As such, Dellsén (2016:76) argues that Einstein, like Alice and Bernie, was clearly also lacking the justification required to *know* that the movements at issue were, in fact, real.

While Dellsén's position is (to the best of my knowledge) unique among philosophers of science, the idea that understanding can be gained absent justification *does* find support within the epistemology literature. In the next section, I examine a case introduced by Kvanvig (2003) which bears a striking resemblance to the three cases appealed to by Dellsén.[25] I also highlight convincing criticism of Kvanvig's account provided by Grimm (2006). The problem for the noetic account of SP is that the examples used by Dellsén appear to be susceptible to this very same line of argument.

III. Kvanvig's Comanche Case

In *The Value of Knowledge and the Pursuit of Understanding* Kvanvig (2003) also argues for a realist notion of understanding. They suggest that, while the truth of one's belief is an important factor for genuine understanding, the way in which one comes to those beliefs,

[25] It is worth noting that while Dellsén (2016, 2017) cites Kvanvig (2003, 2009) in reference to arguments which distinguish knowledge from understanding, he does not make any reference to structural similarities between his own examples and Kvanvig's Comanche case; nor indeed Grimm's (2006) subsequent criticism of it.

their "*etiology*", is irrelevant. To support this claim, Kvanvig provides an example which is analogous to the cases described by Dellsén (2017).

Suppose, Kvanvig (2003:197-198) argues, that we come to an understanding of *why* the Comanche dominated the Southern Plains of North America, through a textbook which we pick up from the local library. Now suppose that, while the textbook through which we gained this understanding is accurate, almost every other book on the topic is full of factual errors. Had we picked one of *these* books off the library shelf, instead of the accurate one, our beliefs about the Comanche would have been almost entirely false.

From this scenario, Kvanvig draws the same conclusion as Dellsén with respect to Alice and Bernie; while we do not have *knowledge* of the Comanche's dominance of the Southern Plains, we do have the relevant *understanding*. As he explains:

> 'It is the internal seeing or appreciating of explanatory relationships in a body of information which is crucial to understanding. When we think about knowledge, however, our focus turns elsewhere immediately, if we have learned our lessons from the Gettier literature: we think about the possibility of fortuitousness, or accidentality, of being right only by chance' (Kvanvig, 2003:198).

Yet, as Grimm (2006) highlights, there is a lack of clarity regarding exactly what is happening in Kvanvig's Comanche case. There appear to be two possible interpretations of the role of fortuitousness here. The first is that we form true beliefs based upon an unreliable *source* of information which just happens to get it right; 'a crystal ball, a pathological liar, etc' (2006:523). The second is that we form true beliefs on the basis of a reliable source of information in an *environment* in which the majority of possible sources are unreliable; 'suppose by luck you happen across the only honest man in a room full of pathological liars' (2006:523).

If we are to follow Kvanvig, what links these interpretations is that wherever truth is accompanied by chance, understanding is possible, while knowledge is ruled out. According to Grimm, however, once these interpretations are fleshed out, it becomes apparent that knowledge and understanding are not so easily prised apart. On neither formulation, Grimm argues, does Kvanvig's Comanche case provide an example of understanding absent knowledge.

In the first instance, taking the Comanche case to be the result of an unreliable source, Grimm (2006) argues that although knowledge doesn't seem to be within reach, neither does understanding. Let us suppose that our textbook merely lays out the facts and leaves the insightful reader to construct explanations on their own: 'the relevant accomplishment – the piecing together, as it were – would then be entirely internal [as Kvanvig claims]' (Grimm, 2006:526). Now imagine that the research undertaken by the author is 'extremely shoddy… [conducted] on the basis of only one sample, and without controlling for the influence of other factors' (Grimm, 2006:526).

Grimm concludes that in such cases, even if it turns out that the information contained in the textbook *just happens* to be entirely correct, 'it is hardly the case that the textbook reader who pieces together and develops an account of Comanche dominance would really come to understand *why* the Comanche dominated the Southern Plains' (2006:262).[26] While we might concede that there is a *kind* of internal accomplishment in this case, it is not, according to Grimm, one of understanding: '[f]or that accomplishment, apparently, a stronger (alternatively: less accidental) connection to the Comanche is required' (2006:526).

[26] Indeed, as Grimm highlights in a footnote, if understanding could be acquired upon the basis of such slipshod way, it becomes difficult to see why disciplines such as sociology, economics and history would need to develop 'their methodologies and established canons for constructing and evaluating explanation' (Grimm, 2006:526, fn.14)

In the second instance, taking Comanche cases to be the result of a reliable source located in an environment of unreliable sources, Grimm (2006) argues that while understanding seems to be within reach, so does knowledge. In this scenario we can assume that the author of our textbook is an expert on the Comanche, with impeccable research acumen, but that every other textbook on the topic which hits upon the truth does so only by accident. The key point to appreciate with this interpretation, according to Grimm, is that 'Kvanvig overstates his case in claiming, without appropriate qualification, that knowledge is incompatible with luck' (2006:527).

A case highlighted by John Hawthorne (2003) draws out the compatibility of knowledge with certain forms of luck here. Imagine that six children are each randomly assigned an atlas, all but one of which contains misinformation regarding the capital of Austria. Despite the fact that these books were assigned at random, Hawthorne (2003) argues, intuition would have it that the child whose book reads 'Vienna' *knows* which city is the capital of Austria. As Grimm explains:

> '[w]hen (as third party evaluators) we have reason to believe that the source of the information is good – here, that the textbook came from a reliable source etc – we tend to focus on the sense in which the belief is not lucky: it was no matter of luck, we think, that the textbook author identified Vienna as the capital of Austria, even if it was a matter of luck that this particular textbook ended up in the student's hands' (2006:528).

Consequently, Grimm concludes that, in a scenario in which we are tempted to think that someone *understands* various things about the Comanche on the basis of a reliable source in an environment of unreliable sources, there is a strong tendency to suggest that the person also *knows* these things about the Comanche (2006:529).

As I highlighted at the end of the previous section, my central interest in Kvanvig's Comanche case, and Grimm's criticism of it, is the obvious structural similarity to the cases provided by Dellsén (2017). It is my contention that all three of Dellsén's cases fall foul of the exact same problem facing the Comanche scenario, being similarly under-described. Once fleshed out, it become apparent that none of these cases serve to sever the connection between understanding and knowledge as Dellsén suggests.

Just as we have seen with respect to the Comanche case, Alice's, Bernie's and Einstein's cases can all be interpreted as instances where either knowledge *and* understanding are gained, or as instances where neither are. To see why, I shall begin by revisiting Bernie's case in the next section, before returning to Alice's case in section V, and Einstein's case in sections VI and VII.

IV. Bernie's Case Revisited

As the reader will no doubt recall, Bernie (a retired mechanic) is in the unenviable position of being confronted with a known confidence trickster who describes the behaviour of his car, allowing Bernie to come to *understand* that the car in question has a broken timing belt. However, since 'he should know better than to trust a convicted con man', it appears that Bernie cannot *know* that the issue is a broken timing belt (Dellsén, 2017:8).

At a first glance, Bernie's case seems to be an obvious analogue of Grimm's first interpretation of Kvanvig's Comanche case. As Dellsén himself highlights, Bernie's case 'involves testimony from source that is known to be untrustworthy' (2017:8). Just as the author of the Comanche textbook *just happens* to provide accurate information despite unreliable methodology, Bernie's con man *just happens* to provide accurate information

regarding the behaviour of his car, despite the general tendency of con men to do quite the opposite.

In response to the Comanche case resulting from an unreliable source, Grimm concludes that neither understanding nor knowledge can be attributed to the reader of the textbook. While there may well be a *kind* of internal accomplishment is such cases, the accidental nature of the connection between the textbook and the Comanche suggests that this accomplishment is not understanding. Given the similarities between this case and Bernie's, one might naturally suppose that the same conclusion can be drawn here.

In response to the fortuitous accuracy of the testimony provided by the con man, one might think that we can conclude that neither understanding nor knowledge should be attributed to Bernie either. In later sections, I shall argue that a thorough analysis of Einstein's case suggests that this interpretation of Comanche-style cases is the correct one. However, for the time being, I wish to address an obvious response to this line of argument; that Bernie's understanding in the case described by Dellsén appears to be far less accidental than the reader's understanding in the analogous case described by Grimm.

While it is true that, in general, the testimony of a con man ought not be trusted, it is no mere accident that Bernie is able to understand the issue to be a broken timing belt. Bernie is, after all, a retired mechanic: an expert in his field. As such, one might think that Grimm's argument doesn't hold for Bernie. A closer inspection of the role of this expertise, however, reveals that this line of argument cannot save Dellsén's claim that Bernie's case serves to differentiate understanding from knowledge.

The importance of Bernie's expertise to Dellsén's argument can be seen if we consider an alternative scenario in which the con man describes the behaviour of his car to *me*, rather than to Bernie. Given that I am not even aware of where one might find a timing belt, let

alone its function, it is patently obvious that any talk of understanding (or knowledge) is off the table in this case.

The understanding which Dellsén ascribes to Bernie, then, can only be the result of this expertise. As such, one might argue that Bernie *is* in possession of a certain type of understanding. It is presumably true of all retired mechanics (at least the competent ones), that they have an impeccable grasp of the *general* explanatory relations concerning the causes and symptoms of broken timing belts.[27]

The problem with this perspective is that, in shifting focus away from the fortuitous nature of the con man's accurate testimony and towards the sense in which Bernie's understanding is no mere luck, we run into Grimm's second interpretation of Comanche-style cases. Here, of course, Grimm argues the expertise and reliable methodology of the author outweigh the luck in our having picked up the only accurate textbook.

As such, this interpretation pulls our intuition towards the ascription of *both* understanding and knowledge of the Comanche's dominance of the Southern Plains. Similarly, any understanding which we attribute to Bernie based upon his expertise as a retired mechanic comes preloaded with *justification,* built up over a lifetime of first-hand experience dealing with similar issues in other vehicles.

On the face of it, it appears that Bernie's case is a simple analogue of Grimm's first interpretation of Kvanvig's Comanche case, both of which result from an unreliable source. Given the fortuitous connection between the source of information and the broken timing-belt, one might conclude that neither knowledge nor understanding are present in this case.

[27] As Dellsén notes: 'Bernie's justification for believing [the con man's] testimony about the car, forms no part of the explanation of the car's breaking down' (2017:9).

However, there is a clear sense in which Bernie's apparent understanding is no mere accident. One could argue that his expertise in automobile mechanics makes this example of Dellsén's sufficiently robust to avoid this argument of Grimm's. Although, just as the child's knowledge survives the random assignment of an atlas which correctly labels the capital of Austria, in light of Bernie's expertise, one might think that his own knowledge survives the fortuitous nature of the con man's testimony.

Clearly, whether Bernie is taken to be capable of *both* understanding and knowledge or *neither*, will depend upon the extent to which Bernie's expertise are thought to "trump" the fortuitousness of the con man's testimony. What is important for my current purposes, of course, is that neither option suggests that Bernie's case serves to sever the connection between understanding and knowledge: 'on any way of filling out the details knowledge and understanding seem to sway together' (Grimm, 2006:520).

As I outlined earlier, in sections VI and VII, I shall argue that that there are pragmatic considerations which suggests that, in general, Comanche-style cases ought to be interpreted as instances of neither understanding nor knowledge. For the time being, however, I shall return to Alice's case, and show that it too is analogous to Kvanvig's Comanche case.

V. Alice's Case Revisited

To briefly refresh our memories, Alice is a student who manages to successfully derive the Pythagorean theorem using a version of the original proof. In light of Alice's dismal track record of failure in previous school assignments, Dellsén argues that she is not justified in believing that her attempt at deriving the Pythagorean theorem is successful. As such, even

though Alice 'clearly' understands the Pythagorean theorem she has no knowledge of it, lacking the requisite justification (Dellsén, 2017:7).

It should come as no surprise that Alice's case is open to the same interpretations as Bernie's. The most obvious of these interpretations presents Alice's case as an analogue of Grimm's second interpretation of the Comanche case. Here, we might note that Alice comes to understand the Pythagorean theorem using 'a version of Pythagoras's original proof' (Dellsén, 2017:6). In this sense, it appears that the reliability of Alice's source of information is beyond dispute, being Pythagoras himself.[28]

Just as when we focus upon the reliability of the author in the Comanche case, or on Bernie's expertise, it could be argued here that Alice's track record of academic failure is outweighed by the reliability of her information source. Thus, while we can grant that Alice understands the Pythagorean theorem, she also seems to *know* the Pythagorean theorem.

However, in response to a similar line of reasoning (put forward by an anonymous reviewer), Dellsén explicitly rejects this interpretation of Alice's case. Instead, he provides an alternative description of Alice's case which suggests that it ought to be interpreted in line with Grimm's first interpretation of Kvanvig's Comanche case:

> 'we could easily stipulate that Alice, for whatever reason, got lucky in her attempt to prove the Pythagorean theorem and that she fails to construct similar proofs in geometry on other occasions (and/or that she would have failed to construct the proof of the Pythagorean theorem in most nearby possible worlds), in which case *her belief-*

[28] This claim is not, strictly speaking accurate, since the original proof of Pythagorean theorem was given not by Pythagoras, but by Euclid in the first book of his *Elements*. Indeed, the statement of the theorem was known (without proof) to the Babylonians long before either Euclid or Pythagoras were born. I take it that one could happily substitute 'Euclid' for 'Pythagoras' without any loss of content to the argument above. However, to avoid any confusion, I have ignored this minor historical inaccuracy in the main body of the text.

> *forming process would not have been reliable* (Dellsén 2017:7, fn.10, italics are our own).

On Dellsén's preferred interpretation then, we should take Alice herself to be the source of information in this case, with her track record of past failure in school assignments serving to establish Alice as an *unreliable source*. This reading, of course, makes Alice's case analogous to the Comanche case where the author of the textbook bases their research upon extremely shoddy methodology.

In this instance, Grimm (2006) argues, even if the information contained in the textbook is correct, the reader of the textbook cannot be said to understand *why* the Comanche dominated the Southern Plains. It is clear to me that the same conclusion results from this interpretation of Alice's case. To see why, we will need to briefly return to a specific aspect of Bernie's case discussed above.

In the previous section, we saw that Bernie's expertise (resulting from his being a retired mechanic) are of crucial importance to Dellsén's claim that he can be said to understand the problem with the con man's car. Imagining a scenario where *I* was in Bernie's shoes, knowing practically nothing about the proper functioning of cars, any thought that I could be said to understand the problem is, quite clearly, off the table.

Suppose, however, that the con man pushes us to provide a diagnosis for his car's issue (perhaps, somewhat ironically, he does not believe our protestations of vehicular ignorance and instead thinks us overly modest). Fortuitously, I had been watching a documentary about cars just that morning in which broken timing belts were mentioned, and since this is the first thing which comes to our minds, I put this forward as a possible explanation of the car's behaviour prior to the breakdown.

The fact that our proposed explanation is correct here does nothing to alter our original verdict that I *do not* understand that the issue with the con man's car is a broken timing belt. This situation clearly mirrors Alice's. Just as I 'got lucky' in our attempt to explain the problem with the con man's car, Alice 'got lucky' in her attempt to prove the Pythagorean theorem' (Dellsén, 2017:7, fn. 10). Given that understanding is quite clearly absent in the former case, there is little reason to think that Alice understands Pythagorean theorem in the latter.

The issue here, once again, is that it is simply not plausible to suppose that understanding is compatible with this level of epistemic luck. Recall that Dellsén suggests that Alice 'would have failed to construct the proof of Pythagorean theorem in most nearby possible worlds' (Dellsén, 2017:7, fn. 10). This modal fragility appears to be wildly at odds with Dellsén's further description of the *process* of Alice's understanding, that she 'grasped how to prove the Pythagorean theorem, realizing how later steps in the proof follow from earlier steps' (Dellsén, 2017:7, fn. 10).

If Alice truly grasps the interaction between these steps of her proof, it seems highly likely that she *would* be able to perform these same steps in nearby possible worlds. Putting aside the contradictory nature of these accounts of Alice's case, this latter idea does not help Dellsén. Just as with Bernie's case, in drawing our attention away from the sense in which Alice got lucky, we arrive back at the interpretation discussed at the beginning of this section (which Dellsén rejects), that Alice is in possession of both understanding *and* knowledge.

So far, I have argued that neither Alice's case, nor Bernie's, support the thesis that understanding is separable from knowledge. Given that the separability of these notions is a necessary condition of Dellsén's claim that SP tracks understanding *rather than* knowledge, the noetic account certainly seems to be on shaky ground. Despite this, my task is not yet

complete, since one example of Dellsén's remains untested: Einstein's use of the kinetic theory of heat to explain Brownian motion.

In the next section, I provide an analysis of this case, showing that it too is an analogue of Kvanvig's Comanche case. What this means is that, like those cases described so far, Einstein's case does not stand as a counterexample to the orthodox stance within philosophy of science, which holds that understanding is a species of knowledge. However, I also highlight an illuminating comparison between Einstein's case and a recent SARS-COV-2 study by Lourenço *et al* (*ms*). In section VII, I show that this comparison suggests that Comanche cases ought to be interpreted in line with Grimm's (2006) initial analysis, as instances of neither understanding nor knowledge.

VI. Einstein's Case Revisited

As we saw in section II, Dellsén's third case concerns Albert Einstein's (1905/1956) explanation of Brownian motion in terms of the kinetic theory of heat. Since the information available at the time was lacking, Dellsén (2016:76) argues that Einstein was clearly also lacking the justification required to *know* that the movements at issue were, in fact, real.

Consequently, on the epistemic account, despite the fact that Einstein's work on this topic 'is widely considered to be one of the most significant achievements of one of history's greatest scientists' (Dellsén, 2016:77), it does not count as progressive. The noetic account, on the other hand, can seemingly make sense of Einstein's progress here, 'because he enabled us to grasp how to correctly explain Brownian motion, thereby providing us with understanding of something that we were previously unable to understand' (Dellsén, 2016:76).

However, it is our contention that Einstein's case, like Alice's and Bernie's, is open to two conflicting interpretations. One might think, given Einstein's undeniable status as one of history's greatest scientists, that he has the relevant expertise to warrant the ascription of understanding. Of course, in focusing upon the extent to which Einstein's explanation was no mere fluke, we are driven towards the conclusion, as in Bernie's case, that Einstein also *knew* Brownian motion to be explainable in terms of the kinetic theory of heat.

However, given the information available to Einstein in 1905, there is a very real sense in which he 'got lucky' in his explanation of Brownian motion. As Dellsén highlights, not only was Einstein's information regarding Brownian motion lacking precision, but '[t]he kinetic theory of heat was very much up for debate at the turn of the 20th century' (2016:76). In this sense, following Grimm's (2006) reasoning, with respect to his first interpretation of the Comanche case, it might seem that we should attribute neither understanding nor knowledge to Einstein in 1905.

So far, I have not shown a preference for either of these interpretations. Such a stance is not, strictly speaking, necessary in order to show that Dellsén's noetic account of SP is undermotivated. What matters for this argument is, as I have shown, that neither interpretation supports the thesis that understanding is separable from knowledge. However, I believe that a closer analysis of Einstein's case suggests that it is Grimm's first interpretation which should be adopted with respect to Comanche-style cases. My reasoning here is based upon a particular feature of understanding recently highlighted by Daniel Wilkenfeld (2017).

In his 2017 paper 'MUDdy understanding', Wilkenfeld notes that 'if we are not already in a situation where we know whether… the state of affairs represented by the explanans obtain, then we cannot know whether a particular explanation is *actually* conducive to

maximal understanding' (2017:1290).[29] In all three of the examples appealed to by Dellsén we know that the affairs represented by the explanans obtain.

In Alice and Bernie's cases, this is simply a stipulation of the scenario. In Einstein's case, we look back with the benefit of hindsight upon an explanation now universally accepted to be (at least approximately) true by the scientific community. However, when we consider an example where the accuracy of a given explanation is unknown, we come to realise that Comanche-style cases have only one viable interpretation.

Our chosen example concerns a recent study, carried out by Lourenço *et al* (*ms*) at the University of Oxford, which caused something of a media frenzy, in reporting that half of the UK population could have been infected with SARS-CoV-2 as early as 19/03/2020.[30] The authors of the study use the reported deaths from COVID-19, in the UK and Italy, to back-calculate the number of people infected with SARS-CoV-2.

As Hunter (2020) describes the study, it presents 'an SIR (Susceptible-Infected-Recovered) model that [the authors] calibrate to the epidemic trajectory in both Italy and the UK using Bayesian approaches. Using a range of assumptions, they conclude that already a large proportion of the UK population, [possibly] up to 68% may already have been exposed to infection'.

However, this conclusion has met with some scepticism from the wider scientific community. According to Nairsmith (2020), for example, the study 'rests on a key

[29] Wilkenfeld's (2017) paper concerns de Regt's (2015) antirealist notion of understanding, which was briefly mentioned in section II. As far as I am aware, no one has yet developed such considerations with respect to understanding as a characterization of scientific progress.

[30] This story was originally published in *Financial Times* (see: https://www.ft.com/content/5ff6469a-6dd8-11ea-89df-41bea055720b), although it was later picked up by *The Times, Daily Express, Evening Standard, Daily Mail* and *The Sun.* It is important to note that, at time of writing, this research has still yet to be published, or even peer-reviewed. In an interview with *Wired* (see: https://www.wired.co.uk/article/coronavirus-infections-oxford-study-immunity?intcid=inline_amp), Tim Colbourn (2020) highlights that '[i]t is a little concerning that they've taken it straight to the media… [i]t has not been properly sense-checked against any data'. It is precisely this unusual situation which makes this case-study fit for my purposes.

assumption which may or may not be the case'. Expanding upon this theme, Wood (2020) notes that '[t]he work merely makes assumptions about asymptomatic infection and mortality rates, but cannot measure them'. A further assumption made by the authors, highlighted by Gubbins (2020), is the proportion of the population at risk of severe disease, a factor which is unknown.

Despite this, there does appear to be a relative consensus regarding the key take away from the study, which is the need for more thorough serological studies in areas where epidemic spread has occurred. As Woolhouse (2020) puts it: the study's conclusion is a *hypothesis*, not a fact: 'a proper test will come from serological surveys [or large-scale surveys of virus genome diversity] – which will tell us how many people have been exposed.'

Like Einstein's explanation of Brownian motion, the study conducted by Lourenço *et al* puts forward an explanation which is compatible with, but underdetermined by, the evidence available at the time. Following Dellsén's reasoning in Einstein's case, we can assume that, at time of writing, Lourenço *et al* lack justification for believing that as much as 68% of the population of the UK had been infected with SARS-CoV-2 by 19/03/2020. What this means is that Lourenço *et al*, like Einstein in 1905, cannot *know* this to be the case and thus, on the epistemic account, this study does not count as progressive.

The key difference between these two cases is that the hypothesis put forward by Lourenço *et al* (*ms*) remains just that: *a hypothesis*. At this point, we have no way of knowing how accurate their model is. Unlike the cases concerning Alice, Bernie and Einstein, we do not know that the states of affairs represented by the explanans obtain. As a result, it appears that we cannot make any pronouncement about the ascription of progress on the noetic account either. This is because Dellsén's quasi-factive notion of understanding requires that 'the explanatorily/predictively essential elements of a theory must be true' (2016:73 fn.6). It

is precisely these explanatorily/predictively essential element which are called into question by Nairsmith (2020), Wood (2020) and Gubbins (2020).

However, it is important to note that this is the same position in which we find Einstein in 1905. Just as we are currently unable to provide a verdict with respect to the progressive nature of the SARS-CoV-2 study, in 1905, it would have been impossible to provide a verdict with respect to Einstein's explanation of Brownian motion. With the kinetic theory of heat being hotly contested at the time, and Einstein's information regarding Brownian motion lacking precision, the verisimilitude of the explanatorily/predictively essential elements of this theory would be epistemically inaccessible.[31]

VII. The Problem of Epistemic Access

The obvious next question concerns what it would take for us to be able to say whether or not the hypothesis put forward by Lourenço *et al* (*ms*) constitutes SP. It appears that the answer to this question is the same for both the epistemic and noetic accounts. This situation would require what Woolhouse (2020) calls a 'proper test', that is, widespread serological, or virus genome diversity, surveys.

Let us call the time at which an original hypothesis is put forward t_h and the point at which a hypothesis is corroborated by evidence t_c. Suppose that in several months, we arrive at t_c, the results of the appropriate surveys are in, and we have confirmation of the accuracy of the assumptions made by Lourenço *et al* (*ms*) at t_h. At this point, Dellsén would presumably be happy to say that this study constitutes SP since, we would now be able to

[31] There may well be room for debate upon this point although, as Dellsén himself highlights, 'we can easily imagine a world in which Einstein's explanation was put forward before the kinetic theory of heat became sufficiently justified to be known (e.g. shortly after James Clerk Maxwell first presented his kinetic theory in 1859)' (2016:76).

assert that Lourenço *et al* (*ms*) provided a correct explanation/prediction despite, at the time, lacking justification for believing it.

The problem here, is that the results of future surveys which would vindicate their position, allowing Dellsén to grant that the explanatorily/predictively essential elements of their theory are true (and thus that theirs is a genuine case of SP), would also provide precisely the sort of *justification* which is required for SP on the epistemic account.

So, we are currently unable to ascribe either understanding or knowledge to Lourenço *et al,* and as a result, we cannot make a pronouncement as to the progressive (or otherwise) nature of their research according to either account of SP. However, the point at which we would be able to reasonably maintain that this piece of research constitutes scientific understanding, t_c, is also the point at which we would be able to reasonably maintain that this piece of research constitutes scientific knowledge. Once again, knowledge and understanding sway together.

This same situation appears to hold for Einstein's explanation of Brownian motion. Dellsén is in a position to grant that Einstein's hypothesis constitutes progress only because it is now thought to be true. However, this situation is largely due to the later work of Jean Perrin, whose experimental verification of Einstein's explanation of Brownian motion (among other advancements) earned him a Nobel Prize in 1926.

Perrin's (1909) work plays the exact same role with respect to Einstein's case as widespread serological or virus genome diversity surveys would with respect to the case of Lourenço *et al:* providing *justification* for the belief that Brownian motion can be explained by the kinetic theory of heat. In 1905, when Einstein first put this hypothesis forward, we would have been unable to make a pronouncement as to the progressive nature of his proposed explanation. By the time the explanatorily/predictively essential elements of

Einstein's theory could reasonably be accepted as true, and thus qualify for understanding, they could also be justifiably believed; both elements being grounded in Perrin's (1909) experimentation.

It would appear that, in order for us to be able to attribute understanding to the likes of Einstein or Lourenço *et al*, we must first be in a position to assert that the explanatorily/predictively essential elements of the relevant theory are true. However, the truth of such elements will remain epistemically inaccessible until such time as confirmation, sufficient to supply epistemic justification, has been attained.

A natural response to this argument, would be to point out that even though it would have been impossible to make such a judgement at the time, the noetic account can still *retrospectively* assign understanding to Einstein, since we *now* know his explanation to be (a least approximately) true. The epistemic account can make no such claim, since in 1905 justification was lacking and, unlike understanding, justification cannot be retrofitted. However, it is difficult to see how this idea stays true to the central tenet of the noetic account of SP, that progress in science tracks understanding *rather than* knowledge.

If, as our argument suggests, Dellsén's realist notion of understanding can only be ascribed *after* a theory has met with confirmation (and thus, justification) of some kind, the idea that understanding is playing the fundamental epistemic role here becomes a hard one to swallow. The case of Lourenço *et al* (*ms*) highlights this nicely. When we think about what it would take for us to able to grant that this study constitutes progress, our intuitions pull us towards those criteria which would also provide justification for believing their hypothesis, i.e. widespread serological, or virus genome diversity, surveys. That this knowledge would allow us to retroactively assign understanding on behalf of Lourenço *et al* at t_h, does not

change the fact that the tell-tale sign of progress in this case is *justification,* and thus*, knowledge*.

This issue also appears to raise another, concerning how the noetic account is to make sense of confirming instances themselves. According to Dellsén, remember, Einstein makes scientific progress in 1905 by proposing an explanation of Brownian motion in terms of the kinetic theory of heat. Nonetheless, confirmation of Einstein's theory is not attained until around 1908. Clearly, this confirmation is of crucial importance to Einstein's theory since, as we have already seen, it is Perrin's (1909) work which allows for the ascription of knowledge at t_c *and* Dellsén's retroactive attribution of understanding at t_h.

However, it is difficult to see how the noetic account can make sense of this, clearly progressive, step. After all, if science progresses through increasing understanding, and Einstein can be said to understand Brownian motion in 1905, what work is there left for Perrin's experimentation to do? Verification constitutes a significant epistemic boon for Einstein's theory; although, the noetic account of SP seems to be incapable of accounting for it: the proverbial horse has already bolted.

Concluding Remarks

In this paper, I have argued that the noetic account of SP, which suggests that progress in science tracks understanding rather than knowledge, is undermotivated. It is undermotivated because the examples which Dellsén uses to support the idea that understanding is separable from knowledge fail in this purpose. I have shown that the cases of Alice, Bernie and Einstein are, in fact, (somewhat disguised) Comanche-style cases. As a result, these scenarios fall foul of Grimm's criticism of Kvanvig's original Comanche case.

In section III, we saw that Grimm (2006) highlights two conflicting interpretations of such cases. They can be seen as either: instances where neither knowledge nor understanding are present; or as instances in which both are present. The problem for Kvanvig (and, as a result, Dellsén) being that neither interpretation supports the thesis that such cases can distinguish understanding from knowledge. In sections IV and V, I showed how Grimm's argument can be applied to the Alice's and Bernie's cases yielding the same result: neither supports Dellsén's position.

In sections VI and VI, I argued that Einstein's case is also susceptible to Grimm's argument against Kvanvig's Comanche case, rendering this scenario similarly silent with respect to a separation of understanding and knowledge. However, I maintain that Einstein's case allows us to show that Grimm's first interpretation of such Comanche-style cases is the correct one.

Through a comparison with a recent study by Lourenço *et al* (*ms*), I argued that neither understanding nor knowledge ought to be attributed to Einstein in 1905, because the accuracy of the explanatorily/predictively essential elements of his theory were not known, and SP (in both cases) appears to be intuitively tied to those episodes with bring about such knowledge. What's more, by tying SP to understanding, the noetic account appears to jump the gun with respect to these episodes, being incapable characterizing the clear sense in which they *are* progressive.

Chapter 2

The Case for Intervention Liberalism

One rather surprising feature of recent interventionist analysis of noncausal explanation, is that they typically jettison the core feature of interventionism: interventions. Indeed, the prevailing opinion within the philosophy of science literature suggests that interventions exclusively demarcate causal relationships. This position is so prevalent that, until now, no one has even thought to name it. I call it "intervention puritanism". In this paper, I mount the first sustained defence of the idea that there are distinctively noncausal explanations which can be characterized in terms of possible interventions, and thus that I-puritanism is false. I call the resulting position "intervention liberalism" (I-liberalism, for short). While many have followed James Woodward (2003) in committing to I-puritanism, I trace support for I-liberalism back to the work of Jaegwon Kim (1974). Furthermore, I analyse two more recent sources of scepticism regarding I-liberalism: debate surrounding mechanistic constitution, and recent attempts to provide a monistic account of explanation. I demonstrate that neither literature provides compelling reasons for adopting I-puritanism. Finally, I present a novel taxonomy of available positions upon the role of possible interventions in explanation: weak causal imperialism; strong causal imperialism; monist intervention puritanism; pluralist intervention puritanism; monist intervention liberalism; and finally, the specific position defended in this paper, pluralist intervention liberalism.

I. Introduction

Recent years have seen increasing interest in the prospect of modifying the interventionist analyses of causal explanation, popularized by Woodward (2003), in order to characterize explanations which are seemingly *noncausal* in nature.[32] However, one odd feature typically shared by such accounts is that they jettison the core feature of interventionism: *interventions*.

[32] While Woodward's (2003) interventionist analysis of causation is generally taken to be 'the standard philosophical account' (Wilson, 2018:18), Woodward himself attributes the term "intervention" to Meek & Glymour (1994) and Pearl (2000). Also see: Hitchock (2001), Pearl (2009) and Briggs (2012).

Indeed, the dominant position within the philosophy of science literature suggests that, roughly speaking, where it is possible to intervene upon *X*, in such a way that changes the value of *Y*, *X causes Y*. Which is to say that interventions exclusively demarcate *causal* relationships.[33]

So prevalent is this position that, until now, no one has seen fit to name it. I call it "intervention puritanism" (*I*-puritanism, for short). While dissenting voices have begun to appear (including Woodward [2018b] himself), this paper represents the first sustained defence of the idea that there are distinctively *non*causal explanations which can be characterized in terms of such interventions; in other words, that *possible* interventions do not carve nature at its causal joints.[34] I call this position "intervention liberalism" (*I*-liberalism, for short).

Given the relatively recent emergence of interest in interventionism with respect to noncausal explanation, it might come as some surprise to discover that precedence for *I*-liberalism can be found as far back as the 1970s.[35] In a series of (largely overlooked) papers,

[33] See, for example: Woodward 2003, 2015; Bokulich 2011; Leuridan 2012; Saatsi and Pexton 2013; Harinen 2014; Pexton 2014; Baumgartner & Gebharter 2015; Rice 2015; Romero 2015; Reutlinger 2016, 2017, 2018; Baumgartner & Casini 2017; French and Saatsi 2018; Khalifa *et al* 2018, 2020; Saatsi 2018; Woodward 2018b; Jansson & Saatsi 2019; Lange 2019, 2021.

[34] There is currently ongoing debate regarding the explanatory status of *impossible* interventions, with several authors having recently argued for their application in noncausal explanations across mathematics, logic, and metaphysics (see e.g., Schaffer (2016, 2017); Wilson (2018a, 2018b, 2021); Baron *et al* (2017); Baron *et al* (2020); Reutlinger *et al* (2020); Baron & Colyvan (2021); and Baron (2022). It is widely understood that such impossible interventions require the rejection of traditional counterfactual semantics (see e.g., Baron & Colyvan 2021:564-567). For the purposes of *this* paper, however, the reader ought to assume that where I use the term "intervention", unless explicitly stated otherwise, I mean "physically possible intervention". One *prima facie* reason for limiting my account in this way is that (as we shall soon see) such cases do not require any substantial modification to be made to the typical interventionist methodology, and thus constitute the most robust form of counterexample to *I*-puritanism. What is more, as will be discussed further in sections V. & VIII., appealing to the role of impossible interventions does not, by itself, constitute a rejection of *I*-puritanism.

[35] In fact, I believe that something like this idea can be traced back to C. S. Peirce (1931-58), who argues that even pure mathematics and logic concern 'operations on diagrams, whether external or imaginary, [which] take the place of the experiments upon real things that one performs in chemical or physical research' (*Collected Papers* 4:530; 1905). Also see Peirce (*Writings* 3:41;1872). While certainly worthy of further investigation, such a task is strictly beyond the scope of this paper.

Kim argues against *causal imperialism*, the view that *all* explanations track causal relations.[36] In 'Causes and Counterfactuals' Kim (1973) highlights cases of asymmetric counterfactual dependence which motivate the existence of distinctly noncausal explanations. In 'Noncausal Connections', Kim (1974) goes further, arguing that both causal and noncausal dependence can be characterized in terms of the "bringing about" relation. Where *A* depends upon *B* (causally or otherwise), Kim suggests, we can bring about *B* by first bringing about *A,* but not *vice versa.*[37]

More recently, Woodward's (2003) manipulationist account of causal explanation has given Kim's intuitive conception of "bringing about" a formal characterization through the notion of a possible intervention; making use of structural equation models to encode an asymmetric pattern of interventionist counterfactuals. In the next two sections of this paper, I revisit Kim's analysis of noncausal explanatory dependence and attempt to reconcile his position within a contemporary interventionist framework.

In section II., I outline Kim's motivation for suggesting that noncausal counterfactual dependence can be characterized in terms of "bringing about". In section III., I introduce Woodward's (2003) analysis of causal explanation and demonstrate how Kim's analysis of noncausal explanation can be fruitfully accommodated with the framework of structural equations models and interventionist counterfactuals. As it transpires, the sort of noncausal

[36] Causal imperialism (a term borrowed from Bokulich [2018]) is most closely associated with Railton (1981); Lewis (1986); Strevens (2008); and Skow (2014). This position is to be distinguished from *I*-puritanism, which is neutral with respect to whether all explanations are causal explanations. As we shall see, many hold that while noncausal explanations *do* exist, they are not characterizable in terms of interventionist counterfactuals. I discuss these distinctions in more detail in section VII.

[37] A popular position within both philosophy of science and metaphysics suggests that 'explanations must depict dependence relations' (Potochnik, 2017:105). While Kim's analysis focuses on causal and noncausal *dependence* specifically, following the likes of Ruben (1990); Kim (1994); Strevens (2008); Audi (2012); Craver (2014) Schaffer (2016); and Kovacs (2017), I shall assume that wherever one finds a dependence relation of the sort in question, an explanation follows. While there remain dissenting voices (Dasgupta 2017; Taylor 2018; Khalifa *et al* 2018; Thompson 2018), the vast majority of those involved in debate surrounding noncausal explanation would at least concede something close to this position.

explanatory dependence highlighted by Kim can be happily cashed out in terms of possible interventions.

I-liberalism has been staunchly opposed within the recent philosophy of science literature, and there are two distinct (but related) debates which have served as focal points of this scepticism. The first concerns a particular type of noncausal explanation: constitutive explanation. In response to Craver's (2007a, 2007b) attempts to characterise constitutive explanation in terms of *mutual manipulability*, one common counter has been that, since manipulability is an essentially causal notion, Craver's account fails as a characterisation of *non*causal explanation.[38]

The second such debate concerns recent interest in providing a *monistic* account of the asymmetry of causal and noncausal explanation. Here it is once again widely assumed that possible interventions characterize only causal relationships and, as such, that they must be jettisoned when providing an account of explanation which unifies causal and noncausal instances.[39] In both debates, however, motivation for *I*-puritanism is remarkably thin on the ground, principally relying upon Woodward's (2003) own commitment to this position.

In section IV., I demonstrate that *I*-liberalism *can* accurately characterise an archetypal case of constitutive explanation: the nastic movement of *Mimosa Pudica.* In section V., I argue that, although Woodward (2003) does appear to commit himself to *I*-puritanism, such an interpretation of his interventionist framework is far from obligatory. What Kim's intuition regarding the "bringing about" relation shows, is that causal relationships do not exhaustively describe the ways in which agents can manipulate the world around them.

[38] See, e.g. Craver 2007a, 2007b; Leuridan 2012; Harinen 2014; Romero 2015; Baumgartner & Gebharter 2016; Cassini & Baumgartner 2016; and Krickel 2018.
[39] E.g. Saatsi and Pexton (2013); Jansson (2015); Reutlinger (2016, 2017); French and Saatsi (2018); Lange (2019); Khalifa *et al* (2020).

In section VI., I consider one of the few arguments against *I*-liberalism which does not rely upon Woodward's own *I*-puritanism. In the process of arguing against explanatory monism, Khalifa *et al* (2020) suggest that *I*-liberalism is untenable precisely because it cannot distinguish between cases of genuine noncausal dependence and cases analogous to spurious correlation resulting from some common explanatory source. Conversely, I argue that, although analogous spurious correlations *do* arise with respect to noncausal explanation, constitutive explanations are not among them and that, where they occur, *I*-liberalism is perfectly capable of dealing with such cases.

In the final section, I provide a novel taxonomy of available positions upon the role of interventions in explanation. I highlight six such positions: weak causal imperialism; strong causal imperialism; monist intervention puritanism; pluralist intervention puritanism; monist intervention liberalism; and finally, the position defended in this paper, *pluralist intervention liberalism.*

II. Kim on Noncausal Connections

In 'Causes and Counterfactuals', Kim (1973) highlights several cases which appear to undermine the causal imperialist claim that counterfactuals exclusively express *causal* dependencies.[40] Take the relationship between Xanthippe's becoming a widow and the death of her husband, Socrates. Does Socrates' death *cause* Xanthippe's widowhood? There are reasons to think not. First and foremost, these events are spatially discontiguous. As Kim highlights, it would be an unforgivable afront to physics to accept that such causal action could be 'propagated instantaneously through spatial distance' (1974/1993:13).[41] Second,

[40] Lewis (1973), being Kim's explicit target.
[41] It is interesting to note that Woodward believes that his interventionist methodology provides reason for denying that spatiotemporal contiguity is a defining characteristic of causation (2003:36). Although, in more

presuming that individual causal relations instantiate nomic regularities, it is difficult to think of a contingent empirical law capable of subsuming these events (Kim, 1974/1993:13).

If Socrates' death is not the cause of Xanthippe's becoming a widow, then perhaps whatever caused Socrates death was *itself* the cause of Xanthippe's widowhood. These events might, then, be the joint effects of a common cause: Socrates' having ingested hemlock. Indeed, this interpretation of the situation seems to allow for a 'nice Humean law' which subsumes hemlock consumption and widowhood: 'given the law, let us assume, that anyone who drinks hemlock dies, we have the law – at least a Humean regularity – that anyone whose husband drinks hemlock becomes a widow' (1974/1993:30).

The problem with this interpretation is that the only route from hemlock to widowhood seems to *go through* death. If Socrates' having ingested hemlock causes Xanthippe's widowhood, it does so only by first causing his death, which puts us back where we started. Consequently, if neither Socrates' drinking hemlock, nor any other apparent cause of his death could count as the cause of Xanthippe's becoming a widow, we can only conclude, according to Kim, that Xanthippe's being widowed has no cause at all.

And yet, these events are obviously connected in some sense. Indeed, notwithstanding those features discussed above, there are some clear similarities between this type of noncausal connection and archetypal causal explanations. First and foremost, Kim argues, the sort of explanatory asymmetry which we might expect from a cause and effect relationship can be drawn out when considering the counterfactual conditionals related to these events:

> If Socrates had not died at *t*, Xanthippe would not have become a widow at *t*.
>
> If Xanthippe had not become a widow at *t*, Socrates would not have died at *t.*

---

recent work, Woodward (2018b) accepts this particular example as an instance of distinctively noncausal dependence.

The counterfactual dependence between these two events is 'irreversible'; while the former is straightforwardly true, in response to latter, Kim notes, 'we would more likely alter the marital condition of Socrates than tamper with the fact of his death at *t*' (1974/1993:24). This irreversibility becomes clearer when we examine a second form of asymmetry which Kim notes with respect these events: asymmetry in the agency relation. Consider the following counterfactuals:

> By bringing about Socrates' death, we could bring about Xanthippe's widowhood.
>
> By bringing about Xanthippe's widowhood, we could bring about Socrates' death.

What Kim's intuition regarding these counterfactuals suggests, is that if we wished to make Xanthippe a widow, facilitating Socrates' death would be the best (indeed only) way to go about doing it. On the other hand, attempting to make Xanthippe a widow would not be an *effective strategy* (to use a term of Cartwright's [1979]) to bring about Socrates' death. In much the same way, while we might increase the length of a pendulum to bring about an alteration in its period of swing, altering its period of swing would not be an effective strategy to bring about an increase in its length. Such events are not the result of coincidence or brute fact, but 'are determined by other events; their occurrence is completely dependent on the occurrence of others, but this is not to say that they are causally determined by them' (Kim 1974/1993:30).

In concluding these observations, Kim tentatively puts forward the thesis that both causal and noncausal connections might be characterised, monistically, in terms of a single unifying relation "*R*": 'a broad relation of dependency that subsumes as special cases the causal relation and other dependency relations' (1974/1993:27). Unfortunately, Kim

(1974/1993) does not provide a substantive account of what this relation might be.[42] With the benefit of Woodward's (2003) interventionist methodology, however, I believe that a more formal characterization to Kim's intuitions regarding the "bringing about" relation can be given in terms of interventionist counterfactuals and structural equations.

III. An Interventionist Account of "Bringing About"

According to Woodward, any attempt to characterise causal dependence ought to begin by considering the practical utility of our notion of causation; what does causal knowledge allow us to achieve that information about mere regularity or correlation, will not? (2003:28). In answer to this question, Woodward suggests that 'it is heuristically useful to think of explanatory and causal relationships as relationships that are potentially exploitable for the purposes of manipulation and control' (2003:25). It is manipulability, then, that distinguishes explanatory counterfactuals from nonexplanatory counterfactuals (the latter of which arise as the result of *mere* correlation).[43]

To say that *X* causes *Y*, on this picture, is to say that *Y* would change in value under some suitable intervention that changed the value of *X*.[44] Where an intervention on *X* with respect to *Y* 'changes the value of *X* in such a way that if any change occurs in *Y*, it occurs

[42] This is likely due, at least in part, to historical timing. While others (e.g. Collingwood 1944; Gasking 1955; and von Wright 1975) had already argued that 'causes are, as it were, levers for moving effects', such "agential" or "manipulationist" accounts of causal explanation were thought to run into 'intractable difficulties', and were largely abandoned (Hausman, 1982:45). It should come as no surprise then, that Kim's project of accounting for both causal and noncausal explanation in terms "bringing about" found little contemporaneous support. By the turn of the twentieth century, however, the tide had well and truly turned. Thanks, in no small part, to a spirited defence by Menzies and Price (1993), which updated several crucial elements of previous agential theories (such as relinquishing a commitment to determinism), and convincingly circumvented many of the seemingly intractable difficulties mentioned above.

[43] I shall discuss the sort of nonexplanatory counterfactuals which arise as a result of mere correlation, at much greater length, in section 6.

[44] Where it is possible to intervene upon *X* with respect to *Y* in this way, one might alternatively say that *X* is exploitable for the purposes of manipulation *Y*. As such, I shall use the terms *intervention* and *manipulation* interchangeably in what follows.

only as a result of the change in the value of *X* and not from any other source' (Woodward, 2003:14). More formally, *I* is an intervention on *X* iff:

I. *I* causes *X*;

II. *I* acts as a switch for all other variables that cause *X*. That is, certain values of *I* are such that when *I* attains those values, *X* ceases to depend on the values of other variables that cause *X* and instead depends only on the value taken by *I;*

III. Any directed path from *I* to *Y* goes through *X*. That is, *I* does not directly cause *Y* and is not a cause of any causes of *Y* that are distinct from *X* except, of course, for those causes of *Y*, if any, that are built into the $I \rightarrow X \rightarrow\rightarrow Y$ connection itself; that is, except for (a) any causes of *Y* that are effects of *X* (i.e., variables that are causally between *X* and *Y*) and (b) any causes of *Y* that are between *I* and *X* and have no effect on *Y* independently of *X*;

IV. *I* is (statistically) independent of any variable *Z* that causes *Y* and that is on a direct path that does not go through *X*. (Woodward, 2003:98):

As we have already seen, Kim is well aware of this agential dynamic to dependence. Indeed, his analysis of the asymmetric relationship between dependent events comes tantalizingly close to the core claim of Woodward's interventionism. Kim argues that the asymmetry of the agency relation, the sense in which 'by bringing about the cause, you bring about the effect', is a *result* of the asymmetry 'between states or events brought about by the action' (1974/1993:25).

While he does not use the term, it is clear that *manipulation* is something like the notion which Kim is intending to highlight with his discussion of the connection between agency and dependence. Indeed, the relationship between Socrates' death and Xanthippe

widowhood, appears to fit nicely into the sort of structural equations utilized by interventionists in modelling causation. Such a model consists of:

- A set of variables representing features of reality, in this case:

  C: Whether Socrates dies.
  E: Whether Xanthippe is a widow.

- A set of structural equations linking the values of these variables according to reality's causal structure, where '→' expresses counterfactual dependence:

  $E \rightarrow C$

- And, an assignment function specifying which values the variables actually take:

  $C = 1; E = 1$

For C to be considered a cause of E, it must be possible to intervene upon C, altering its value from 'C = 1' to 'C = 0', in such a way that will result in a change in the value of 'E = 1' to 'E = 0'. This means that it ought to be possible to intervene upon Socrates death in such a way that also prevents Xanthippe's becoming a widow. Suppose, for example, that Crito was to knock the hemlock from Socrates' hand before it could be consumed.[45] In this scenario, as a direct result of Crito's intervention, Socrates would survive, so 'C = 0' (assuming all other variables are held fixed), what is more, as a direct result of Socrates' survival, Xanthippe would not become widowed, so 'E = 0'.

This interventionist analysis of Socrates and Xanthippe's situation also allows us to cash out the sort of asymmetry which Kim highlights as a common factor in cases of both causal and noncausal dependence. In line with *III* above, it is precisely because any possible

[45] While Crito's knocking the hemlock from Socrates' hand is itself a causal process, the dependency relation which it speaks to, holding between Socrates death and Xanthippe's widowhood, is clearly noncausal. The idea that causal processes can give rise to noncausal explanations is not novel. I discuss further instances of this surprising detail, related to constitutive explanation, in section 4.

intervention upon Xanthippe's widowhood *must* go through Socrates' death, which suggests that the dependency here is asymmetric. Which is to say, Xanthippe's widowhood is not exploitable for the purposes of manipulating Socrates' death. And yet, for Kim, there are clear reasons for thinking that Socrates' death and Xanthippe's widowhood are *not* related as cause and effect in any ordinary sense.

What this appears to suggest is that possible interventions cannot stand as a useful dividing line between causal and noncausal explanation. The possibility of intervening simply does not carve nature at its causal joints. This conclusion will obviously come as a blow to *I*-puritans; without possible interventions to play this role, it is not obvious how we are to distinguish these different types of explanation.[46] For those without such pre-theoretical commitments, however, *I*-liberalism ought to hold some intuitive appeal. Afterall, possible interventions stand as an addition to the metaphysical toolkit through which noncausal dependence can be analysed.

One area where my analysis has obvious application is the ongoing discussion surrounding *constitutive* explanation, a popular hunting ground for *I*-puritans. In response to Craver's (2007a, 2007b) attempts to define constitution in terms of symmetrical interventions, many have argued that this account fails on *I*-puritan grounds. Since interventions are assumed to designate causal relations, and causal relations *alone*, it is argued that Craver cannot make sense of the noncausal dynamic to a phenomenon's being constituted by its spatiotemporal parts. However, in the next section, I apply the *I*-liberal methodology described above to an example of constitutive explanation.

[46] See Wilson (2020) for a thorough survey of plausible means by which one might seek to classify causal and noncausal dependence.

IV. Intervention-Liberalism and Constitutive Explanation

Among philosophers of science, constitutive explanations are typically taken to be a form of *mechanistic* explanation, where a mechanism consists of entities/part/objects and their activities/interactions/operations.[47] Constitutive mechanistic explanation is often distinguished from *etiological* mechanistic explanation. In the latter case, some mechanism explains a phenomenon for which it is *causally* responsible, whereas in the former, a phenomenon is explained by the underlying mechanism which *constitutes* it.

Take, for example, the nastic movement of *Mimosa pudica*.[48] Nastic movements occur in plants and fungi as a response to environmental stimuli (*thigmonasty*), with *Mimosa* being the most heralded example due to the dramatic nature of the response. Such movement is constituted by a release of potassium ions in the plant's pulvini cells, which lowers the cell's turgor pressure (pressure exerted on the cell wall due to exosmosis) and, in turn, collapses the cell's parenchyma tissue, constricting the vascular strand serving as a hinge (Esau, 1965).

This explanation allows us to identify the three parts of *Mimosa* that are involved in the phenomena of nastic movement (E*): the potassium ions in the pulvini cells (C^*_1), the turgor pressure of the pulvini cells (C^*_2), and the parenchyma tissue of the pulvini cells (C^*_3). As we saw in the previous section, in order to capture the noncausal dependence at play here, we ought to be able intervene upon the *Mimosa's* pulvini cells in such a way that will also affect the plant's nastic movement, but not *vice versa.* And this is exactly what we see.

---

[47] See, e.g. Machamer, Darden & Craver (2000); Craver & Darden (2002); Craver (2007b); Illari & Williamson (2012); Glennan (2017).

[48] Other examples of constitutive mechanistic explanation abound: Spatial memory (Craver 2007b; Bechtel 2008); action potential (Craver 2007b); the heart (Bechtel and Abrahamsen 2005; Glennan 2010; Craver & Darden 2013); cells synthesizing proteins (Machamer, Darden & Craver 2000; Darden 2002; Craver and Darden 2013); long-term potentiation at synapses of neurons (Machamer, Darden & Craver 2000; Craver & Darden 2001; Craver 2007b; Craver & Darden 2013).

Variables:

C^*_n: Whether the parenchyma tissue of the *Mimosa's* pulvini cells collapse.[49]

E*: Whether the *Mimosa* exhibits nastic movement.

Structural equations:

$E^* \rightarrow C^*_n$

Assignment:

$C^*_n=1$; $E^*=1$

Just as with the case of Socrates and Xanthippe, the nastic movement of the *Mimosa* can be manipulated through an intervention upon its pulvini cells. For example, administering potassium channel blockers (such as peptides containing the integrin-binding sequence RGD [Arg-Gly-Asp] [Jaffe *et al* 2002]), restricts potassium ions in the pulvini cells from affecting the cell's turgor pressure and, as such, prevents nastic movement from occurring; so here '$C^*_n=0$' and, as a result, '$E^*=0$'.

This interventionist approach to constitutive explanation also allows us to cash out the sort of explanatory asymmetry which Kim highlights as a common factor in cases of both causal and noncausal dependence. Once again, in line with condition *III*, it is precisely because there is no possible intervention upon the *Mimosa's* nastic movement that does not *go through* the *Mimosa's* pulvini cells, which suggests that the dependence here is asymmetric. Which is to say, the *thigmonsasty* of the *Mimosa* is not exploitable for the purposes of manipulating its pulvini cells.

[49] Here, I have condensed C^*_1-C^*_3 into a single variable. The relationship between spatio-temporal parts of a constitutive mechanism is typically taken to be causal (i.e. release of potassium ions in the pulvini cells causes the cell's turgor pressure to drop). However, my principal interest is in the *non*causal relationship between the constitutive mechanism and the phenomenon to be explained, as such combining these variables allows us to maintain the noncausal character of the arrow in the structural equations. Presuming, for the purposes of argument, that under experimental conditions, we can guarantee that the pulvini cells' parenchyma tissue will collapse *only* when the release of potassium ions decreases the turgor pressure within the cell, this ought to make no difference to my argument.

However, in their interventionist account of constitutive explanation, Craver (2007a, 2007b) has suggested that it is in fact *mutual* manipulability that defines such noncausal mechanisms. Craver argues that while 'one can change the explanandum phenomenon by intervening to change a component [of a mechanism]', one can *also* (contrary to my account) 'manipulate the component by intervening to change the explanandum phenomenon' (2007b:153). As such, Craver concludes that all constitutive dependency relationships are "bidirectional".

Yet, Craver's account is clearly problematic on two related fronts.[50] First, as Romero (2015), Baumgartner & Gebharter (2016), and Krickel (2018) highlight, and the example of *Mimosa* demonstrates, manipulations of the latter variant, whereby a component is manipulated via the explanandum phenomenon, are impossible by Woodward's (2003) definition of an intervention.[51] Second, supposing such "top-down" interventions were possible, given that interventions are intended to characterise explanatory relations, this would suggest that constitution entails explanatory *symmetry* (Schindler, 2013). And, as Khalifa *et al* note, 'a surefire way to embarrass a theory of explanation is to show that it fails to respect the commonsense idea that explanation is an asymmetric relation' (2018:1). [52]

The *I*-liberal interpretation of such mechanisms presented above, can easily avoid both of these issues, preserving the explanatory asymmetry which forms the heart of Kim's desire

[50] For further criticisms, not discussed here, see: Harinen 2014; and Cassini & Baumgartner 2016.
[51] Romero (2015) and Baumgartner & Gebharter (2016) argue that such interventions are actually 'fat-handed' rather than outright impossible. A fat-handed intervention is an intervention which violates "*III*" in as much as it manipulates both the mechanistic components and the phenomena in question *at the same time*, effectively serving as a common cause of both. While Woodward's (2003) definition of an intervention can, according to Romero (2015) and Baumgartner & Gebharter (2016) be altered to accommodate such a notion, Krickel (2018) has argued that this approach has severe limitations. In so far as my own position takes such interventions to be impossible, and thus beyond the scope of *I*-liberalism, it preserves more of the core of Woodward's (2003) original definition, and as such, stay truer to the character of his manipulationist account of explanation.
[52] One might well wonder why Craver introduces the notion of mutual manipulation at all. His motivation is principally to try and make sense of important "top down" research strategies within the life sciences, distinguishing between interference experiments, stimulation experiments and activation experiments (2007b:146-157). However, Baumgartner and Gebharter (2016) have recently argued that top-down experimentation can be made sense of without the need for top-down interventions.

for a unifying account of causal and noncausal dependence while, at the same time, ruling out the sort of top-down intervention which Craver (problematically) believes to be characteristic of constitutive explanation.[53] There is a further apparent issue with Craver's account, however, which is of much greater interest to us. In 'Three Problems for the Mutual Manipulability Account of Constitutive Relevance in Mechanisms' Bert Leuridan (2012) suggests that Craver's account entails that constitutive explanations are, in fact, *causal* explanations.

Leuridan's (2012) argument is that one cannot get away with embedding an account of constitutive explanation within an interventionist framework and emerge with a characterisation of *non*causal explanation: interventions, in other words, highlight *only* causal relations. Indeed, Baumgartner & Gebharter (2016) agree, noting the following slogan from Woodward: 'no causal difference without a difference in manipulability relations, and no difference in manipulability relations without a causal difference' (2003:61). This slogan is, of course, the central tenet of *I*-puritanism.

The literature surrounding constitutive explanation is not the only area where we find support for *I*-puritanism. Another topic which has elicited a great deal of discussion surrounding the essentially causal character of interventions is the broader project of providing a *monistic* account of explanation: 'an analysis that accommodates causal and noncausal explanations, and accounts for the asymmetries of both' (Khalifa *et al*, 2018). Here too, Woodward's (2003) interventionism has taken centre stage.

[53] The question of what distinguishes constitutive explanation from both other forms of noncausal dependence, and causal dependence, is an interesting one. Unfortunately, there is not the space here to discuss this topic at length. However, I would point out that there are obvious features of constitutive explanation which could serve as useful essential characteristics, e.g. a mechanism and the phenomena which it explains are typically taken to share the same spatio-temporal location, which distinguishes such cases from noncausal explanations like Kim's example of Socrates death and Xanthippe's widowhood, which are not so constrained.

Debate surrounding explanatory monism, and debate surrounding mechanistic explanation are closely connected. For example, a successful account of explanatory monism would, presumably, apply to mechanisms (both constitutive and etiological) as a limiting case.[54] It is no surprise, then, to find that both debates have motivated their commitment to *I*-puritanism along very similar lines. The principal motivation for the essentially causal character of dependencies characterized by interventions, is that Woodward (2003) himself assumes such an *I*-puritan stance. In the next section, however, after discussing explanatory monism in more detail, I argue that, although Woodward (2003) does support *I*-puritanism, this conclusion is not *entailed* by his interventionist analysis of causation.

V. Woodward's Intervention-Puritanism

An *almost* universal feature of recent attempts to provide a monistic interventionist account of explanation has been the abandoning of structural equation models and interventionist counterfactuals with respect to noncausal explanation.[55] Indeed, this appears to be a rare point of agreement, even among those who take the monist framework to be something other than counterfactual (e.g. Khalifa *et al* 2018), and those who advocate explanatory pluralism (e.g. Lange 2019). To take a popular example, Alexander Reutlinger (2017) proposes the following non-interventionist counterfactual criteria for a monistic explanation, where 'G_1,…, G_m' comprise generalizations, and 'S_1,…, S_n' comprise auxiliary statements:

Veridicality condition: G1, …, Gm, S1, …, Sn, and E are (approximately) true.

[54] Although, this is not to say that the existence of noncausal mechanistic explanation is committal with respect to monism.

[55] E.g. Saatsi and Pexton (2013); Jansson (2015); Reutlinger (2016, 2017); French and Saatsi (2018); Lange (2019); Khalifa *et al* (2020).

> *Implication condition*: G1, …, Gm and S1, …, Sn logically entail E or a conditional probability P(E|S1, …, Sn) – where the conditional probability need not be 'high' in contrast to Hempel's covering-law account.
>
> *Dependency condition*: G1, …, Gm support at least one counterfactual of the form: had S1, …, Sn been different than they actually are (in at least one way deemed possible in the light of the generalizations), then E or the conditional probability of E would have been different as well.

And yet, as Roski (2020) highlights, in stripping Woodward's (2003) account of the mechanism which characterizes causal asymmetry, namely interventions, Reutlinger's account appears to suffer from the same embarrassing explanatory symmetry which plagues Craver's (2007a, 2007b) account of constitutive noncausal explanation. Reutlinger is not alone along monists here.[56] As Marc Lange argues with respect to many other recent attempts to characterize explanatory monism:

> 'These attempts recognize that even when there is explanatory asymmetry, there may be symmetry in counterfactual dependence. Therefore, something more than mere counterfactual dependence is needed to account for explanatory asymmetry' (2019:1).

In light of the predicament in which symmetry places monist accounts of explanation, it would seem natural to expect to find some convincing reasons for rejecting the story which was told in the first half of this paper; that interventions stand to characterize certain explanatory asymmetries across both causal and noncausal instances. Strangely, however, there is very little in the way of argument put forward in defence of this stance.

[56] I use this analysis as an example because Reutlinger's (2016, 2017, 2018) account stands as an exception to the rule that 'precise formulations of [explanatory monism] are few and far between' (Khalifa *et al*, 2018:2).

The principal motivation for this position appeals to Woodward's (2003) claim that possible interventions serve to illuminate exclusively causal dependencies. This intuition is frequently deployed within the recent literature. As Lange himself suggests that 'the notion of an intervention is a causal notion and so is not obviously applicable to non-causal explanation' (2019:2).[57] Lina Jansson also argues that interventions cannot help in characterising the asymmetry of noncausal explanation, since 'the solution is given in terms of interventions, [and] these are cashed out in causal terms in Woodward [2003]' (2015:22, fn. 48). Similarly, Juha Saatsi and Mark Pexton argue that although Woodward 'happily welcomes the possibility that the counterfactual aspect of his account may come apart from its causal aspect', this element of Woodward's account 'should not be wedded to a causal manipulationist interpretation of explanatory modal information' (2013:614).[58]

However, while Woodward (2003) does indeed support such an *I*-puritan reading of his interventionist framework, an *I*-liberal interpretation is by no means ruled out. In summarizing the interventionist mantra, he suggests that *any* successful explanation ought to be accompanied by 'a hypothetical or counterfactual experiment that shows us that and how manipulation of the factors mentioned in the explanation… would be a way of manipulating or altering the phenomenon explained' (Woodward, 2003:11). What the discussion of Kim's work suggest, is that interventions do not exclusively demarcate causal relationships.

Woodward (2003) holds that what distinguishes explanatory counterfactuals from non-explanatory counterfactuals (which highlight *mere* correlations) is that only the former allow for the possibility of manipulation. This is not to say, of course, that all interventionist counterfactuals pick out causal relations *per se,* but instead, that no interventionist

[57] It is important to note, however, that Lange does not support explanatory monism. Rather he suggests that 'the order of explanatory priority is fixed by different considerations in different non-causal explanations' (2019:24).
[58] As Woodward himself highlights, 'Woodward (2003) (tacitly and without explicit discussion) adopted the common philosophical view that causal (and causal explanatory) relationships contrast with relationships of dependence that hold for purely conceptual, logical, or mathematical reasons' (2018:121).

counterfactuals pick out relations of mere correlation.[59] This understanding of interventions is perfectly compatible with the idea that *some* interventionist counterfactuals highlight possible manipulations which are not distinctly causal in nature.

The idea that interventions might stand as a useful distinguishing factor between causal and noncausal explanations can also be traced to Woodward (2003:220-221). While both causal and noncausal patterns of dependence ought to be able to support counterfactuals (which in turn support "what-if-things-had-been-different" questions), in the latter case, according to Woodward, these counterfactuals cannot be interpreted in terms of interventions. It is important to note that this conclusion is reached on the basis of a single example: the dependence of the stability of planetary orbits on the dimensionality of space-time. Woodward argues that 'it seems implausible to interpret such derivations as telling us what will happen under interventions on the dimensionality of space-time' (2003:220).

The exact reason for this implausibility is not mentioned, although the most obvious candidate is that such an intervention is nomologically impossible.[60] However, Woodward (2003) does not consider any interventions of the type discussed in the previous sections. There is nothing nomologically impossible, for example, about the prospect of intervening upon the *Mimosa's* pulvini cells in order to elicit *thigmonasty*. And, as we saw in section 2. intervening upon Socrates' death for the purposes of manipulating Xanthippe's widowhood is, not only possible, but provides an accurate characterization of the explanatory asymmetry which interests Kim (1974:1993).

---

[59] While, as Baumgartner and Gebharter (2016) point out, Woodward (2003:61) suggests that there can be 'no difference in manipulability relations without a causal difference' this claim is *not* entailed by the definition of an intervention discussed in section 3. of this paper. If this position is correct, it is not obviously so.

[60] Interestingly, elsewhere Woodward seems to have little issue with the notion of impossible interventions. He argues, for example, that '[e]ven in purely theoretical contexts, causal claims should be understood as telling us about the results of hypothetical manipulations; it is just that we cannot, at least at present, carry out these manipulations' (2003:37).

Given the frequency with which Woodward (2003) is appealed to in defence of *I*-puritanism, there a quiet irony in his having recently weakened his own commitment to this stance. In 'Some Varieties of Non-Causal Explanation', Woodward briefly discusses cases of noncausal explanations where 'at least some of the variable figuring in the candidate *explanans* are possible targets for manipulation… but the *connection* between these and candidate *explanandum* seems (in some sense) purely mathematical' (2018:130).

However, as Reutlinger *et al* explain, '[i]n the mathematical case, this involves supposing that mathematical facts were different. But on the standard philosophical accounts of mathematics, mathematical truths are necessary', as such intervening in such cases 'would seem to be deeply problematic' (2020:10). Lange (2019) explicitly references the example used by Woodward (2018b) (the traversibility of Königsberg's famous bridges), as requiring interventions which are impossible to perform. This is not to say that there have not been attempts to make sense of the explanatory potential of such impossible interventions, but such accounts are strictly speaking beyond the scope of this paper.[61]

Despite this, there are reasons to think that Woodward (2018b) would be sympathetic to the central aim of this paper: providing the first robust defence of the claim that there are distinctively noncausal explanations which can be characterized in terms of *possible* interventions. Woodward *does* now consider interventions of the type discussed in the previous sections: 'there is an obvious sense in which it is true that by manipulating whether or not Socrates dies, one can alter whether Xantippe is a widow' (2018:121).[62] Further suggesting that such explanations are more naturally described by locutions such as "brings

[61] See e.g., Baron *et al* (2017); Baron *et al* (2020); Reutlinger *et al* (2020); Baron & Colyvan (2021); and Baron (*forthcoming*). I discuss how impossible interventions fit into the landscape of stances on the nature of the distinction between causal and noncausal explanation in more detail in section 7.

[62] It is interesting to note that, while Woodward (2018b) does not reference Kim (1974) as his source for this example, both papers share this unusual spelling of "Xanthippe". Upon advice from an anonymous referee, outside quotations, I have adopted the orthodox spelling throughout the paper.

about by" than "causes" (2018:131). While Woodward does not provide a great deal of detail concerning how he expects such interventions to fit within his earlier interventionist framework, I can see little reason for him to reject the methodology laid out in this paper.

More recently still, Khalifa *et al* (2020) refer to a position very close to our *I*-liberalism, by the (none-too-pithy) title *quasi-interventionist change-relating counterfactual monism* (QCM). They note that the only revision to Woodward's original characterization of an intervention which is required here, is replacing *I* with something like 'X counterfactually depends on *I*' (Khalifa *et al*, 2020:6).[63] This interpretation seems to capture the spirit of "*R*" and Kim's desire to subsume 'as special cases the causal relation and other dependency relations' (1974/1993:27). As Khalifa *et al* suggest, 'if this quasi-interventionist approach captured every kind of explanation, causes would just be a limiting case' (2020:6).[64]

Interestingly, however, Khalifa *et al* (2020) introduce QCM in the process of arguing that an account of explanatory monism based upon it is untenable, because such a methodology is incapable of dealing with a familiar type of noncausal explanation: constitutive explanation. The principal motivation for this stance being that possible interventions are apparently incapably of distinguish between constitutive explanations and spurious correlations resulting from common explanatory dependence.

While a defence of monism itself is beyond the scope of this paper, in section IV., I argued that a benefit of our own position is that it *can* account for the explanatory asymmetry of constitutive explanation. As such, in the next section, I demonstrate that *I*-liberalism is

---

[63] It is worth noting that this revision is a much less dramatic one than that proposed by Romero (2015) and Baumgartner and Gebharter (2016) in order to accommodate fat-handed interventions.

[64] It is not clear to me that replacing the claim that '*I* causes *X'* with the claim that *'I* counterfactually depends on *X*' is at all necessary for *I*-liberalism. As I mentioned in footnote 12, I am perfectly happy with the relationship between *I* and *X* being a causal one. In the case of Socrates' death and Xanthippe's widowhood, any intervention upon Socrates' death *will* be a causal process. What is important for *I*-liberalism, is that such an intervention establishes a subsequent *noncausal* explanatory dependence between the death of Socrates and Xanthippe's becoming a widow. In this sense, *I*-liberalism appears to require even less modification to Woodward's original manipulationist framework than QCM.

perfectly capable of drawing a distinction between genuine noncausal constitutive dependence and spurious correlations arising as a result of a common explanatory source.

VI. Noncausal Interventions and Common Explanatory Dependence

Khalifa *et al* (2020) argue, citing Petri Ylikoski (2013), that while 'causal relata are metaphysically independent entities, constitutive relata 'are not independent existences, so one cannot think of an intervention on the basis that would not also be an intervention on the system' [2013:284]' (2019:7). As a result, they suggest that an intervention upon a system's components or organization (the explanans), would also be a direct intervention upon the explanandum. According to Khalifa *et al* (2020), an intervention in this case would violate Woodward's principle that '[a]ny directed path from *I* to *Y* goes through *X*' (2003:93). This is because such an intervention would apparently act as a "common cause" of both *X* and *Y*.

If true, this would indeed be a troubling result. Difficulties in distinguishing instances where variables are spuriously correlated, owing to a common explanatory dependency, have historically plagued theories of explanation. Indeed, it is a principal motivation for the adoption of interventionist methodology that, where *A* and *B* are correlated, it allows for the distinction to be reliably drawn between scenarios where: *A* causes *B*; *B* causes *A:*or *A* and *B* are both caused by *C*.[65]

[65] Ylikoski's (2013) argument is specifically aimed at Craver's (2007a, 2007b) mutual manipulability account of constitutive explanations. In this context, the argument that constitutive explanation underdetermines explanatory relations in virtue of being unable to distinguish such cases from spurious correlations arising from common causes is, as far as I am concerned, perfectly sound. As is highlighted above, the type of "top-down" interventions which result from Craver's mutual manipulability, *can* be interpreted in such a way that they do appear to be the result of common causes. However, Khalifa *et al* (2020) take this argument further, seemingly suggesting that even the sort of "bottom up" intervention which I have taken to be unproblematic gives rise to this same explanatory confusion. Woodward (2018b) has also suggested (although in vaguer terms) that in abandoning interventions, explanatory monists face something like this problem. However, as we shall see below, by adopting *I*-liberalism, this result can be avoided.

While analogous spurious correlations *do* arise with respect to noncausal dependence, it is clear to us that instances of constitutive explanation are not among them. If Khalifa *et al* were correct in their claim that an interventionist account of noncausal explanation will misdiagnose instances of constitutive explanation as spurious correlation, then the *I*-liberal would not to be able to draw a distinction between these cases. As I shall now demonstrate, however, my analysis is perfectly capable of distinguishing between spurious (noncausal) correlations and constitutive explanations.

As an example of a genuinely spurious noncausal correlation, take the relationship between Xanthippe's widowhood, and the existence of Socrates' Singleton. It seems that there is a necessary inverse correlation here. All else being equal, there are no possible worlds in which Xanthippe is a widow and Singleton Socrates exists, and there are no possible worlds in which Singleton Socrates does not exist and Xanthippe is not a widow. Yet, we would not want to say that Xanthippe's widowhood *depends* upon, or is *explained by*, the nonexistence of Singleton Socrates (or *vice versa*).

Indeed, the correct story here seems rather obvious: the existence of Socrates' Singleton and Xanthippe's widowhood are *both* determined by a single common factor: the existence of Socrates. Where Socrates exists, it is necessarily the case that Single Socrates exists and that Xanthippe is not a widow (and *vice versa*). And, while Socrates' existence explains both Singleton Socrates' existence and Xanthippe's not being widowed, neither of the latter facts explain each other. This case, then, looks like an instance of genuine noncausal common dependence of the type which Khalifa *et al* (2020) intend to highlight.

That the correlation between the existence of Singleton Socrates and Xanthippe's widowhood is spurious can be quite happily cashed out in terms of interventions. In line with *III*, try as one might, it is simply not possible to intervene upon either Xanthippe's

widowhood, or the existence of Singleton Socrates, in order to manipulate the other; any such intervention *must* go through Socrates' existence. This tells us that it is Socrates' existence which is doing all of the determining here, and hence, all of the explanatory work.[66] For the sake of simplicity, let's consider another (more basic) example of constitutive explanation:

Variables:
C: Whether the diamond's constituent carbon atoms are thus-and-so arranged.
E: Whether the diamond is hard.
Structural equations:
$E \rightarrow C$
Assignment:
C=1; E=1

We can now see that this case of constitutive explanation is quite different to the spurious correlation highlighted above, and that this difference can be drawn out in terms of possible interventions. If Khalifa *et al* (2020) are correct, and the arrangement of the diamond's constituent carbon atoms, and its hardness, are wrongly characterised by *I*-liberalism as being jointly depend upon some common factor, it ought to be the case that there are no possible interventions upon either variable that goes through the other.

This line of thought mirrors Kim's in considering Socrates' death and Xanthippe's widowhood as common effects of Socrates' having ingested hemlock. In this case, Kim suggested that the problem with this idea is that the only route from hemlock to widowhood seems to go through death. In the case of Xanthippe's widowhood and the existence of Singleton Socrates, on the other hand, any possible intervention which attempts to manipulate

[66] In 'Supervenience as a Philosophical Concept' (1990/1993) Kim uses a very similar argument to defend the thesis that supervenience, like correlation, is not an explanatory relation. Just as correlation is insufficient to establish an explanatory causal dependence relationship between two variables, Kim argues that supervenience is insufficient to establish an explanatory *noncausal* dependence relationship between two variables. The analogy between supervenience and correlation could well be an illuminating one, especially given a recent resurgence of interest in the explanatory status of supervenience (e.g. Kovacs 2019). Unfortunately, however, we do not have the space to explore this connection here.

the former using the latter, is *mediated* by the existence of Socrates. If Khalifa *et al* (2020) are correct, then cases of constitutive explanation ought to look more like the relationship between Xanthippe's widowhood and the existence of Singleton Socrates, that the relationship between Xanthippe's widowhood and Socrates' death. However, this is not what we find.

Just as in the case of Socrates' death and Xanthippe's widowhood, in line with *III*, any possible intervention upon the hardness of a cut diamond *must* go through its constituent carbon atoms.[67] It is simply impossible to alter the hardness of a diamond without altering the arrangement of its constituent carbon atoms. This is just what we saw in relation to the *Mimosa's* in section 4., any intervention upon the plant's nastic movement must go through its pulvini cells. Were these cases of constitutive explanation the result of a common explanatory source, there would be some third variable doing the actual explanatory work. But this is not so. Thus, an account of noncausal explanation which makes use of interventions *can,* in fact, draw an illuminating distinction between constitutive explanation and spuriously correlated variables resulting from a common explanatory dependence relation.

VII. Remarks on Taxonomy

In this paper, I have attempted to mount the first sustained defence of *I*-liberalism, against its more popular rival, *I*-puritanism. In section II., I introduced Kim's argument against causal imperialism, the claim that all explanation is essentially causal in nature. Causal imperialism is false, according to Kim, because there are clear cases of asymmetric explanatory counterfactual conditionals which are not the result of causal relationships. Crucially, what

[67] Such an intervention could be performed by subjecting our diamond to around 10 million times ordinary atmospheric pressure, for example (see e.g. Knudson *et al* 2008)

these causal and noncausal counterfactuals share, is a close connection to the notion of "bringing about": where *A* depends upon *B*, we can bring about *A* by first bringing about *B*.

In section III., introduced Woodward's interventionist analysis of causal explanation and argued that Kim's intuition regarding the asymmetry of noncausal explanation and the "bringing about" relation can be neatly characterized in terms of interventionist counterfactuals and structural equation models. What this shows, I argued, is that the notion of a possible intervention does not line up with the distinction between causal and noncausal explanation.

Having characterized the core claim of *I*-liberalism, that possible interventions can characterize certain noncausal explanations, in section IV., I moved on to apply this methodology to an archetypal instance of such explanation: a constitutive mechanism. I argued that *I*-liberalism avoids two central issues facing Craver's own mutual manipulability analysis: the need for impossible interventions; and the 'embarrassing' explanatory symmetry which results. I also observed that constitutive explanation has been a key breeding ground of *I*-puritan sentiments, with the likes of Leuridan (2012) having dismissed Craver's account *tout court*, simply because it attempts to characterise noncausal explanation in terms of interventions.

As I also highlighted, however, this is not the only area where we find such dismissive attitudes. More recently still, debate surrounding the viability of explanatory monism has invoked similar responses. In section V., I noted that the *I*-puritan stance has led to difficulty in characterising the obvious asymmetry of noncausal explanation. Despite this, I showed that motivation for *I*-puritanism among those involved in characterizing *both* monistic and constitutive explanation typically appeals to Woodward's (2003) own defence of this position. In response, I argued that, although Woodward (2003) supports *I*-puritanism,

nothing in his account mandates this interpretation. Indeed, Woodward's (2018) most recent foray into the topic of noncausal explanation appears to roll back this stance and take a significant step towards *I*-liberalism.

In section VI., I discussed an argument, recently put forward by Khalifa *et al* (2020), which suggests that an interventionist account of noncausal explanation would be unable to distinguish between genuinely explanatory relationships and spurious correlations resulting from common causes. On the contrary, I argued the *I*-liberal is perfectly capable of drawing a distinction, in interventionist terms, between genuine explanatory noncausal dependence relations on the one hand, and unexplanatory spurious correlations arising from a common dependence relation, on the other. In what remains of this final section, I wish to provide something of a taxonomy of the various positions which have been discussed in this paper and highlight exactly what our own position commits us to.

While the idea that all explanation is causal explanation has largely fallen out of favour, it is worth noting that causal imperialism is fully compatible with *I*-puritanism as I have described it. One might think that all explanation is causal explanation *and* that all such explanation can be characterized in interventionist terms. Let's call this position "strong causal puritanism". Of course, causal imperialists need not necessarily think that all causal explanation is characterizable in interventionist terms, just that wherever one *can* intervene upon X with respect to Y in such a way that changes the value of Y, X causes Y. This leaves open the possibility that, while all explanations are causal explanations, some such explanations defy interventionist analysis. Let's call this position "weak causal puritanism".

The former of these positions, strong causal puritanism, implies explanatory monism. If all explanation is causal explanation, and all causal explanation is characterizable in terms of interventions, then we have a single unifying (interventionist) account of explanation.

Conversely, the latter position, weak causal puritanism, implies explanatory pluralism. Even if all explanation is causal in nature, if some such explanations are not characterizable in interventionist terms, then explanatory monism must be false. As far as I am aware, however, no one has committed themselves to either of these positions in the recent literature.

Those of an *I*-puritan persuasion are likely to reject causal imperialism on the grounds that noncausal explanations are possible. Monist *I*-puritans will argue that although causal and noncausal explanation can be captured using a single unifying thesis, said thesis will not reference interventions.[68] Pluralist *I*-puritans, on the other hand, will accept that noncausal explanation is possible, but reject the idea that both types of explanation can be captured in a single unifying thesis, as Lange puts it: 'the order of explanatory priority is fixed by different considerations in different non-causal explanations' (2019:24).

I-liberalism is obviously incompatible with causal imperialism on two fronts. First, *I*-liberalism presupposes that noncausal explanations are possible, and second, it argues (*contra* *I*-puritanism) that (at least) some noncausal explanations can be characterized in terms of the possibility of intervening upon the explanans variable. Given that causal imperialism consists in the denial the first of these claims, the second is clearly a nonstarter. Although, as with causal imperialism and *I*-puritanism, *I*-liberalism is noncommittal with respect to explanatory monism vs explanatory pluralism.

A monist *I*-liberal methodology would imply, not only that possible interventions can characterize *certain* instances of noncausal explanation, but that possible interventions are capable of characterizing *all* instances of explanation. The most obvious reason for rejecting this hard-line *I*-liberalism, are scenarios of the sort highlighted by Woodward (2003) as a

[68] As we have seen, such theories typically take counterfactuals to be the central unifying feature of causal and noncausal explanation, although Khalifa *et al* (2018) have recently argued for a monistic *inferential* account of explanation.

reason for rejecting *I*-liberalism all together. These involve counterfactual conditionals whose antecedents hold with necessity. Intervening upon a variable which holds its value of necessity (like the dimensionality of space-time) will, of course, be *at least* nomologically (although often also logically and/or metaphysically) impossible.

Such counterpossible counterfactuals have received a great deal of attention within the philosophy of science literature.[69] Indeed, one of the most interesting recent developments has been the idea that we can cash out the explanatory potential of counterpossibles in terms of interventions, even though such interventions are impossible. For example, Baron *et al* (2017), Reutlinger *et al* (2020) and Baron *et al* (2020) have argued that mathematical explanations can be understood in terms of interventions which are, strictly speaking, (metaphysically) impossible to perform.[70]

It is important to note, however, that this position is compatible with *I*-puritanism. Indeed, a pluralist *I*-puritan might well accept that certain noncausal explanations are characterizable in terms of impossible interventions, but nonetheless maintain their central thesis, that wherever it is *possible* to intervene upon X in such a way that changes the value of Y, then X causes Y. In this sense, there would remain hope for the *I*-puritan that a neat dividing line can be drawn between causal and noncausal explanation in terms of *possible/impossible* interventions. It is this idea which I have attempted to undermine.[71]

---

[69] See e.g. Handfield (2004); Brogaard & Salerno (2013); Berto and Jago (2013); Priest (2016); Tan (2019); Kimpton-Nye (2020); Wilson (2020); and Hicks (*ms*).

[70] Schaffer (2016, 2017) and Wilson (2018a, 2018b) have argued that metaphysical explanations similarly require the analysis of impossible interventions, and Baron & Colyvan (2021) and Baron (2022) have argued that certain ontological and logical explanations (respectively) are in the same boat. In all these cases, adopting such a stance requires a commitment to the non-triviality of counterpossibles and the abandoning of the traditional semantics for counterfactuals (see e.g., Stalnaker 1968; Lewis 1973). However, as Schaffer (2016) highlights, there are already good reasons for thinking that counterpossible scenarios require non-trivial evaluation (see, e.g. Restall 1997; Goodman 2004; Priest 2005; and Jago 2015).

[71] My thanks go to an anonymous reviewer at *Erkenntnis* for pressing this important point.

This paper has sought to mount a thorough defence of only a weak form of *pluralist I*-liberalism, which suggests that possible interventions do not demarcate causal relations *alone*. In other words, my pluralist *I*-liberal thesis suggest merely that *some* noncausal explanations, of the types highlighted herein, are characterizable in terms of interventions which are possible to perform. This weaker *I*-liberal thesis is, of course, sufficient to prove the falsity of *I*-puritanism. If even a single noncausal explanation permits of a manipulationist analysis, then it cannot be the case that possible interventions serve only to characterize causal relations.

Chapter 3

Plumbing Metaphysical Explanatory Depth

One of the core strengths of interventionist analyses of causal explanation is the ability to provide a convincing account of explanatory depth; the sense in which explanations come in degrees. In contrast, however, recent attempts to provide interventionist treatments of metaphysical explanation have left this notion of depth almost entirely unexplored. In this paper I shall attempt to rectify this oversight by motivating an interventionist analysis of metaphysical explanatory depth (MED), in terms of the range of interventions under which a metaphysically explanatory generalization remains invariant. After elucidating the notion through a toy-example, I demonstrate the important work which MED can perform in characterizing debate within contemporary metaphysics. Focusing upon rival approaches to explaining the identity and distinctness of concrete objects, I argue that the progress achieved in this debate can be characterized in terms of increasing explanatory depth. Having made an initial case for the utility of MED, I then turn this analysis to the metaphysics of explanation itself. By adopting an interventionist framework with respect to MED, I will show that we can assess the depth of competing theories of explanation. This application has two interesting results: first, it suggests that an interventionist analysis of explanation provides deeper explanations of the connection between explanans and explanandum than rival accounts; and second, it suggests that explanations provided by interventionism become deeper still, if one accepts that this methodology ranges over metaphysical, as well as causal, instances.

I. Introduction

The last decade has seen growing interest in the prospect of modifying interventionist analyses of causal explanation in order to characterize *non*causal explanations within metaphysics.[72] One area where interventionism traditionally shines, is in accounting for the

[72] See e.g., Schaffer (2016, 2017); Reutlinger (2017); Wilson (2018a, 2018b); Miller & Norton (2022a, 2022b)

sense in which explanation comes in degrees, or as Hitchcock & Woodward (2003b) put it, the *depth* of an explanation. However, an analogous notion of depth is entirely absent from parallel debate concerning metaphysical explanation. In this paper, I seek to rectify this oversight. I argue that *metaphysical* explanatory depth ought to be analogously understood in terms of the range of interventions under which an explanatory generalization remains invariant.

I shall proceed as follows. In the next section I provide a detailed account of the interventionist analysis of explanatory depth, and the benefits of this approach over rival inferential analyses, which define depth in terms of *scope*. In section III., I apply these contrasting notions of depth to a toy-example of metaphysical explanation and show that the benefits of an interventionist analysis carry across to explanations within metaphysics.

With the preliminaries out of the way, I put this account to work, applying the interventionist understanding of explanatory depth to two case-studies of metaphysical explanation *in the wild* (as it were). As these case-studies show, an interventionist account of metaphysical explanatory depth provides us with novel tools with which we can characterize live debate within contemporary metaphysics.

In section IV., I focus upon a recent approach to explaining the identity and distinctness of concrete objects, the *quantitative properties proposal*, put forward by Erica Shumener (2020). I argue that Shumener's thesis can be seen as progressive with respect to prior proposals in terms of both *qualitative properties* (Black 1952, Rocca 2005) and *weak discernibility* (Saunders 2006), precisely because it provides greater metaphysical explanatory depth.

In section V., I argue that this same methodology can be applied to debate surrounding the nature of explanation *itself*. We can, I suggest, think of analyses of explanation as

providing metaphysical explanations for the connection between explanans and explanandum. Here, competing accounts of explanation provide contrasting meta-explanatory generalizations of the following form: "*for any x and y, if …, then x explains y".* Through the adoption of an interventionist framework with respect to metaphysical explanation, I will show that we can assess the depth of rival accounts of explanation.

This novel application of the notion of explanatory depth appears to have two interesting results. First, it accurately characterizes the interventionist analysis of explanations as providing greater depth than both inferential (Hempel 1965; Kitcher 1981) and conserved quantity (Salmon, 1984, 1989; Dowe 1992, 2000) accounts. And second, it suggests that interventionism can be shown to provide still deeper explanations of the connection between explanans and explanandum, if one additionally accepts that this analysis rangers over metaphysical, as well as causal, instances.

In the final section I discuss three methodological issues which arise as a result of this interventionist analysis of metaphysical explanatory depth. The first concerns how we are to make sense of the distinction between same- and other-object counterfactuals in the context of identity and distinctness; the second, raises the issue of how we are to understand the role of interventions when it comes to metaphysical explanations more generally; and the third asks whether shallow metaphysical explanations are actually explanations at all. In each case, I argue that the interventionist can provide a satisfying response.

II. Explanatory Depth: Scope vs Invariance

For much of the 20th century, debate surrounding the nature of explanation was dominated by broadly inferential analyses. On Hempel's deductive-nomological model, for example, explanation is centrally concerned with *expectation*. An explanation, on this account,

involves a set of auxiliary statements $C_1 \ldots C_n$, asserting the occurrence of events and a law, L, from which we can logically deduce a statement asserting the occurrence of the event to be explained E. Such arguments are intended to show that 'given the particular circumstances and the law in question, the occurrence of the phenomenon *was to be expected*' (Hempel, 1965:337).

However, Kitcher (1981, 1989) argues that behind Hempel's "official view" of explanation as nomic expectation, there lies an "unofficial" view of explanation as *unification*.[73] As Kitcher interprets this unofficial view, laws are explanatory in virtue of 'showing us how to derive descriptions of many phenomena, using the same patterns of derivation again and again' (1989:432).[74] Kitcher's key insight here, is that in order for a generalization to be explanatory (as opposed to a merely accidental) it ought to apply to a *range of different cases*.[75]

This idea provides us with a natural way of understanding explanatory depth: the wider the range of cases to which an explanation applies, the deeper the explanation. Where DN and unificationist accounts diverge from an interventionist analysis, however, is with respect to *which* cases are taken to be salient in determining depth. On both the DN and unificationist accounts, this range is understood in terms of *scope*: the set of objects or systems that fall under the antecedent of a given law.

Such laws, according to Hempel (1965) and Kitcher (1981), take the form of universally quantified conditional claims like "All *A*s are *B*s". As Hitchcock & Woodward

---

[73] Also see Friedman (1974).

[74] As Hempel himself suggests, explanation is achieved 'by a systematic unification, by exhibiting the phenomenon as manifestations of common, underlying structures and processes' (1966:83).

[75] Brad Weslake highlights that this approach appears to have been, at least tacitly, endorsed by Hempel himself: 'the most natural way to incorporate an account of explanatory depth into the DN account is itself suggested by Hempel [1959:302-303], who mentions in passing the predictive possibilities afforded by laws in situations other than the one under consideration' (2010:276).

(2003a) highlight, such conditionals support "other-object" counterfactuals of the following form: 'if some object *o** that is different from *o* and does not possess property *A* were to be an *A*, then it would be a *B*' (2003:a19).[76] In this sense, the wider the range of *other objects* to which an explanatory generalization applies, the more unifying it will be, and thus the deeper the explanation it will provide.

There is, however, a fundamental problem facing a scope-based account of explanatory depth and the traditional understanding of laws as the 'universal premisses that occur in explanatory patterns' (Kitcher, 1989:447). On this interpretation of laws, it becomes difficult to see quite how one generalization could provide for deeper explanations than any other. As Hitchcock & Woodward explain, 'a true generalization is either a universal law, in which case it can facilitate explanations, or it is accidental, in which case it cannot – there are no other options' (2003b:183).[77]

Fortunately, the interventionist analysis avoids this problem because it does not require generalizations to be universal or exceptionless in order to be explanatory. Rather, according to Hitchcock & Woodward, 'it is only if a generalization is invariant under testing interventions that it conveys information about how one variable depends on another' (2003a:19). For a generalization to be invariant under testing interventions, it must 'describe a relationship which holds for certain *hypothetical* values of *X* and *Y possessed by the very object o…* where the value of *X* is changed by an intervention' (Hitchcock & Woodward, 2003a:20).

By interpreting the explanatory character of generalizations in these terms, we see that '[a]mong those generalizations that are invariant, some will be more invariant than others,

[76] Also see Woodward (2003:279-288).
[77] Hitchcock & Woodward are not the first to mount this objection. Jaegwon Kim, for example, similarly argues that '[u]nderstanding and explanatoriness are matters of degree… The DN model makes explanation an all-or-nothing affair' (1994:59).

and they will correspondingly provide deeper explanations' (Hitchcock & Woodward, 2003b:183-184). As such, the interventionist account of explanatory depth is not concerned with the range of *other objects* for which a given generalization holds, but rather, the range of changes to the *actual object being explained* under which a given generalization remains invariant. In other words, the greater the range of *same-object* counterfactual scenarios under which a generalization holds, the deeper the explanation.

In the introduction, I noted a recent trend which seeks to characterize metaphysical explanation, in interventionist terms, as analogous to causal explanation. Despite what Kment calls the 'far-reaching and structural analogy' (2014:5) between metaphysical and causal explanation, the role of explanatory depth within metaphysics remains, as yet, unexplored. One initial reason for thinking that this area deserves further investigation is that recent interventionist theories of metaphysical explanation already provide us with all of the methodological tools necessary to characterise a notion of explanatory depth operative in the metaphysical domain.

Schaffer, for example, has recently echoed Hitchcock & Woodward (2003a, 2003b) in arguing that the generalizations which govern metaphysical explanation needn't be fundamental or exceptionless. According to Schaffer, to qualify as explanatory a "metaphysical law" must merely support an appropriate pattern of counterfactuals, as 'it is through counterfactual-supporting generalizations that one can calculate the impact of potential interventions' (2017:306). So, just as Hitchcock & Woodward argue with respect to causal explanation, Schaffer (2017) argues that *metaphysical* explanation requires generalizations which will remain invariant under testing interventions.

From here, it is a small step to arrive at an analogous notion of *metaphysical* explanatory depth. For one metaphysical explanation to be deeper than another, is for the

corresponding generalization to be invariant under a wider range of same-object counterfactuals. In the next section, I shall apply this interventionist analysis of explanatory depth to a toy-example of metaphysical explanation, and demonstrate that the benefits of this analysis, over a scope-based account, appear to carry across into the metaphysical domain.

III. A Toy-Example

Consider, as a preliminary example, the fact that Jeff Bezos is a billionaire.[78] One might well ask why this is the case, what explains Bezos's being billionaire? A natural explanation of this fact could simply highlight his net worth: $182B. Another might cite the fact that his net worth is ≥$1B.

(a) *For any person x, if x has a net worth of $182B, then x is a billionaire.*

(b) *For any person x, if x has a net worth of ≥$1B, then x is a billionaire.*

At least on the face of it both (a) and (b) appear to qualify as candidate generalizations figuring in an explanation of Bezos's being a billionaire. However, it is my contention that the second of these generalizations provides the deeper explanation, and that the

[78] One might query whether this example is actually an instance of *metaphysical* explanation at all. While some, like Dasgupta (2017) take metaphysical explanation to be synonymous with *constitutive* explanation (a category into which this example appears to fall), it is not clear that metaphysics has a monopoly on such explanations. Within the literature on noncausal explanation more generally, constitutive explanations are typically taken to be a form of *mechanistic* explanation and are distinguished from *etiological* mechanistic explanation. In the latter case, some mechanism explains a phenomenon for which is *causally* responsible, whereas in the former, a phenomenon is explained by the underlying mechanism which *constitutes* it (e.g., Machamer, Darden & Craver (2000); Darden & Craver (2002); Craver (2007a, 2007b); Glennan (2010, 2017). For what it's worth, I take *all* explanation to be an instance of *either* metaphysical or causal explanation although, as Wilson (2020) highlights, even this distinction can be difficult to parse. While the above explanation is clearly not causal in nature, for those unwilling to accept it as an instance of metaphysical explanation, I would hope that it at least serves a useful purpose in helping to draw out the difference between same- and other-object counterfactuals in relation to explanatory depth. In sections IV. and V., I will consider two more complex examples of explanation which are less controversially metaphysical in character.

interventionist analysis of explanatory depth is better able to make sense of this claim, than a scope-based analysis.

As we saw in the previous section, on a scope-based account, explanatory depth is determined by the range of *other objects* to which an explanation applies. Here, in assessing the depth of the above explanations, we thus consider the range of other objects under which (a) and (b) hold. The greater the range of objects which can be subsumed under each generalization, the deeper the explanation. This gives rise to a problem, however. It seems that, in terms of scope, these candidate generalizations are equally explanatory.

Take Boris Johnson, for example. As far as it is possible to tell, Johnson is not a billionaire; his net worth appears to be somewhere between $2M and $4M, depending on who you ask. Despite this, *were* Johnson's net worth $182B, then he *would* be a billionaire. In this case, the generalization specified by (a) would still hold. In fact, (a) will continue to hold for any counterfactual scenario in which we replace Bezos with another person. *Anyone* would be a billionaire if their net worth were $182B. The problem is that the same is true of (b). Were Johnson's, net worth ≥$1B, then he would be a billionaire. Were *my* net worth ≥$1B then I would be a billionaire. There is no "other object" with which we can replace Bezos which will allows us to draw a distinction between (a) and (b) in terms of scope.[79]

By adopting an interventionist analysis of explanatory depth, however, we can provide a satisfying characterization of why (b) provides a deeper explanation of Bezos's being a

[79] One might worry that the generalizations which are specified in (a) and (b) are not *universal*. Since it was specified above that inferential accounts takes laws to be universal generalizations, it could be argued that this example is unfairly tipped in the interventionist's favour. I would point out, however, that these generalizations can be easily rephrased in order to apply universally: "(a*) *For any x, if x has a net worth of $182B, then x is a billionaire".* However, this formulation unnecessarily complicates matters, opening up the possibility of counterfactual scenarios which become impossible to interpret with any clarity. For example, what if *x* is an inanimate object, or an animal? Would a rock, a record player or a rhinoceros be a billionaire if they had a net worth of $182B? Hitchcock and Woodward (2003a) argue that it is mysterious what such counterfactual scenarios are even supposed to *mean*. More importantly, such scenarios still do not allow us to draw a distinction, in terms of depth, between (a) and (b). Where one generalization breaks down, so will the other.

billionaire than (a). On this account, rather than observing the range of other objects under which (a) and (b) hold, we must assess changes to the *actual* object (or system) in question. What we are chiefly looking for, in justifying the intuition that (b) provides a deeper explanation than (a), are hypothetical counterfactual scenarios in which (a) is violated but (b) remains invariant. Such cases are not difficult to come up with.

Imagine, for example, that all Amazon warehouse staff unionize and, with investors concerned about the impact of good working conditions on profitability, Amazon's share price falls, leaving Bezos with a net worth of $150B. In this scenario, it appears that (a) no longer explains why Bezos is a billionaire; it tells us nothing about counterfactual scenarios in which a person's net worth is anything other than $182B. So, while Bezos is obviously still a billionaire in this scenario, (a) cannot be used to explain this fact. On the other hand, (b) will remain invariant under counterfactual scenarios which see interventions upon Bezos's net worth, so long as his net worth remains ≥$1B. As such (b) holds under a wider range of counterfactual scenarios than (a) and can thus be considered a deeper explanation of Bezos's being a billionaire.

To my mind, examples such as this do a good job of motivating the idea that there is a notion of explanatory depth, analogous to that identified by Hitchcock & Woodward (2003b) operative in the metaphysical domain. What's more, as was noted above, several well fleshed-out interventionist theories of metaphysical explanation are already on the table, theories which would require little alteration in order to make use of this notion.[80]

However, the recent literature is also replete with attempts to characterize metaphysical explanation in analogy with inferential accounts of scientific explanation. Dasgupta (2014, 2017), for example has recently motivated an 'analogue of the "DN model"' upon which 'the

[80] See e.g., Schaffer (2016, 2017); Reutlinger (2017); Wilson (2018a, 2018b); Miller & Norton (2022a, 2022b).

thing to be explained follows from the explainer together with something like a "metaphysical law"' (2017:80).[81] Similarly, Kovacs (2020) follows Kitcher (1981, 1989) in claiming that metaphysical explanation results from seeing 'how a large number of phenomena are the consequences of a small number of basic facts, from which they can be derived using relatively few and similar patterns of derivation' (2020:1673).[82]

Insofar as any adequate analysis of explanation ought to be able to account for the sense in which explanation comes in degrees, an interventionist methodology clearly outshines its inferential counterparts. This is not to say that those who provide inferential accounts of metaphysical explanation explicitly endorse a scope-based account of depth; in fact, any reference to depth appears to be entirely absent from this literature. Nonetheless, until an alternative to a scope-based account is put forward, the superiority of an interventionist analysis of explanatory depth, over its inferential rivals, represents a key motivation for its adoption in cases of both causal and metaphysical explanation.

Admittedly, the examples discussed thus far may strike the reader as an uninteresting application of the notion at issue. As a result, my analysis of explanatory depth within metaphysics might seem like a relatively modest benefit of an interventionist methodology. In response to this thought, I now want to turn to two much more interesting cases of metaphysical explanatory depth *in the wild*. In the next section, I shall demonstrate that debate within a lively area of contemporary metaphysics research, concerning the identity

---

[81] Also see Wilsch (2015, 2016). Reutlinger similarly notes that 'a friendly amendment of the covering law account may even allow for… metaphysical covering laws (such as general statements about one kind of facts grounding another kind of facts)' (2017:241). Unlike Dasgupta (2017), however, Reutlinger quickly dismisses this account as a viable option, owing to 'well-known problems of the covering-law account' (2017:241). See Salmon (1989) for a survey of these problems.

[82] Baron & Norton (2021) also defend a unificationist account of metaphysical explanation which utilizes Friedman (1974) and Kitcher's (1981, 1989) analogous models of scientific explanation. Of course, I would not want to claim that unification is an irrelevant factor in assessing competing explanation. The important point, however, is that unification is better understood in interventionist terms as being a result of invariance under testing interventions; the wider this range, the more unifying the explanation. See Hitchcock & Woodward (2003b:192-194) and Woodward (2003:356-371) for more on the role of interventions in characterising the unificatory dynamic of explanation.

and distinctness of concrete objects, can be interpreted through the framework provided above: in terms of *progression* through metaphysical explanations of increasing depth. In section 5, I then turn this methodology to debate surrounding the nature of explanation itself.

IV. Explaining Identity and Distinctness

In a recent paper, Shumener (2020) seeks to provide a novel metaphysical explanation of the identity and distinctness of concrete objects. Before presenting her own account, Shumener discusses two others. The first, the *qualitative properties proposal* suggests that identity facts of the form [$x = y$] are explained by the fact that x and y share all of their qualitative properties.[83] The second, the *weak discernibility proposal,* suggests that such identity facts are explained by the fact that x and y stand in only *reflexive* relations to one another.[84] According to Shumener's own *quantitative properties proposal* on the other hand, identity facts are explained by the fact that x and y stand in *quantitative* relations to each other *non-fundamentally*.[85]

Unfortunately, a critical analysis of Shumener's argument is beyond the scope of this paper. My interest in this account is, rather, in the dialectical trajectory of the debate; an intuitive interpretation of which can be given in interventionist terms, as progression through metaphysical explanations of increasing depth. First, note that each of the above explanations of identity, and the converse explanations of distinctness, constitute something approaching

[83] A precise definition of a qualitative feature is difficult to come by, although, Shumener suggests that: 'qualitative features are those that do not involve the identity relation or involve specific relations. So, for example, *5km mass, adjacent to, same colors as* are qualitative features' (2020:2079). By specifying that only qualitative properties ground identity, we avoid the trivial possibility that the property of identity *itself* grounds facts about what is identical to what. See Black (1952:11).

[84] See Saunders (2006).

[85] Quantitative properties or features are, like qualitative features, difficult to formally define. However, of specific interest to us here, are those taken from our physical theories, having a determinate-determinable structure and admitting of degrees: 'examples of determinate quantitative relations include: *five meters away from, twice as massive as, opposite charge as* and the like' (Shumener, 2020:2084).

an explanatory generalization. We can reinterpret each to give a generalization which fits better with the example already discussed:

(c) **The qualitative properties proposal**: *for any objects x and y, if x and y share all of their qualitative features, then x is identical to y; and if x has some qualitative feature that y lacks, then x and y are distinct.*

(d) **The weak discernibility proposal**: *for any objects x and y, if x and y only stand in reflexive relations to one another, then x is identical to y; and if x stands in an irreflexive relation to y, then x and y are distinct.*

(e) **The quantitative properties proposal**: *for any objects x and y, if for any quantitative relation R that x and y stand in, x and y stand in R to one another non-fundamentally, then x is identical to y; and if for at least one quantitative relation Q that x and y stand in, x and y stand in Q to one another fundamentally, then x and y are distinct.*

By adopting the interventionist analysis of depth, we can provide a satisfying characterisation of why (e) provides a deeper metaphysical explanation of identity facts than (d) and, in turn, why (d) provides a deeper metaphysical explanation of identity facts than (c). Once again, what we are chiefly looking for, in justifying the claim that the *quantitative properties proposal* provides the deepest explanation of identity and distinctness, are hypothetical counterfactual scenarios in which one generalization is violated while another remains invariant.

In order to show how the *weak discernibility proposal* provides greater metaphysical explanatory depth than the *qualitative properties proposal*, we require a counterfactual scenario in which it can be shown that qualitatively identical objects can be numerically distinct. Such a scenario will be one under which the generalisation specified in (c) is *not* invariant. This scenario must, of course, be one in which it remains the case that the objects involved stand only in irreflexive relations to one another, thus securing the invariance of (d). As luck would have it, Max Black has popularized a case which fits the bill.

Black (1952) imagines a possible world containing only two spatially separated objects, *A* and *B,* which possess *different* qualitative properties. For the sake of argument, let's assume that *A* is spherical, and *B* is cuboid. Now, according to the generalization specified by (c), if *x* and *y* share all of their qualitative properties, then they are identical; and if *x* has some qualitative feature which *y* lacks, then *x* and *y* are distinct. Since *A* possesses the property of 'being spherical', which *B* lacks, (c) appears to be a candidate explanation of the distinctness of *A* and *B*.

In order to assess the *depth* of this explanation, however, we need to assess the range of same-object counterfactuals under which the relevant generalization will remain invariant. So, now imagine that we intervene upon *A* or *B* (or both), resulting in a counterfactual scenario in which they now share all of their qualitative properties (they are the same size, shape, mass etc), yet remain spatially separated. They are, in other words, indistinguishable in terms of their qualitative properties.

In such a scenario, (c) is no longer explanatory. *A* does not possesses any qualitative properties which *B* lacks, and yet *A* and *B* are not identical (since they remain spatially separated). As Shumener argues, '[t]he Qualitative Properties Proposal cannot account for the distinctness of the spheres because there is no qualitative feature that one sphere has that the other lacks' (2020:2080). Here then, we have a testing intervention under which (c) is not invariant. Since *x* and *y* share all of their qualitative features and yet they are not identical.

Conversely, as Saunders (2006) has argued, (d) remains invariant in such cases. This is because, while all of the *qualitative* relations in which *x* and *y* stand to one another are reflexive, 'the spheres stand in irreflexive relations like *five meters away from* to one another' (Shumener, 2020:2080). Thus, the explanatory generalization specified by (d) remains invariant under a testing intervention which (c) is not. As a result, on Hitchcock &

Woodward's (2003b) understanding of explanatory depth, the *weak discernibility proposal* appears to provide a deeper explanation of the identity and distinctness of objects than the *qualitative properties proposal*.[86]

Just as before, in arguing that Shumener's (2020) *quantitative properties proposal* provides a deeper metaphysical explanation of identity and distinctness than the *weak discernibility proposal*, we need to locate a counterfactual scenario in which (e) remains invariant while (d) does not. Shumener provides us with another example which fits the bill: "Deluxe Max Black cases", involving 'metaphysically possible scenarios in which there are co-located, qualitatively indiscernible objects' (2020:2081).

So, if we return to our qualitatively indiscernible objects *A* and *B*, and now imagine a counterfactual scenario which involves them sharing the exact same spatiotemporal location, then (d) will no longer be explanatory. Since *A* and *B* will now be co-located, we cannot differentiate between them based upon the irreflexive spatiotemporal relations which they stand in to one another.[87] As such, on an interventionist account of metaphysical explanation, (d) is not invariant under testing interventions resulting in Deluxe Max Black scenarios.

Shumener's (2020) own *quantitative properties proposal* relies upon a novel distinction between an object having properties fundamentally and non-fundamentally, which differs from a property *itself* being either fundamental or non-fundamental. While this

[86] The reader may well question whether (c) is actually an *explanatory* generalization at all. Doesn't showing that (d) provides greater explanatory depth than (c) also involve showing that (c) is, in fact, false? Assuming that explanation is factive, this means that the qualitative properties proposal doesn't merely provide a shallower explanation of the identity and distinctness of objects, it provides no explanation at all. As a result, one might think that my account of explanatory depth differs from Hitchcock & Woodward's in an important respect. I am entirely sympathetic to this concern and, in section 6, I explain that this situation is exactly what we ought to expect from analogous interventionist analyses of explanation within methodologically divergent domains.

[87] As Shumener points out, certain "symmetrized states" of quantum particles, appear to be cases of the Deluxe Max Black variety, containing multiple subatomic particles 'which are not distinguished on the basis of their positions' (2020:2082). See, e.g. French (1989).

distinction is of crucial importance to Shumener's argument in favour of (e), a detailed discussion of the intricacies of this approach is beyond the scope of this paper.

What is important, for our purposes, is that Shumener argues that the *quantitative properties proposal* can correctly characterize the objects involved in both Max Black and Deluxe Max Black cases as being distinct; despite sharing all of their qualitative properties in the former case, and additionally not standing in irreflexive spatiotemporal relations to one another in the latter. In both scenarios, the objects can be distinguished on the basis that they stand in *quantitative* relations to one another *fundamentally.* Were they identical, then they would stand in such relations to one another *non-fundamentally*.

To briefly recap, (c) is not invariant under interventions resulting in either Max Black and Deluxe Max Black cases. The *qualitative properties proposal* is thus invariant under the narrowest range of testing interventions, providing the shallowest of the three candidate metaphysical explanations highlighted by Shumener (2020). Because (e) holds with respect to both Max Black and Deluxe Max Black cases, Shumener's *quantitative properties proposal* is invariant under the widest range of same-object counterfactual scenarios and, as a result, can be seen to provide the deepest metaphysical explanation of the identity and distinctness of concrete objects. And finally, since (d) is invariant under counterfactual scenarios resulting Max Black cases, but not Deluxe Max Black cases, the *weak discernibility proposal* occupies the middle-ground in terms of metaphysical explanatory depth. (d) has a wider range of invariance than (c), but a narrower range of invariance than (e).

I believe that this application of metaphysical explanatory depth naturally gestures towards another, even more interesting application. This second application becomes apparent once we notice that debate surrounding the nature of explanation is *itself* metaphysical in character. As I shall argue in the next section, what this means is that by

adopting an interventionist account of metaphysical explanation, along with the account of metaphysical explanatory depth put forward in this paper, we can show that the explanations provided by interventionism are deeper than those supplied by rival accounts.

What is more, I will argue that in the very act of adopting interventionism with respect to metaphysical explanation, the range of counterfactual scenarios under which such meta-explanatory generalizations remain invariant, widens. Which is to say, interventionism provides deeper explanations of the connection between explanans and explanandum, if one additionally accepts that this interventionist analysis ranges over metaphysical, as well as causal, instances.

V. Depth in The Metaphysics of Explanation

So far, we have encountered two opposing positions upon the nature of explanation. On the one hand, DN and unificationist accounts suggests that explanation is *inference*. On the other, the interventionist account suggests that explanation is intimately connected to *manipulation*. However, a third theory, popularized by Salmon (1984, 1989) and Dowe (1992, 2000), suggests that 'causal processes, causal interactions and causal laws provide the mechanisms by which the world works' (Salmon, 1984:132). On this, *conserved quantity* (CQ) account, an explanation of *y* in terms of *x* is the result of a causal interaction resulting in the exchange of a conserved quantity (e.g., energy, momentum, charge etc).

In the previous section we saw that Shumener (2020) takes conflicting accounts of the identity and distinctness of concrete objects to give competing explanations for facts of the form [$x = y$]. Similarly, one can view each of the above theories as attempting to motivate alternative explanations for facts of the form [*x* explains *y*]. Here, differing accounts of what it takes to explain a phenomenon can be seen to provide meta-explanatory generalizations:

(f) **The Inferential Proposal:** *For any x and y, if y can be logically deduced from x and the laws in question, then x explains y.*

(g) **The Conserved Quantity Proposal:** *For any x and y, if x and y causally interact in such a way that facilitates the exchange of a conserved quantity from x to y, then x explains y.*

(h) **The Interventionist Proposal:** *For any x and y, if it is possible to intervene on x with respect to y, in such a way that changes the value of y, then x explains y*.

Using the analysis detailed above, I believe that we can show that explanations which make use of (h) are deeper than those which make use of (f) and (g). Our first step is to show that interventionism provides deeper explanations than inferential accounts.[88] This is not a difficult task since the latter face well known problems in characterizing the intuitive asymmetry of explanation. Bromberger (1965) provides the most familiar example of such a case, concerning a flagpole and its shadow.[89]

Bromberger's example illustrates that, given the length of the flagpole's shadow (*x*) (and the angle of elevation of the sun), we can deduce the height of the flagpole (*y*). And yet, the length of a flagpole's shadow does not *explain* its height. What this means, in the parlance of Hitchcock & Woodward (2003b), is that the generalization specified by (f), will not remain invariant in such cases, where *y* can be deduced given *x* and the laws in question, despite the fact that *x* does not explain *y*.

The interventionist generalization specified by (h) fairs much better here. Part of the central motivation for interventionism itself stems from its ability to accurately characterise the asymmetric character of such explanations. As (h) suggests, while intervening upon the

[88] While the DN and unificationist pictures of explanation do differ in several respects, de Regt highlights that 'Kitcher's unificationist model turns out to be a sophisticated version of Hempel's deductive-nomological model, preserving the basic features of deductive argument and subsumption under laws' (2017:53). As a result, for my current purposes it is not necessary to draw a meaningful distinction between the two theories.
[89] Also see Barnes (1992).

height of the flagpole would allow us to manipulate the length of its shadow, the converse relation does not hold. That is, one cannot manipulate the height of the flagpole by intervening upon its shadow. Any such intervention would, itself, have to *go through* the height of the flagpole.[90] Consequently, (h) correctly characterises Bromberger's example as unexplanatory. Thus, the interventionist account will remain invariant under a wider range of counterfactual scenarios than inferential accounts and can be considered to provide a deeper *metaphysical* explanation of the connection between explanans and explanandum as a result.

The CQ proposal faces similar problems to inferential accounts.[91] As de Regt argues, Salmon's concept of causality is problematic at the deepest level of physical reality, where standard interpretations of quantum mechanics leave no room for 'continuous space-time trajectories along which energy and momentum are transported' (2017:61). Similarly, Woodward argues that:

> '[t]here are explanations, such as those involving causation by omission or by double prevention, that do not involve a physically interesting form of action at a distance, but are nonetheless cases of causal connection without intervening spatiotemporally continuous processes or transfer of energy momentum from cause to effect' (2003:353).

Of course, such cases will constitute interventions in which the criteria specified by (g) are not invariant since, although the connection between *x* and *y* is (at least widely accepted to be) explanatory, the criteria specified by the relevant explanatory generalization are not met. However, Woodward (2003) argues that interventionism can correctly characterize such instances, providing principled reason to deny that explanation is connected to the transfer of a conserved quantity along continuous space-time trajectories. As such, we can once again

[90] See Woodward (2003:98-102).
[91] Hitchcock (1995) notes that the CC proposal is, in fact, unable to make sense of many of the counterexamples to the DN model put forward by Salmon (1984) himself.

see that the range of interventions under which (h) will remain invariant is wider than the equivalent range for (g). Thus, interventionism provides deeper explanations of the connections between explanans and explanandum than the CC proposal.

So, I have argued that by adopting an interventionist reading of metaphysical explanation, along with the account of metaphysical explanatory depth outlined in this paper, we can show that interventionism provides the deepest account of the nature of explanation *itself*. However, in applying interventionism at this meta-explanatory level, to generalizations purporting to account for facts of the form [*x* explains *y*], it might not be immediately obvious what the "object" of the requisite "same-object" counterfactuals actually is.

On the traditional account of interventionism, discussed in section 2, an intervention involves altering the value of an explanans variable in order to manipulate an explanandum variable. In the case discussed above, we can intervene upon the height of the flagpole in order to manipulate the length of the shadow. It is precisely because flagpoles can be used to manipulate shadows that we take flagpole height to explain shadow length, and not *vice versa*. In terms of same-object counterfactuals, the "object" here is the flagpole.

However, interventionism (on my view, at least) does not merely claim to be a theory of the relationship between flagpoles and shadows, it claims to be a theory about the relationship between explanans and explanandum *tout court.* As such, in attempting to test the invariance of (f), (g) and (h) as accounts of facts of the form [*x* explains *y*], the "object" of the relevant "same-object" counterfactuals cannot be the flagpole alone. Briefly revisiting the example from the previous section will help us to get a grip on what is happening at this meta-explanatory level.

Recall that the generalizations (c), (d) and (e) purport to explain something about the relationship between *x* and *y*, namely [$x = y$]. I have argued that on an interventionist account

of metaphysical explanation, we can assess the depth of these generalizations by considering the range of interventions under which they remain invariant. However, such interventions are admittedly slightly different from those described in section 2. In the case of the identity and distinctness of concrete objects, we do not intervene on *A with respect to B*; our goal is not to attempt to manipulate one of our spheres *by* intervening upon the other.

Rather, we intervene upon the features of *A* and *B* and their relations to one another (*x*), in order to attempt to "manipulate" the relation of identity and distinctness itself (*y*). In this sense, the "object" of the same-object counterfactuals in question ought to be thought of as the combination of *A*, *B* and the features they possess. It is by intervening on this *system*, constituting *x*, that we are able to assess the range of interventions under which (c), (d) and (e) remain invariant and thus, the depth of the explanations they provide. As I see it, the "object" in the meta-explanatory case is the same.

On an *inferential proposal*, for example, the fact that we can deduce the length of the flagpole's shadow from its height and the relevant laws, itself explains why we take flagpoles to explain shadows. In this sense, the object of the relevant counterfactual is the system constituted by the flagpole, its shadow and the deductive relationship in which they stand; with the explanation relation itself being the thing which we are attempting to manipulate. As we have already seen, however, the fact that we can construct counterfactual scenarios, involving the same objects, in which we can deduce the height of the flagpole from the length of its shadow suggest that (f) provides for relatively shallow explanations of the relationship between explanans and explanandum here.[92]

[92] It is important to note that while the antecedents of such counterfactuals consist of multiple different objects and relations these objects stand in to one another, the objects themselves remain the same in each counterfactual considered. While we can alter these objects in various ways in order to assess the impact upon the relation of explanation which holds (or doesn't) between them, this is a far-cry from the sort of scope based "other-object" counterfactual model discussed in section II. It would perhaps be more accurate to call the counterfactuals involved in instances of metaphysical explanation "same-object[*s*]" counterfactuals, however I

As the above discussion shows, examples of causal explanation are enough to prove that (h) will remain invariant under a wider range of interventions that either (f) or (g). However, the *true* depth of the interventionist account of explanation becomes apparent only once one acknowledges that metaphysical instances ought to contribute to the range of counterfactual scenarios against which competing theories of explanation are measured.

To take an archetypal example, given the existence of Socrates, and a relevant law (e.g., *set formation* as embedded in Zermelo-Fraenkel set theory, see Shaffer 2017:309-310), we can deduce that {Socrates} existed; there are no possible worlds in which Socrates existed, but {Socrates} did not. However, the necessary connection here runs in both directions. Given that {Socrates} existed, Socrates' existence can also be deduced. So, according to (f), *x* explains *y* and *y* explains *x*. Yet the consensus suggests that it is the existence of Socrates explains the existence of {Socrates}, but not *vice versa*. Once again, an inferential account will not be invariant in such cases, where the occurrence of *x* can be deduced, given *y* and the laws in question, despite the fact that *y* does not explain *x*.

While the DN model sees explanation where we typically take there to be none, the CQ account faces the opposite dilemma. For reasons which we have already seen, the CQ account will have great difficulty making sense of the explanatory connection between the existence of Socrates and {Socrates}. The connection between *x* and *y*, here is not mediated by spatiotemporally continuous processes of conserved quantity transfer. Indeed, Woodward is cognizant of this difficulty: '[t]here are reasons to doubt that [the CQ account] is an extensionally adequate theory, in the sense that it correctly distinguishes between causal and noncausal interactions' (2003:30).

---

think that this would prove more confusing than helpful. For more on the possibility of the same/other-object distinction breaking down in the case of metaphysical explanation, see section VII.

What this means, of course, is that while (f) is violated in cases where metaphysical explanation is intuitively asymmetric, (g) will fail to capture the explanatory character of metaphysical explanations *altogether*. Since, as has been seen at some length, (h) accurately characterises such explanations with metaphysical character, every such instance represents a widening of the range of interventions under which (h) will be invariant when compared to (f) and (g). Thus, any metaphysical explanation will be an addition to range of invariance for (h), but not (f) and (g).

There is, however, a further benefit of accepting that metaphysical explanations ought to contribute to the range of counterfactual scenarios against which depth is measured. In the very act of adopting interventionism with respect to metaphysical explanation, the range of testing interventions under which explanations of the connection between explanans and explanandum remain invariant, widens. Which is to say, interventionism *itself* provides even deeper explanations of the connection between explanans and explanandum if one additionally accepts that this analysis ranges over metaphysical, as well as causal instances.

As a result, the account of metaphysical explanatory depth which I have attempted to elucidate appears to provide some motivation for those interventionists still on the fence with respect to metaphysical explanation. Accepting that metaphysical explanations can be given an interventionist treatment significantly widens the range of counterfactual scenarios under which (h) remains invariant. Thus, explanations provided by interventionists who reject an analogous analysis of metaphysical explanation will be *shallower* than those provided by interventionists who accept this analysis.

VI. Methodological Concerns

In this final section, I wish to discuss several methodological questions which arise as a result of the account of metaphysical explanatory depth which I have attempted to motivate. The

first concerns how we are to make sense of the distinction between same- and other-object counterfactuals in the context of identity and distinctness; the second, raises the issue of how we are to understand the role of interventions when it comes to metaphysical explanations more generally; and the third asks whether shallow metaphysical explanations are actually *explanations* at all.

With regards to the first concern, the worry here is that when considering explanations of identity and distinctness, the border between same- and other-object counterfactuals breaks down. If so, then it appears that the interventionist analysis would collapse into a scope-based account. In order to see why this is not the case, we will need to assess how a scope-based account of depth would cope in such scenarios. So, let's return to our possible world containing spatiotemporally separated qualitatively discernible objects: *A,* a sphere; and *B* a cube.

According to the *qualitative properties proposal*, since *A* and *B* are qualitatively discernible, they are distinct. However, in order to assess the depth of this explanation, on a scope-based account, we are invited to assess the range of other-object counterfactuals under which (c) will continue to hold. Imagine, for example that we substitute *A* and *B*, for *C* and *D*; objects which are *indiscernible* in terms of their qualitative properties.

Despite the fact that *C* and *D* share all of their qualitative properties, *were* it the case that *C* possessed the qualitative properties of *A*, and *D* possessed the qualitative properties of *B*, *C* and *D* would be distinct. In this scenario, because *C* would possess the property of being spherical, which *D* lacks, (c) would continue to hold. The problem for scope-based accounts, is that on an interventionist analysis of explanatory depth in terms of *same*-object counterfactuals, (c) *does not* remain invariant in scenarios concerning qualitatively indiscernible, spatially separated objects.

On an interventionist analysis, we do not substitute *A* and *B* for other objects in order to assess the scope of the relevant generalization, but rather imagine intervening upon *A* and *B* (or both) in such a way that results in a situation in which they share all of their qualitative properties. Here (c) is violated; it will mischaracterize *A* and *B* as being identical when we know them to be distinct (being spatially separated). So, it appears that a scope-based account is once again unable to draw a meaningful distinction between these competing accounts of the identity and distinctness of concrete objects in terms of depth; a distinction which *can* be drawn if we adopt an interventionist notion of depth.[93] As a result, it is clear that the contrast between same- and other-object counterfactual analyses of depth holds firm, even in scenarios involving the identity and distinctness of concrete objects.

In order to address the second methodological concern, I would first like to make an admission: I am what Emmerson (2021) has recently labelled an "intervention liberal". It is my view that interventions do not carve nature at its causal joints. I take it that, in at least *some* cases, interventions provide us with a useful tool in characterising metaphysical explanation. As we saw in section 3, the hypothetical interventions used to assess the depth of competing explanations for Jeff Bezos's being a billionaire, are entirely "possible" in the requisite sense, corresponding to 'conceptually possible or well-defined physical manipulations' (Woodward, 2018:122).

Even in the (admittedly more contentious) case concerning the identity and distinctness of concrete objects, I see little difficulty in conceptualizing the relevant interventions. It

[93] It is important to note that the same situation holds with respect to the *weak discernibility proposal.* Supposing that *A* and *B* are qualitatively indiscernible, spatially separated objects, but that *C* and *D* are qualitatively indiscernible, co-located objects. *Were* it the case that *C* and *D* stood in the same relations to one another as *A* and *B*, then *C* and *D* would be distinct according to (d). On an interventionist analysis (d) would be violated in this case, while (e) would remain invariant. As such, on a scope-based account, we are unable to draw meaningful distinction between the *weak discernibility* and *quantitative properties proposals* in terms of their depth.

seems entirely possible to imagine a hypothetical scenario in which, at t_1, only a single object exists (to which *A* and *B* both refer), and then to consider the implications of a manipulation which sees this single object split into two spatially separated objects, at t_2; objects which are qualitatively discernible from each other, *and* from the original object which existed at t_1.[94]

Despite this, there is not universal agreement on this point. Others, labelled "intervention puritans" by Emmerson (2021), believe that interventions exclusively serve to demarcate causal relationships, and that the notion becomes problematic in the context of noncausal explanation.[95] The worry is that, in cases of metaphysical explanation, the requisite interventions are neither well-defined, nor (in some cases at least) logically or metaphysically possible.[96] It is, however, beyond the scope of this paper to mount a sustained defence of the role of interventions with respect to noncausal explanation. For my purposes it suffices that there are already well-developed interventionist approaches analyses of metaphysical explanation on the table which are able to accommodate the account of metaphysical explanatory depth which I have provided above[97].

---

[94] It seems that I am in good company on this point. As I mentioned in section II., Schaffer (2017) explicitly characterises "metaphysical laws" in terms of invariance under testing interventions. Similarly, Wilson (2018) describes a wide variety of metaphysical explanations which can be accurately characterized using interventionist counterfactuals and structural equation models. Also see Schaffer (2016); Wilson (2018a, 2018b); and Miller & Norton (2022a, 2022b)

[95] Proponents of intervention puritanism include: Bokulich (2011); Leuridan (2012); Saatsi & Pexton (2013); Harinen (2014); Pexton (2014); Jansson (2015); Romero (2015); Baumgartner & Gebharter (2016); Baumgartner & Casini (2017); French and Saatsi (2018); Khalifa *et al* (2018, 2020); Reutlinger (2018); Saatsi (2018); Lange (2019).

[96] To the extent that intervention liberalism appears to require a commitment to *counterpossible nontriviality,* Schaffer (2016) highlights that there are already good reasons for thinking that counterpossible scenarios require non-trivial evaluation (see e.g., Restall 1997; Goodman 2004; Priest 2005; Berto & Jago 2013; Jago 2015; Jago *et al* 2018). What's more, largely as a result of their perceived utility in scientific explanation, recent years have seen a dramatic increase in attempts to motivate non-trivial counterpossibility (e.g., Tan 2019; Baron *et al* 2020; Kimpton-Nye 2020; Reutlinger *et al* 2020; Wilson 2021; Baron & Colyvan 2021). Consequently, it is clear that counterpossibles pose a problem of interventionism *in general*, not just for interventionist interpretations of metaphysical explanation.

[97] As a final note on this topic, I would highlight that regardless of whether one is able to *imagine,* or *conceive of*, the sort of hypothetical manipulation required by an interventionist analysis of metaphysical explanation, Schaffer argues that there is nothing formally problematic here: '[t]he mathematics doesn't "know" if an intervention is countermetaphysical or counter logical. It just sees adjusted values to variables and adjusted functions, which it solves as before' (2016:71).

And this brings us to our final methodological concern: whether shallow metaphysical explanations are actually explanatory at all. Consider an example used by Hitchcock & Woodward (2003b) to illustrate the notion of depth with respect to causal explanation: the laws of Newtonian mechanics and Einstein's relativistic correction to those laws. When applied to objects with a velocity that is relatively small compared to that of light, generalizations generated by Newtonian mechanics will remain invariant under a range of interventions, *R*, on that velocity.

However, the special relativistic correction to these laws will remain invariant under a much wider range of interventions *R**, where *R** strictly contains *R*, but also contains interventions upon velocities closer to that of light. In this sense the special relativistic corrections to Newtonian mechanics provide for deeper explanations insofar as they remain invariant under a wider range of testing interventions, *R**, despite the fact that Newton's laws *are* explanatory within the narrower range *R*.

In this example, it seems that we are comparing two *explanatory* generalizations; the "laws" of Newtonian mechanics appear to provide explanations within a given domain (concerning objects with velocities which are relatively small compared to that of light), despite failing to be invariant when this domain is expanded. In the cases of metaphysical explanation discussed throughout this paper, however, one might think that something rather different is going on. We have not been comparing *explanatory* generalizations at all. Such generalizations *compete* and, as a result, only one of them can be true and thus, explanatory.

By showing that the *quantitative properties proposal* provides deeper explanations than both the *qualitative properties proposal* and the *weak discernibility proposal*, what we are actually doing here is showing that the latter theories are false, and thus couldn't have been explanatory in the first place. Presuming that explanation is factive, by demonstrating that (e)

is invariant under a wider range of interventions, we provide *counterexamples* to both (c) and (d), which means they must be false and cannot qualify as metaphysically explanatory generalizations.

Woodward (2021) has recently argued that we should not think of invariance as *evidence* of truth. Rather, 'invariance in relationships is a matter of the holding of certain kinds of truths – truths that we regard as particularly important to discover, rather than something that competes with truth or is evidence of truth' (Woodward, 2021:266). That *x*, if true, would best explain *y*, is no reason to think that *x is* true, according to Woodward, because to show that the premises of an explanation are true 'we need to appeal to independent evidence in support of such truth claims' (2021:266). However, in metaphysics, such independent evidence will typically underdetermine which, if any, of the premises of our candidate explanations are true.

Instead, in analysing competing metaphysical theories, we are typically required to "grant" or "assume" their truth *for the sake of argument*, and then assess what each theory would commit us to were it, in fact, true. Such theorizing is not uncommon and arises throughout scientific practice. As Wilson (2021) highlights, physics is difficult, and false theories abound. If we are to have any hope of progressing towards the correct fundamental theory, thinking critically about various competing possibilities, and evaluating them by contrasting their consequences, is a methodological imperative. However, we are not required to judge hypothetical scenarios as *objectively* possible in order to investigate such theories. Wilson (2021) draws upon a helpful analogy with reductio arguments in mathematics to make this point.

Classically, mathematical statements are taken to be true if possible. As a result, in order to reason nontrivially about false mathematical claims, we must be able to reason

nontrivially about the impossible: '[m]athematicians may use a reductio argument to establish the falsity of a claim that they already know to be false (e.g., when teaching students)' (Wilson, 2021:1121). Reasoning in this way clearly requires that mathematicians be able to temporarily grant that the claim in question is true and hence, possible. In this sense, physicists, mathematicians (and metaphysicians) can 'adopt a noncommittal pretence of possibility for the sake of argument' (Wilson, 2021:1121).

When presented with the qualitative properties and weak discernibility proposals, the metaphysician does not have epistemic access to the truth or falsity of (c) and (d). Without the relevant empirical evidence to help us, the metaphysician can instead adopt a "noncommittal pretence of possibility for the sake of argument" and then proceed to examine the range of testing interventions under which each remains invariant. In this way, the metaphysician can demonstrate that the weak discernibility proposal provides deeper potential explanations of the identity and distinctness of concrete objects that the qualitative properties proposal, without needing to accept either theory as true.

Of course, exactly how we are to cash out the process involved in adopting a noncommittal pretence of possibility is another question. We could, for example, follow Toby Handfield (2004) and embed problematic counterfactuals in indicative conditionals. Alternatively, we might, as Wilson puts it, "go metatheoretical" and replace counterfactual reasoning with 'direct theorizing about models' (2021:1119); or appeal to fictionalism in the make-believe style of Kendall Walton (1990), Roman Frigg (2010), and Sam Kimpton-Nye (2020).

While the reader might not be enamoured with any of these options (and others are available), the purpose of this discussion is not to promote any particular methodology.[98] Rather my aim is merely to highlight that this process, granting the truth of a theory for the sake of argument, is by no means uncommon. Given that the practices of both mathematics and physics appear to require us to account for such theorizing already, I take it as no concession at all that the analysis of metaphysical explanatory depth which I have provided here might also be able to make use of such an account.

To return to the original point, it is clear that the connection between truth and depth is much stronger in the case of metaphysical explanation than in the case of causal explanation. While Woodward (2021) argues that depth and truth are entirely disconnected in the causal case, in the case of metaphysics depth itself can provide (defeasible) reason to believe that a given theory is true. When we adopt a noncommittal pretence of possibility with respect to a metaphysical theory, and can find no testing intervention which violates it, this provides us with at least *some* justification for our belief in its truth.

Conversely, if the metaphysician comes across a testing intervention under which the theory appears to be violated, they have good reason to believe the theory to be false; although this needn't prevent us from cogently talking about the structure of the world that the theory describes. However, while the notion of depth appears to play a different role within metaphysics and science, this difference does not lead to a difference in the

[98] More controversially, one might consider adopting something like the notion of *acceptance* put forward by Cohen (1992). According to Cohen, one *accepts* that *p* when one treats it as given, i.e., when one 'adopts a policy of… including [*p*] among one's premises for deciding what to do or think in a particular context' (1992:4). For alternative accounts of acceptance, and how this notion might differ from belief, see e.g., Van Fraassen (1980); Bratman (1992) and Maher (1993). Finnur Dellsén (2017), has recently argued that understanding, the cognitive achievement which results from grasping an explanation, can be accompanied by mere acceptance, rather than full-blown belief: 'belief and acceptance will coincide in most cases. However, they can come apart, viz. when one decides to adopt a policy of treating something as given despite being indisposed to feel that it is true' (2017:14). Dellsén argues that we can "treat" a theory as given and use it in our explanations of various natural phenomenon – thus accepting it for explanatory purposes – despite not believing it to be true and even, in some cases, where we believe it to be false.

methodology of *assessing* or *calculating* explanatory depth across these domains. As I have shown, an interventionist analysis requires little alteration in order to characterize the depth of both causal and metaphysical explanations.

Chapter 4

Interventionism, Understanding and Explanatory Knowledge

While the last decade seen growing interest in scientific understanding and its connection to explanation, those involved have typically adopted a methodological stance which privileges the former over the latter; attempting to provide analyses of understanding which remain silent on upon the nature of explanation. Despite widespread agreement among philosophers of science that understanding is constituted by explanatory knowledge, the prevalence of this "understanding-first" methodology has meant that theories of explanation are seldom evaluated explicitly with respect to how successfully they account for the connection between explanation and understanding. In this paper, however, I defend an alternative "explanation-first" methodology, which allows theories of explanation themselves to guide analyses of understanding. What's more, I maintain that when such an approach is adopted, it is interventionism that emerges as our best hope for a theory of scientific explanation capable of characterizing scientific understanding.

I. Introduction:

There is a broad consensus among philosophers of science that understanding is merely a species of knowledge; explanatory knowledge.[99] To *understand why p* occurs, on this view, one must possess an explanation of *p*; a proposition equivalent to "*p* because *q*". As Bird argues, for example, '[t]o understand why something occurred is to *know* what causes, processes, or laws brough it about.' (2007:84).[100] Despite the prevalence of this position,

[99] A distinction is often drawn between "objectual understanding" and "understanding-why". In this paper, my focus shall be on this latter notion, typically taken to be a type of cognitive achievement expressed by sentences like "*S* understands why *E*", where *E* is an explanandum.

[100] Also see e.g., Salmon (1984, 1989); Lewis (1986); Miller (1987); Hitchcock & Woodward (2003a); Woodward (2003); Lipton (2003); Strevens (2008); Khalifa (2017). This position is by no means universal, however. A number of epistemologists have become dissatisfied with this traditional picture, arguing that understanding is distinct from knowledge; a notion of *intrinsic* epistemic value. See e.g., Elgin (1996, 2004, 2007); Zagzebski (2001); and Kvanvig (2003, 2009); Pritchard (2008, 2010); Hills (2009, 2015); Gardiner (2012); and Mizrahi (2012). As a directed result, the landscape of debate surrounding scientific understanding

however, *theories* of explanation are rarely evaluated explicitly with respect to how successfully they characterize the connection to understanding. Indeed, in the only sustained attempt to engage with this issue, Kim (1994) comes to the startling conclusion that extant accounts of explanation resolutely *fail* as analyses of understanding.

While recent years have seen a significant amount of intellectual effort expended on the connection between explanation and understanding, those involved have tended to adopt a methodological stance which *privileges* the latter. de Regt, for example, suggest that 'we need a general theory of scientific understanding that is independent of a specific model of explanation but allows for the possibility that understanding can be achieved via markedly different explanatory strategies' (2017:86). Similarly, Khalifa maintains that he 'would like the larger points about understanding to swing freely of any of my idiosyncrasies about explanation, this favours being relatively noncommittal about the nature of explanation' (2017:6). In this paper, however, I attempt to revive the *explanation-first* methodology utilized, but ultimately abandoned, by Kim.

Engagement with Kim's (1994) work has tended to focus upon his ancillary claim that causation is merely one of a variety of dependence relations serving as the "objective correlates" of explanation. In light of recent interest in the notion of *grounding*, thought by many to be a noncausal dependence relation which provides the objective correlate for a distinctive form of metaphysical explanation, this focus is unsurprising.[101] However, successful accounting for the contribution of noncausal dependence relations to

has also begun to shift, with several leading philosophers of science having broken ranks. Chief among them are Dellsén (2016, 2017, 2018, 2021, *forthcoming*) and de Regt (2015, 2017, 2020, 2022), both of whom argue that scientific understanding is distinct from scientific knowledge and that it is the former upon which the epistemology of science ought focus.

[101] See e.g., Audi (2012); Trogdon (2013, 2018); Schaffer (2012, 2016, 2017); Maurin (2018); Wilson (2018a, 2018b); Trogdon & Skiles (2021).

understanding is, according to Kim (1994), merely *one* hurdle over which a theory of explanation must jump in order to qualify as a theory of understanding.

Equally important is the need to explain *how* and *why* knowledge of such relations contributes to understanding. In answering this question, Kim (1994) argues that a successful explanation-first account of understanding ought to be able to: show how understanding differs from merely descriptive knowledge; illuminate the unificatory role of dependence relations, in reducing the number of independent phenomena we need to recognize as fundamental; and clarify the important sense in which explanatoriness and understanding come in degrees. Disappointingly, however, at this tantalizing juncture Kim throws in the towel:

> 'To build a bridge from unity and simplicity to understanding and intelligibility, we need an epistemology of understanding, something that has by and large been neglected by contemporary analytic epistemology and philosophy of science' (1994:69).

Given Kim's conclusion, the recent focus upon *understanding-first* approaches, which attempt to provide analyses of understanding which are silent upon the nature of explanation, seem entirely justifiable. Nonetheless, I believe that a theory of explanation is now available which can meet all of Kim's criteria for a successful analysis of understanding. The theory is question is *interventionism*, as popularized by Hitchcock and Woodward (Hitchcock & Woodward 2003a, 2003b; Woodward 2003).

The structure of this paper is as follows. In sections II-IV, I outline Kim's motivation in claiming that extant theories fail to adequately account for this connection. Section II deals with Hempel's (1965) inferential analysis; section III with the unificationist accounts of Friedman (1974) and Kitcher (1981); and section IV with the explicitly causal theories of Salmon (1984) and Lewis (1986). In sections V-VIII I motivate an interventionist

explanation-first analysis of understanding. Roughly speaking, this picture suggests that understanding consists in grasping the range of interventions under which an explanatory generalization remains in variant.[102]

In section V, I show that interventionism provides a satisfying account of the contribution of explanatory knowledge to understanding, and how it differs from merely descriptive knowledge. On the interventionist account, the distinctive value of explanatory knowledge lies in enabling us to 'distinguish those relationships that are potentially exploitable for the purposes of manipulation and control from those that are not' (Woodward, 2003:36).

In section VI, I argue that the interventionist notion of *explanatory depth* allows us to accurately characterize the sense in which explanatoriness and understanding come in degrees. The depth of an explanation is a matter of the range of interventions under which it is *invariant*. The greater this range, the deeper the explanation. I propose that the degree of understanding which can be attributed to an agent, *S*, is directly proportional to the depth of the explanation known to *S*.

Given the widespread assumption that interventions are a purely causal notion, Kim's position on the role of *non*causal dependence in facilitating understanding might seem at odds with our own interventionist commitments. However, in section VII, I argue that interventionism is best understood as an analysis of explanatory dependence *simpliciter*; with

[102] The idea that "grasping" is the psychological notion necessary in order to achieve understanding is commonplace. Strevens, for example, suggests that an 'individual has scientific understanding of a phenomenon just in case they grasp a correct scientific explanation of that phenemonen' (2013:510); and Khalifa, that a 'natural suggestion is that explanatory understanding is the possession or "grasp" of an explanation' (2017:6). Also see e.g., Kvanvig (2003, 2009); Grimm (2006, 2010, 2014); Khalifa (2013); Strevens (2013); Wilkenfeld (2013); Hills (2015); Bourget (2017).

interventions functioning to provide knowledge of *causes* in the case of causal explanation and knowledge of *grounds* in the case of metaphysical explanation.

In section VIII, I turn to the role of unification in understanding. In a recent paper, Kovacs (2020) suggests that appealing to dependence relations does little to elucidate the connection between explanation and understanding. Instead, Kovacs argues that if we are to take this connection seriously, in science or metaphysics, then we ought to adopt a *unificationist* analysis of explanation. In contrast, I propose that an interventionist methodology can better articulate the role of unification in explanation and understanding owing *precisely* to its focus upon the illumination of dependence relations.

I do not claim that this interventionist analysis is inherently superior to understanding-first approaches. Rather, my intention here is merely to highlight that advances in our conception of explanation justify a second look at Kim's explanation-first programme; and that when taken seriously, this programme identifies interventionism as a clear frontrunner. I fully accept that the likes of de Regt and Khalifa are unlikely to be swayed; as Khalifa admits, 'gaze deep into my soul, and you will see a card-carrying explanatory pluralist staring back at you' (2017:8). Gaze into *my* soul, however, and you will see a dyed-in-the-wool interventionist staring back. In this sense, what follows can be read as defence of interventionism against the common understanding-first strategy which seeks to side-line questions about the nature of explanation.

II. Understanding as Nomic Expectability

Despite the close connection typically taken to hold between explanation and understanding, Kim notes that terms like "understanding", "intelligibility" and "explanatory knowledge" are often jettisoned when 'serious theory construction begins… [and] seldom make an

appearance once the initial stage-setting is over' (1994:52). One reason for this, is the long-held suspicion that understanding is a purely "psychological", or "pragmatic" (Bunge 1973).[103]

This position can be traced back (at least) as far as Hempel (1965) who argues that, in the context of understanding, explanation is 'a relative notion: something can be significantly said to constitute an explanation in this sense only for this or that individual' (1965:426). In contrast, by focusing upon *nomic expectability*, Hempel intends his own analysis to reflect the "objective" aspect of explanation: 'the sense of being independent of idiosyncratic beliefs and attitudes on the part of the scientific investigators' (1983/2001:374).

On Hempel's deductive-nomological (DN) model, an explanation is an argument consisting of a premiss *G* (a statement representing the occurrence of an event *g*), a conclusion *E* (a statement representing the occurrence of an event *e*), and a law *L* (a statement specifying a universal generalization) allowing for the deduction of *E* from *G*. Crucially, Hempel argues that 'the argument shows that, given the particular circumstances and the laws in question the occurrence of the phenomenon *was to be expected*' (1965:337).

Given this focus, it ought to come as little surprise that Kim finds the DN model ill-placed to illuminate the role of explanatory knowledge in promoting understanding. Kim argues that Hempel's view collapses the explanatory relation between *g* and *e* into a logical relation between their descriptions, *G* and *E*. Explanation becomes a matter of "logico-linguistic" connections between *descriptions* of events, and 'the job of formulating an explanation consists, it seems, in merely re-arranging appropriate items in the body of

[103] Trout (2002), for example, has argued that understanding is nothing more than a feeling of confidence or satisfaction gained when one has seemingly answered a question adequately. And Humphreys similarly suggests that a focus upon understanding will lead to a 'relativization of explanations to an individual' and that, as such, we ought to 'set aside that whole issue of what constitutes or promotes understanding' (1989:127).

propositions that constitute our total knowledge at the time' (Kim, 1994:55-56). Understanding, on the DN model, takes place entirely *within* the "epistemic system" and, as a result, 'on the "subjective" side of the divide between knowledge and the reality known, or between representation and represented' (Kim, 1994:56).

What's more, such "explanatory *internalism*" fails to capture the important sense in which understanding and explanatoriness appear to come in degrees.[104] On the DN-model, explanation (and therefore, understanding) results from the explanandum statement being *derivable* from, or *entailed* by, the explanans statements. These logical notions are an "all-or-nothing" affair. *E* is either entailed by *L* and *G* or it is not; *G* is either a successful explanation of *E,* capable of promoting understanding, or it is not. There is no middle ground.

More promising, however, are the *unificationist* accounts of Friedman (1974) and Kitcher (1981), both of whom appear to be cognizant of the importance of understanding to an account of explanation. Friedman, for example, demands that 'a theory of scientific explanation tell us what it is about the explanation relation that produces understanding' (1974:6); and Kitcher criticises the DN model for failing to explain 'why those derivations which employ laws advance our understanding' (1981:168).

However, despite this apparent desire to carve out a central role for understanding, Friedman and Kitcher's analyses also tie explanation to law-based derivations. As a result, while Kim agrees upon the importance of unification to our practices of explanation and understanding, he ultimately concludes that these unificationist accounts suffer from similar problems to the DN model.

[104] As Hills notes: 'we do talk about understanding why (in at least some contexts) as if it comes in degrees and there is some suggestive linguistic evidence supporting this. "Understanding why p" is "gradable", that is, it is similar to verbs such as "regret" or adjectives such as "tall" (2015:665). Also see e.g., Kvanvig (2003, 2009); Khalifa (2013); Wilkenfeld (2013); Kelp (2015).

III. Understanding as Unification

According to Kitcher, it is unification, rather than nomic expectation, which is doing the explanatory work on the DN model. Behind Hempel's "official" nomological view, we are invited to recognize an "unofficial" view which regards explanation as unification (Kitcher, 1981:167). As Hempel himself acknowledges, explanation aims at:

> 'an objective kind of insight that is achieved by a systematic unification, by exhibiting the phenomenon as manifestations of common, underlying structures and processes that conform to specific, testable, basic principles' (1966:83).[105]

In pursuing this "unofficial view", Kitcher argues that an argument or derivation is explanatory in virtue of its membership of a class which *best unifies* our system of beliefs.[106] Explanation provides understanding by 'showing us how to derive descriptions of many phenomena, using the same patterns of derivation again and again', and in doing so, 'teaches us how to reduce the number of types of facts we have to accept as ultimate (or brute)' (Kitcher, 1989:432).

Crucially for Kim, however, while Kitcher endorses Friedman's demand that a theory of explanation give an account of understanding, neither suggest that explanation involves discovering or imparting, *additional* knowledge. In both cases, explanation is once again an activity which consists in constructing derivations, the steps of which are logically related to the rest of our belief system. Unification, then, also appears to depend 'solely on factors

[105] Also see Hempel (1965:345 & 444).
[106] Friedman (1974) provides a similar analysis which focuses on laws, rather than classes. As Kim highlights: '[f]or Friedman whether or not a given law explains another is determined crucially b the unifying power of the explaining law, and the concepts in terms of which the latter is explained are exclusively logical ones (equivalence, implication, etc.) and evidential ones ("independent acceptability")' (1994:64).

internal to the epistemic system, such as the number of argument patterns required to generate the given class of arguments, the "stringency of patterns, etc"' (Kim, 1994:64).

In taking the epistemic virtues of understanding to concern 'our representations of the world, but not the world itself', Kim argues that unificationist accounts of explanation suffer from similar problems to the DN model (1994:64). Explanation thus becomes a matter of the structure and organization of our entire belief system, rather than of the content of the propositions which constitute it; a "holistic" process, which makes it 'impossible to say anything precise and useful about the epistemic gain to be associated with individual explanations' (Kim, 1994:64).

As a result, unificationist treatments of explanation also face difficulty in accounting for the sense in which explanation and understanding come in degrees. On Kitcher's model, for example, explanatory power cannot be determined locally. Rather, a derivation is explanatory only if it belongs to the set which 'collectively provides the best systematization of our beliefs' (Kitcher, 1989:430). However, set membership is an all-or-nothing notion. A given derivation is therefore either: a member of the set of that best unifies our belief system, in which case it can provide understanding; or it is not a member of this set, in which case it cannot provide understanding.

This is not to say, however, that these unificationist accounts of explanation get *everything* wrong. Indeed, Kim is sympathetic to the idea that unification is a central function of our explanatory practices, although, he suggests that the unifying effect of explanations can be seen at work in *both,* our belief system, and in the world. In order to capture this worldly sense of unification, Kim argues that we require an account of explanation which not only connects propositions within our body knowledge, but also the "objective correlates" of such propositions, which lie *outside* of it (1994:56).

IV. Understanding as Knowledge of Causes

According to Salmon and Lewis, what warrants the use of a proposition *G,* in explaining an event *e,* is that the event which *G* represents, *g*, is a *cause* of *e*. Lewis (1986), for example, argues that *every* explanatory claim is a causal claim, and that an explanation is informative precisely insofar as it is informative about the causal history of *e.* Similarly, Salmon suggests that 'causal processes, causal interactions, and causal laws provide the mechanisms by which the world works; to understand why certain things happen, we need to see how they are produced by these mechanisms' (1984:132).

One problem is that it is far from obvious just *how* understanding "springs" from our ability to 'discover and formulate causal judgements of the sort that Salmon [and Lewis] would consider explanatory' (Kim, 1994:61). Why, to use an example of Kim's, does knowledge that Socrates' death was caused by his drinking hemlock, promote understanding, when knowledge that this event took place in 399BCE, does not? What we need is an account of why some knowledge has a distinctive explanatory character, purely in virtue of its content; an account which neither Salmon, nor Lewis, provide.

As I mentioned at the outset, however, recent interest in Kim's (1994) work has focused upon his ancillary claim that causation is just one of many explanation-grounding relations knowledge of which constitutes understanding. This presents explicitly causal accounts of explanation with another problem; how to account for *non*causal explanations. Causation, Kim argues, is 'one type of dependence, obviously one of central importance', however, he goes on to highlight *mereological* dependence as '[a]nother dependence relation, orthogonal to causal dependence and equally central to our scheme of things' (1994:67).

While Kim appears to take causal and mereological forms to be the most prolific types of dependence serving as "explanation-grounding" relations, he does consider others: Xantippe's widowhood appears to depend upon Socrates' death; mental states are widely regarded as being dependent upon the physical nature of the brain; and evaluative and normative properties seem to depend upon factual, or nonevaluative properties. In each case, Kim argues, the existence of a *non*causal dependence relation serve to 'generate explanations' (1994:68).

These days, such explanation-generating noncausal dependence relations are often thought to be united under the umbrella of "grounding". As Jonathan Schaffer summarizes the connection between grounding and explanation: '[g]rounding connects the more fundamental to the less fundamental, and thereby backs a certain form of [metaphysical] explanation' (2012:122).[107] Indeed, echoing Kim, it has recently been argued that the grounding relation is necessary in formulating physicalism (e.g., Schaffer 2009; Rosen 2010; Dasgupta 2014; Ney 2016), and may well be "indispensable" for normative theorizing more broadly (Berker 2018). What's more, understanding is also often taken to be the epistemic aim of metaphysical explanation.[108]

So, while Salmon and Lewis are right to tie understanding to knowledge of the "objective correlates" of explanations, they are mistaken in thinking that causation exhausts the scope of such relations. However, despite also failing to account for the distinction between explanatory and descriptive knowledge, Kim (1994) argues that these externalist theories make better sense of the unificatory role of explanation than the internalist approaches of Hempel, Friedman and Kitcher.

---

[107] Kim similarly argues that just as causal dependence 'rationalizes our attempt to look for the diachronic, temporally antecedent determinants of phenomena', mereological dependence rationalizes of attempt to 'search for their synchronic micro-determinants' (Kim, 1994:67). Also see: Kim (1984/1993:77).
[108] See e.g., Thompson (2016); Dasgupta (2017); Maurin (2017); Schaffer (2017); Dellsén (2018).

We think of the world, according to Kim, 'as a system with structure, not a mere agglomeration of unconnected items, and much of the structure we seek comes from the pervasive presence of dependence relations' (1994:68). The "ontological contribution" of dependence relations like grounding and causation lies precisely in the role of reducing the number of independent events, states, facts and properties which we need to recognize as fundamental, or brute. In this sense, Kim maintains that unity and structure are intimately connected, that dependence *enhances* unity by generating structure. It is at this point, however, that Kim appears to give up: '[h]ow unification or simplification is to be connected to explanatory understanding is a difficult question, and I will have nothing useful to add to what has already been said by Friedman, Kitcher and others' (1994:67).

As was noted at the outset, given the broad consensus that understanding is knowledge gained by grasping an explanation, philosophical *theories* of explanation ought to be centrally concerned with distinguishing knowledge capable of imparting such understanding, from merely descriptive knowledge. While providing such an analysis of the connection between knowledge and understanding is a key task for a theory of explanation, it is not the *only* task, however. In characterizing the distinction between explanatory and descriptive knowledge, a philosophical theory of explanation also needs to be able to account for three additional aspects of understanding.

These are: the sense in which explanatoriness, and thus understanding, come in degrees; the role of *non*causal dependence relations in generating explanations; and the unificatory element to explanation, facilitated by dependence relations. In what remains of this paper, I shall argue that interventionism can meet each of these requirements, providing a successful account of the connection between explanatory knowledge and understanding. As a result, I maintain that this theory represents our best hope for a coherent explanation-first analysis of understanding.

V. Explanatory Knowledge as Knowledge of Manipulability

Like Kim, Woodward (2003) argues that by characterizing explanation in terms of the law-based derivations, both the DN and unificationist theories fail to adequately distinguish between explanatory and "merely descriptive" knowledge (Woodward, 2003:31; 2003:636). The issue here for Kim (1994) concerns the idea that explanation is entirely a matter of the "logico-linguistic" relations between descriptions of events; factors internal to an epistemic system.

This methodology fails to adequately tie understanding to knowledge of the objective correlates of explanation. Such correlates are, according to Kim (1994), the dependence relations between the events themselves, which underpin the explanatory relation between their descriptions. Hitchcock & Woodward similarly argue that this difficulty in distinguishing between explanatory and descriptive knowledge arises as a result of prior theories failing to tell us precisely what the explanandum 'depends upon' (2003a:18).

Successful explanation, according to Woodward, ought to enable us to see 'what sort of difference it would have made for the explanandum if the factors cited in the explanans had been different in various possible ways' (2003:11). In order to do so, however, an explanation must be *invariant* under testing interventions, which is to say, it must 'describe a relationship which holds for certain *hypothetical* values of *X* and *Y possessed by the very object o*… where *X* is changed by an intervention' (Hitchcock & Woodward, 2003a:20).

Woodward (2003) is also explicit that a philosophical theory of explanation ought to be able to account for the fact that some knowledge plays a special explanatory role purely in virtue of its content. He argues that any such theory ought to be able to explain how, and why, causal knowledge is of practical use: what it enables 'us (or other animals) to achieve

with respect to such practical goals as survival and reproduction that other kinds of knowledge does not' (Woodward, 2003:30).

In answer to this question, Woodward proposes that the distinct benefit of explanatory knowledge has to do with our ability to manipulate and control the world around us. An intervention serves as something like 'a hypothetical or counterfactual experiment that shows us that and how manipulation of the factors mentioned in the explanation… would be a way of manipulating or altering the phenomenon to be explained' (Woodward, 2003:11). It is, furthermore, 'only if a generalization is invariant under testing interventions that it conveys information about how one variable depends upon another' (Hitchcock & Woodward, 2003a:19).

In this sense, Hitchcock and Woodward follow Salmon and Lewis in taking the objective correlates of explanation, relations of dependence, to be '"out there" in nature' (Woodward, 2003:23). Unlike Salmon and Lewis, however, the interventionist can give a compelling account of why some knowledge has a distinctive explanatory character, purely in virtue of its content. For example, our knowledge that Socrates' death was caused by his drinking hemlock promotes understanding *precisely* because we know that certain testing interventions upon this event would result in a different outcome. For example, had Crito intervened and knocked the hemlock to the floor, Socrates' life would have been saved.

On the other hand, knowledge that the event in question took place in 399BCE does not promote understanding, because the year in which Socrates consumed the hemlock has no bearing upon the outcome. Had Crito intervened to have Socrates granted a one-year stay of execution, this event would have taken place in 398BCE, yet the outcome remains the same. All else being equal, Socrates would still have consumed the hemlock and died. Given that this intervention upon the year in which Socrates' drank hemlock does not allow us to

manipulate the outcome, knowledge that the event took place in 399BCE is not explanatory and thus does not contribute to our understanding of why Socrates died.

According to the traditional view, S *understands* why *e* occurs where S knows that *G* explains *E.* To know that *G* explains *E*, on an interventionist account, is to know that it is possible to intervene upon *g,* in such a way that changes the value of *e*. In Kim's terminology, the objective relation connecting events, *g* and *e*, that grounds the explanatory relation between their descriptions, *G* and *E*, is one of dependence, and such dependence relations become known to us through the implementation of testing interventions (1994:56).

The interventionist account of explanation looks a lot more like an account of a type of explanatory knowledge than those considered (and rejected) by Kim (1994). Indeed, unlike the accounts of accounts of Hempel (1965), Friedman (1974), Kitcher (1981), Salmon (1984) and Lewis (1986), interventionism is crystal clear on the distinction between explanatory knowledge and merely descriptive knowledge: explanatory knowledge enables us to 'distinguish those relationships that are potentially exploitable for the purposes of manipulation and control from those that are not' (Woodward, 2003:36). With this first challenge out of the way, in the next section, I examine the notion of explanatory depth available on an interventionist account and show how it can be used to characterise the sense in which explanation and understanding come in degrees.

VI. Degree of Understanding as Explanatory Depth

Once again echoing Kim (1994), Hitchcock & Woodward (2003a, 2003b) argue that both the DN and unificationist theories make explanatoriness an all-or-nothing affair. Traditionally, the laws needed to derive an explanation are taken to be *universal* or *exceptionless* generalizations. However, according to Hitchcock & Woodward, this distinction results in an

"exhaustive dichotomy" of true generalizations: 'a true generalization is either a law, in which case it is explanatory, or it is accidental, in which case it is not explanatory. There are no other possibilities' (2003b:183).

As such, it becomes difficult to see how one generalization could be any more or less explanatory than another. On the DN model, given the explanans, every law makes the explanandum equally *expectable* (Hitchcock & Woodward, 2003a:18-21). On top of this issue, in taking only those derivations comprising the "best" systematization of our beliefs to be explanatory, unificationist accounts conclude that 'only our deepest, most unified theories are explanatory at all; everything else is non-explanatory' (Hitchcock & Woodward, 2003b:194).

However, on an interventionist account, generalizations need not be exceptionless, or universal to be explanatory. Rather, by characterizing explanatory power in terms of invariance under testing interventions, Hitchcock and Woodward maintain that '[a]mong those generalizations that are invariant, some will be more invariant than others, and they will correspondingly provide *deeper* explanation' (2003b:183-184). For one explanation to be to be deeper than another, then, is for it to answer a greater range of *what-if-things-had-been-different-questions* concerning a given target object or system.

In the previous section, I suggested that Socrates' having consumed hemlock *explains* his death, in light of the possibility of using the former event to manipulate the latter: where Crito knocks the hemlock to the floor, Socrates survives. However, *hemlock* consumption is by no means the only explanation of this event available; nor is it the deepest. Consider, as another possible explanation, Socrates' having consumed *coniine*; the poisonous substance found in plants like hemlock, fool's parsley, and the yellow pitcher plant.

On Hitchcock & Woodward's account, an explanation of Socrates' death which references the fact that he consumed coniine will be *deeper* than an explanation of Socrates death which merely references the fact that he consumed hemlock. To see why imagine that, rather than knocking the hemlock from Socrates' hand, Crito now manages to intercept the drinking vessel while the concoction is being prepared and replaces the hemlock with a seemingly innocuous plant he found in the garden (which is, unbeknownst to him, fool's parsley).

Presumably feeling rather smug at having pulled off this feat of horticultural espionage, one can imagine Crito's confusion when Socrates' dies, nonetheless. Here, hemlock consumption clearly no longer explains Socrates' death; this explanation is not *invariant* under Crito's intervention. In contrast the fact that Socrates' ingested coniine, the poisonous chemical found in both hemlock and fool's parsley, explains Socrates' death in both scenarios.

Coniine consumption provides a *deeper* explanation of Socrates' death than hemlock consumption, because the former explanation remains invariant under a wider range of interventions than the latter. Consequently, someone who *knows* that Socrates consumed coniine will be able to answer a wider range of what-if-things-had-been-different questions concerning this scenario, than someone who merely knows that he consumed hemlock.[109] While both parties will be able to answer questions like "what would have happened if Crito had knocked the hemlock to the floor?"; only the former will be able to answer questions like "what would have happened had Socrates consumed fool's parsley, or a yellow pitcher plant, instead of hemlock?".

[109] In "Understanding as Knowledge of Causes", Grimm hints at something like this idea in noting that understanding will typically grant the 'ability to answer a variety of what [Woodward] calls "What if things were different?" questions' (2014:334).

This allows the interventionist to cash out the sense in which explanatoriness, and thus understanding, come in degrees. *S* understands why *e* occurs where *S* knows that *G* explains *E*. However, given that explanations vary in depth, or the range of interventions under which they remain invariant, *S*'s understanding of *e* can also vary in depth. The degree of understanding which can be attributed to *S* is thus, on an interventionist picture, directly proportional to the depth of the explanation known to *S*. Someone who knows that Socrates consumed coniine will thus have a better *understanding* of why Socrates died, than someone who merely knows that Socrates consumed hemlock.

In section VIII, I will argue that this notion of explanatory depth can also be used to characterize the role played by unification in connecting explanation to understanding and, what's more, that this interventionist interpretation of unification does a better job of accounting for this connection than archetypal accounts (e.g., Kitcher 1981). Before doing so, however, in the next section I shall argue that the notion of an intervention can be utilized in making sense of the role of dependence relations in providing the objective correlates for *noncausal* dependence.

VII. Understanding as Knowledge of Causes and Grounds

In section IV, we saw that Salmon (1984) and Lewis (1986) appear to a capture an important element of the connection between explanation and understanding, insofar as they suggest that understanding requires knowledge of the objective correlates of explanation. Despite this, Kim argues that these explicitly causal accounts of explanatory knowledge still face two hurdles. The first is that they fail to explain why some knowledge plays a distinctive explanatory role purely in virtue of its content. In section V, however, I argued that the

interventionist analysis of explanation avoids this problem, by tying explanation to dependence through the notion of an intervention.

The second problem facing causal accounts of explanation, according to Kim (1994), is that not *all* explanations are causal in nature. While Kim agrees that understanding requires knowledge of the objective correlates of explanation, he argues that causation is merely one such relation. On the face of it, this appears to present a problem for my claim that interventionism can meet Kim's (1994) demands of a theory of understanding. This is because it is still widely assumed that where it is possible to intervene upon *X*, in such a way that changes the value of *Y*, *X causes Y.*[110]

As was also highlighted in section IV, however, the sort of explanation-generating noncausal dependence relations in which Kim (1994) is interested, are now often taken to be unified under the umbrella of "grounding". One notable trend within the recent grounding literature, has been to characterise the grounding relation explicitly in terms of Woodward's (2003) interventionist analysis of *causal* explanation.[111]

Schaffer (2016, 2017), for example, argues that metaphysical explanation is, like causal explanation, connected to manipulation. This manipulationist role, according to Schaffer, is elucidated through 'Woodward's (2003) guiding conception of explanations as serving to answer "what if things had been different" questions, and Skyrm's (1980:11) idea that wiggling the value of one variable wiggles the value of another' (2017:206). What's more, Schaffer sees this manipulationist aspect as connected to both the unifying feature of explanation, and its role in providing understanding:

[110] In a recent paper, Emmerson (2021) labels this position "intervention puritanism".
[111] See, for example, Schaffer (2016, 2017); Reutlinger (2017); Wilson (2018a, 2018b); Miller & Norton (2021, *forthcoming*); Emmerson (2022b).

> ‘Laws are the stable patterns which unify phenomena, provide recipes for manipulation and guide understanding. So I conclude that if there are metaphysical explanations, then there must be counterfactual supporting general principles in the metaphysical realm’ (2017:307).

In section V, we saw Hitchcock & Woodward (2003a) argue that we come by explanatory knowledge through the implementation of testing interventions, and that such knowledge is explanatory in virtue of conveying ‘information about how one variable depends upon another’ (Hitchcock & Woodward, 2003a:19). What Schaffer (2017) is suggesting, in appealing to interventionism in characterising grounding, is that the knowledge imparted as a result of a successful intervention is not always *causal* knowledge. While a successful intervention does indeed serve to highlight the existence of a dependence relation, this fact alone underdetermines whether said relation is one of causation, or one of grounding.

Schaffer is by no means to first to recognize the relation to manipulability shared by both causal and noncausal explanation. Indeed, in “Noncausal Connections” Kim (1974/1993) points to the “bringing about” relation as a unifying feature of both causal and noncausal explanation. Portending his later work, Kim argues that ‘there appear to be dependence relations between events that are not causal’ and that the thesis of universal determinism (“every event has a cause”) ought to be revised to include them ‘if we are to have a clear and complete picture of the ways in which events hang together in this world’ (1974/1993:22).

These observations motivate Kim to put forward the thesis that both causal and noncausal connections might be characterized in terms of a single unifying relation “*R*”: ‘a broad relation of event dependency that subsumes as special cases the causal relation and other dependency relations’ (1974/1993:27). While Kim (1974/1993) does not provide a substantive account of what this relation might be, it is clearly the very same relations “*R*”

which he later refers to as the "explanation-grounding" relation (Kim, 1994). What Kim *does* suggest, however, is that both the causal and noncausal connections, subsumed by "*R*", share a close connection to the "agency relation":

> 'the asymmetry of the agency relation in "We could bring about Xantippe's widowhood by bringing about Socrates' death" points to the asymmetry of the dependency relation between Xantippe's widowhood and Socrates' death. As in the causal case, the asymmetry of the former appears to be rooted in the asymmetry of the latter' (1974/1993:25).[112]

Nicholas Emmerson has recently noted that an interventionist methodology provides Kim's intuitive conception of "bringing about" with a more formal characterization, arguing that 'it is clear that *manipulation* is something like the notion which Kim is intending to highlight with his discussion of the connection between agency and dependence' (2021:3). Even Woodward's position on the topic of noncausal explanation appears to have softened of late.[113] He notes, for example, that 'there is an obvious sense in which it is true that by manipulating whether or not Socrates dies, one can alter whether Xantippe is a widow' (Woodward, 2018:121).

What this suggests, I believe, is that interventionism is best understood as a methodology for characterizing explanatory dependence, *simpliciter*; as opposed to a methodology for characterizing *causation*, or causal dependence, specifically. In the case of metaphysical explanations, successful interventions function to provide knowledge of

---

[112] Kim is perhaps most explicit about his manipulationist sympathies in his 1984 paper "Concepts of Supervenience", where he suggests that 'the idea that "real connections" exist and the idea that the world is intelligible and controllable are arguable equivalent ideas', arguing that it is in virtue of such connections that 'the world can be made intelligible; and by exploiting them we are able to intervene in the course of events and alter it to suit our wishes' (1984/1993:53).

[113] While Woodward appears to have been cognizant of the existence of noncausal explanations, he had previously argued that these cannot be given a manipulationist interpretation: '[w]hen a theory or derivation answers a what-if-things-had-been-different question but we cannot interpret this as an answer to a question about what would happen under an intervention, we may have a noncausal explanation of some sort' (2003:221). Also see Woodward (2015).

grounds; and in the case of causal explanation, successful interventions function to provide knowledge of causes.

On this interpretation, interventionism is able to overcome both of the problems identified by Kim with respect to the explicitly causal accounts of Salmon and Lewis. In the first instance, interventionism explains why some knowledge plays a distinctive explanatory role: such knowledge relays information about the possible manipulability of the phenomenon in question. And, in the second, the notion of manipulability can be utilized in characterising explanations resulting from the existence of both causal and grounding relations. Of course, one final hurdle remains: unification.

VIII. Unification as Explanatory Depth:

In a recent paper, David Kovacs criticises grounding theorists for having become overly fixated upon the idea that metaphysical explanations 'track objective, worldly order', while ignoring their epistemological features; the sense in which they 'increase understanding, make the phenomena intelligible, satisfy our curiosity etc' (2020:1661). Indeed, Kovacs echoes Kim's (1994) charge against Salmon and Lewis, in arguing that 'it's not transparent what it is about backing relations that yields understanding, or, if you prefer, answers to why-questions' (2020:1672).

On the former point, we agree. This paper and Kovacs' (2020) share a core goal: to put scientific and metaphysical explanation back on an equal epistemological footing. Where we *dis*agree, however, is upon the best way to accomplish this task. In the second half of this paper, I have attempted to make clear the role played by "backing relation", or "objective correlates" of explanation in yielding understanding. According to Kovacs (2020), however,

rather than getting clear on exactly *how* such relations contribute to our understanding, we should simply abandon the idea that they have a role to play in promoting understanding *altogether*.

Instead, Kovacs argues that it is Kitcher's (1981, 1989) unificationist theory which provides the best hope of re-establishing 'the unduly neglected link between explanation and understanding in the metaphysical realm' (2020:1663). Following Kitcher, Kovacs (2020) claims that *explanatory* knowledge is knowledge belonging to the *most unified* set. Metaphysical explanation is thus a holistic affair, providing a "global" form of understanding which helps us to 'see how a large number of phenomena are the consequences of a small number of basic facts, from which they can be derived using relatively few and similar patterns of basic derivation' (Kovacs, 2020:1673).

However, Kovac's criticism of grounding theorists, for having 'a lot to say about the worldly aspect of explanation but much less about its epistemological aspects' (2020:1962), loses its bite once one remembers that these seemingly disparate aspects of explanation are actually closely connected. As Kim (1994) argues, the ontological contribution of dependence relations like causation and grounding lies precisely in their unificatory role; reducing the number of phenomena which we need to recognize as fundamental. What's more I believe that an interventionist analysis of explanation is able to provide a more convincing account of the unificatory power of explanatory knowledge, *precisely* because of its focus upon the illumination of such worldly relations.

While both unificationism and interventionism share a commitment to the idea that successful explanation ought to apply to a number of different cases, they differ with respect to what they take the relevant cases to be. On Kitcher's (1981, 1989) view, this range of cases are understood in terms of *scope*: the set of objects or systems that fall under the antecedent

of a given law. As Hitchcock & Woodward describe this notion, for a generalization like "all *A*'s are *B*'s" to be explanatory, 'it must 'support' counterfactuals of the following form: if some object *o** that is different from *o* and does not possess property *A* were to be and *A*, then it would be a *B*' (2009a:19).

In this sense, the greater the range of *other objects* to which an explanation applies, the more unifying it is, and thus the greater the understanding which it provides. In contrast, interventionists are not concerned with the range of *other objects* for which a given explanation holds, but rather with the range of changes to the *actual object* being explained. Let's return to the example from section V to see this distinction play out. If you recall, we considered two possible explanations of Socrates' death: (a) Socrates' having ingested *hemlock*; and (b) Socrates' having ingested coniine.

While both the fact that Socrates consumed hemlock, and the fact that he consumed coniine, explain the *actual* event of his death, I argued that the latter provides a deeper explanation. As a result, I maintain that someone who knows that Socrates consumed coniine will have a greater understanding of why he died, than someone who merely knows that Socrates consumed hemlock. However, despite the obvious sense in which (a) provides less understanding than (b), on a unificationist account like that endorsed by Kovacs (2020), however, these explanations are mischaracterized as being equally unifying.

The problem here, lies with Kitcher's (1981) interpretation of unification in terms of scope. Consider, the scenario in which it is Crito, rather than Socrates, who consumes the hemlock. In this case, we would expect Crito to meet the same unfortunate end as Socrates. It is in this sense, on a unificationist account, that (a) can be said to explain Socrates death: because were anyone else to consume hemlock, we would expect them to die as well. The problem is that the same is true of (b): had we, or Crito, consumed *coniine*, death would be

the likely result. Indeed, all else being equal, there is no object with which Socrates can be replaced that will allow us to draw a meaningful distinction between (a) and (b) in terms of scope, and thus, unification.

For this reason, Hitchcock & Woodward (2003b) argue, that we ought to adopt an interventionist account of explanation, where unification is understood in terms of explanatory depth. In the present case, that Socrates consumed coniine provides a deeper explanation of his death, because it remains invariant under a wider range of testing interventions. The appeal to coniine ingestion here thus allows us to unify *all* the scenarios in which (b) is invariant (including hemlock and fool's parsley consumption), under a single explanation. Of course, this notion of unification looks rather different to that advocated by Kovacs and Kitcher.

Indeed, an interventionist analysis vindicates many of the features which Kim deems central to an account of unity. For starters, it is *local*; a feature of the propositions and events involved in individual explanations, rather than a *holistic* or *global* feature of our whole belief system (Kim, 1994:68). We know that coniine consumption unifies Socrates' having ingested hemlock, and Socrates' having ingested fool's parsley, without needing to consult our whole belief system. Such wide-ranging consultation, as Kim points out, 'isn't something we do or need to do; it probably isn't something that any of us *can* do!' (1994:65).

What's more, the interventionist analysis takes unity to be a feature 'of events and facts of the world as well as of our beliefs and propositions' (Kim, 1994:68). As Kim puts it, the world is no 'mere agglomeration of unconnected items', rather, we typically think of it as having structure, structure which 'comes from the pervasive presence of dependence relations' (1994:68). This is a central benefit of the interventionist analysis of explanation; as

was argued in section VI, interventions themselves serve to provide us with knowledge of such relations.

An agent can be said to understand why *e* occurs, where they possess an explanation of *e* equivalent to "*E because G*". On the interventionist picture, to possess such an explanation, is to know (roughly) that it is possible to intervene upon *g*, in such a way that changes the value of *e*. In other words, the content of explanatory knowledge concerns a dependence relation which holds between *e* and *g*; the worldly events (or properties, states, facts etc) represented by *E* and *G*.

Concluding remarks:

At the outset, I identified a curious puzzle arising from a common assumption among philosophers of science. Despite the prevalence of the view that understanding is merely explanatory knowledge, philosophical theories of explanation themselves typically fail to provide a substantive account of how the latter gives rise to the former. Indeed, the recent literature has tended to take the opposite, *understanding-first,* approach to this topic, attempting to provide an analysis of understanding which is silent upon the nature of explanation. At a first glance, Kim's (1994) argument appears to justify this methodological strategy.

Having subjected the DN model, unificationism and causal accounts to his explanation-first methodology, Kim concludes that none of them are able to present a satisfying account of the connection between explanation and understanding. Indeed, as we have seen, Kim himself suggests that the solution to this problem lies in the epistemology of understanding; in characterizing a notion which is both clear and rich enough for useful theorizing.

Nonetheless, in this paper, I have attempted to motivate an alternative to this understanding-first methodology; an explanation-first analysis which can meet all the criteria specified by Kim as being necessary in order for a theory of explanation to qualify as a theory of understanding. In so doing, I have explicitly defended an interventionist explanation-first account of understanding.

First and foremost, interventionism provides a satisfying account of the contribution of explanatory knowledge to understanding; how it differs from merely descriptive knowledge. Explanatory knowledge is distinctively valuable insofar as it enables us to distinguish those relationships which are (at least potentially) exploitable for the purposes of manipulation, from those that are not. What's more, interventionism allows to us accurately characterize the sense in which explanatoriness and understanding come in degree, through the notion of *explanatory depth*. The degree of understanding which can be attributed to *S*, is directly proportional to the depth of the explanation known to *S*.

I have also argued that interventionism is able to account for the role of noncausal dependence relations in promoting understanding. On my preferred interpretation, interventions function to provide knowledge of causes in the case of causal explanation, and knowledge of grounds in the case of metaphysical explanation. Last, but by no means least, I have shown how an interventionist theory is able to articulate the role of unification in explanation and understanding. The resulting notion is both *local* (applying to individual explanations), as well as being a feature of events and facts in the world (not just our beliefs about it).

However, this analysis is not intended to be reductive or eliminative, although the reader is free to take it as such should they wish. Woodard defends his own non-reductive strategy by arguing that we can elucidate a concept 'by tracing its interconnections with or

locating it within a circle of interrelated concepts, but without claiming to analyse the concepts in this circle in terms of concepts that live entirely outside it' (2003:27). I would argue something similar if pressed. My aim here has merely been to justify an explanation-first methodology which *starts* with explanation and traces a path through interrelated concepts to arrive at a viable conception of understanding. As I highlighted at the outset, I do not claim that this notion is in any way more accurate than, or preferable to, alternative notions arrived at through an understanding-first analysis.

Nor do I intend to have the reader believe that all of the arguments put forward here are entirely novel. Woodward (2003), for example, defends the idea that explanatory knowledge is knowledge about manipulation and mounts an interventionist critique of unificationism. Similarly, Lipton (2001) presents a difference-making account of the contribution of causation to explanation. Nowhere, however, have such arguments been brought together into a coherent whole: a detailed defence of an interventionist analysis of understanding as explanatory knowledge. In this sense, the above can be read as a modest attempt to defend an explanation-first analysis of understanding against rival analyses which remain silent on the nature of explanation. That interventionism emerges as our best of hope for an explanation-first analysis of understanding represents an achievement *tout court,* regardless of the viability of the methodology itself.

Chapter 5

Putting Explanation First: Progress in Science and Metaphysics

In this paper, I present a unifying account of progress across science and metaphysics. In order to do so, however, I first motivate a novel "explanation-first" methodology for the analysis of progress, which recognises that different theories of explanation provide for markedly divergent interpretations of progress. I then defend a specific interventionist variant of this methodology. On this analysis, progress occurs when scientists and metaphysicians provide increasingly invariant explanations of target phenomena. What's more, I maintain that scientific and metaphysical theories can be said to "correspondence" where the range of invariance displayed by a progressive theory strictly contains the range of invariance displayed by a superseded theory.

I. Introduction

This paper has three goals:

1. To motivate a novel *explanation-first* approach to analysing scientific progress.
2. To defend a specific *interventionist* variant of this approach.
3. To show that this approach can be utilized to characterize progress in *metaphysics.*

While the claim that science makes progress is uncontroversial, the precise nature of this progress has long been a topic of intense debate.[114] For most of the 20th century two conflicting accounts dominated proceedings. On the one hand, realists like Popper (1962) characterized progress as the accumulation of true beliefs. On the other, antirealists like Kuhn

[114] See e.g., Kuhn (1962); Popper (1962); Laudan (1977); Niiniluoto (1980, 1987, 1999, 2014, 2017); Bird (2007, 2008, 2019, 2022); Rowbottom (2008, 2015); Mizrahi (2012, 2013, 2017, 2021, 2022); Cevolani & Tambolo (2013); Dellsén (2016, 2017, 2018, 2021, 2022); Park (2017; 2020); Sterpetti (2018); Shan (2019, 2022a); Dellsén, Lawler & Norton (2022, *forthcoming*); Emmerson (2022a); Lawler (2022).

(1962) and Laudan (1977), argued that progress is made by solving "puzzles" or "problems". More recently, the focus of this debate has shifted to epistemic considerations, with Bird (2007) and Dellsén (2016) characterizing progress respectively in terms of the accumulation of knowledge and increasing understanding.

While Bird and Dellsén pay lip service to the importance of explanation in progress-making, neither party provides a substantive account of how, exactly, explanation is supposed to facilitate progress. However, such an analysis is necessary, I believe, because different theories of explanation provide for markedly divergent interpretations of progress, and an ill-informed choice in this respect can render epistemically motivated accounts of progress untenable. I shall argue that both knowledge- and understanding-based conceptions of progress are best served by the adoption of an *interventionist* analysis of explanation. On the resulting "interventionist explanation-first" (IEF) analysis, progress is made when scientists provide increasingly *invariant* explanations of target phenomena.

In stark contrast to science, the claim that philosophy makes progress is highly contested.[115] While it is science to which philosophy invariably finds itself compared when the topic of progress comes up, attempts to provide a *unifying* analysis of progress are remarkably thin on the ground.[116] However, my contribution to this broader unificatory project will be relatively modest, concerning only one small corner of philosophy: metaphysics. Historical interest in the apparent *contiguousness* of science and philosophy has largely been motivated by metaphysics, making it the natural place to start in constructing a

[115] Indeed, as Shan (2022b) notes, *pessimism* appears to prevail. This attitude is typically motivated by reference to either: the apparent lack of consensus within philosophy (Horwich 2012; Bourget & Chalmers 2014; Rescher 2014; Chalmers 2015; Shand 2017); or the fact that philosophical theories are rarely superseded (and thus discarded) in the way that scientific theories often are (Sterba 2004; Dietrich 2011; Jones 2017; Slezak 2018).
[116] For comparisons of progress in science and philosophy see e.g., Russell (1912); Rapaport (1982); van Inwagen (2004); Rescher (2014); Chalmers (2015); Gutting (2016); Stoljar (2017); Brock (2017); Kamber (2017); Cappelen (2017); Frances (2017); Jones (2017); Bengson, Cuneo & Shafer-Landau (2019).

unified analysis of progress.[117] If a framework for thinking about scientific progress cannot be applied to metaphysics, the philosophical discipline most closely associated with science, there seems little promise of expanding it into other, less closely associated areas.

The structure of this paper is as follows. In the next section, after briefly outlining historical debate concerning scientific progress, I discuss a recent paper by McKenzie (2020), which claims that analogous progress within metaphysics is in principle impossible. According to McKenzie, scientific progress requires *correspondence*; a notion traditionally understood in terms of approximation. To paraphrase Popper (1959), the central equations of a progressive theory must contain, at least a good approximation of, those of the theory(s) it supersedes. However, McKenzie maintains that this necessary element of scientific progress has no clear analogue outside of mathematicised disciplines. As a result, science appears to have a monopoly on progress.

I agree with McKenzie on the importance of correspondence. Progressive theories must indeed be able to capture the explanatory potential of their predecessors, while adding novel explanatory power of their own. Nonetheless, I believe that this necessary component of scientific progress need not be cashed out in terms of approximation. Rather, by adopting either a knowledge- or understanding-based conception of scientific progress and an interventionist account of explanation, I shall argue that correspondence can be defined in terms of *invariance*. Since this notion of correspondence abstracts away from the central equations of the theories involved, it *can* be applied outside of mathematised disciplines.

I begin my defence of this claim by tackling the first of this paper's aforementioned goals: motivating a novel *explanation-first* approach to analysing scientific progress. In section III., I argue that, despite offering conflicting accounts of scientific progress, Bird

[117] See e.g., Duhem (1904-5); Wittgenstein (1921); Carnap (1932); Quine (1951); Popper (1959). Gillies (1993) provides an excellent analysis of this history.

(2007) and Dellsén (2016) agree on one crucial point: that the most significant episodes of scientific progress involve explanation. Despite this, as noted above, neither Bird nor Dellsén offer a substantive account of how or why explanation gains this privileged epistemic status. In section IV., I highlight the importance of providing such an account by demonstrating that, when combined with either inferential or causal analyses of explanation, both knowledge- and understanding-based conceptions of progress are rendered untenable.

I then move on to the second goal: defending an *interventionist* variant of the explanation-first approach to scientific progress. In section V., I introduce the interventionist methodology popularized by Hitchcock and Woodward (Hitchcock & Woodward 2003a, 2003b; Woodward 2003) and argue that it provides a natural way of characterizing progress. According to the resulting IEF analysis, scientists make progress by providing increasingly invariant explanations of target phenomena. What's more, I argue that correspondence occurs where the range of invariance displayed by a superseded theory is *strictly contained* within the range of invariance displayed by a progressive theory. In section VI., I demonstrate how this IEF account of progress avoids the issues facing rival inferential and causal analyses.

Finally, in section VII., I take on the third goal: utilizing this novel approach to characterize progress in metaphysics. In order to do so, I apply my IEF account to a case-study concerning two rival explanations of the identity and distinctness of concrete objects: the *qualitative properties proposal* (QPP) and the *weak discernibility proposal* (WDP). I argue that progress in this case can be characterized in precisely the same terms as scientific progress. The WDP is progressive with respect to the prior QPP, insofar as the former provides explanations which display greater invariance than those provided by the latter. What's more, since the range of invariance displayed by the QPP is strictly contained within the range of invariance displayed by the WDP, I maintain, *contra* McKenzie (2020), that these metaphysical theories correspond.

II. Progress, Truth and Correspondence

Until relatively recently, the battlelines of debate surrounding the nature of scientific progress had largely been dictated by metascientific considerations, principally those arising from responses to the Pessimistic (Meta)Induction (PMI); the inference from the failure of theories once thought successful, to scepticism regarding the truth of current (and future) theories. For antirealists, like Kuhn (1962) and Laudan (1977), the PMI shows that discovering truths cannot be the goal of science. If science aims at truth, and many (perhaps most) theories once thought true have now been proven false, then science has been *unsuccessful*. If, however, science aims at something other than truth, the falsity of superseded theories cannot be considered a failure of science.

In light of past refutation, Kuhn and Laudan argue that scientific progress must be measured in terms of solving "puzzles" or "problems", success in which can be judged only from within a "paradigm" or "research tradition".[118] It is, according to Laudan (1977), the "workability" of this *functionalist-internalist* analysis of progress which is its greatest virtue. While the PMI supposedly makes determining the truth of a theory impossible, it doesn't impinge on our ability to 'determine whether a given theory does or does not solve a particular problem' (Laudan, 1977:127).[119]

[118] While the term "progress" here is partly evaluative, indicating *improvement* over time, accounts of scientific progress are not intended to capture *every* sense in which science could be said to improve. Niiniluoto (2019) calls the type of progress at issue "cognitive progress" and distinguishes it from methodological, economical, educational, and professional progress. As Dellsén, Lawler and Norton highlight, 'although science would improve by being better funded, by adopting more reliable ways to conduct peer review, or by increasing gender equality among scientists, these types of improvement are not the subject of the aforementioned accounts of progress' (2021:10).

[119] See Shan (2019, 2022a) for a recent attempt to revitalize Kuhn and Laudan's classic antirealist approach to scientific progress.

Many realists also ultimately conceded the conclusion of the PMI, if the not the validity of the argument itself.[120] In response, Popper (1963) introduced the notion of *verisimilitude* in an attempt to capture the idea that 'scientific theories, though born refuted, are nonetheless increasingly getting some things right' (Bird, 2022:43).[121] More recently, Niiniluoto (1980, 1987, 1999) has constructed a philosophy of critical realism on the foundations of his own, more sophisticated, analysis of verisimilitude. According to the resulting *semantic* account, scientific progress occurs between t_1 and t_2, when the theories accepted at t_2 are more verisimilitudinous, or *truthlike*, that those accepted at t_1.[122]

It is this realist response to the PMI which motivates McKenzie's (2020) scepticism regarding the prospect of progress in metaphysics. Despite having been briefly reinstated as a respectable enterprise during the latter half of the 20th century, metaphysics has once again found itself the target of widespread derision. Many appear to concur with Ladyman & Ross that metaphysics simply 'fails to qualify as part of the enlightened pursuit of objective truth, and should be discontinued' (2007:vii). Nonetheless, as McKenzie notes, those making such claims often go on to posit grandiose metaphysical theses: 'Ladyman and Ross [2007] argue for the fundamentality of relational structure given the facts of modern physics; likewise, Maudlin [2007] argues for the fundamentality of laws, hence modality, and the primitivity of the direction of time' (2020:4).

Implicit in much of the recent rhetoric, then, is an apparent concession. It's not that *all* metaphysics is a waste of time; while *a priori* metaphysics isn't worth pursuing, *naturalized* or "scientifically informed" metaphysics *is*. Although even this weaker sceptical argument

[120] Psillos, for example, argues that '[i]n our interactions with the world, the exact truth cannot generally be had… A perfect match between theories and the world is impossible' (1999:276).
[121] The notion of verisimilitude was intended to capture a balance of both informativeness and approximation to the truth. As such, an uninformative but true theory may be *less* truthlike than a theory which is strictly speaking false but nonetheless highly informative.
[122] Also see e.g., Worrall (1989); Oddie (1986); Kuipers (2009); Cevolani & Tambolo (2013); Ladyman (2014).

has been challenged. Williamson (2013), for example, raises the spectre of demarcation, noting that the naturalist's critique is meaningless without a criterion for what it is to count as "science"; and Chakravartty (2017) argues that the idea of being "scientifically informed" is so nebulous as to place practically no constraint on acceptable metaphysics.

Nonetheless, McKenzie maintains that metaphysics faces a more fundamental, yet all too familiar, sceptical challenge:

> '[S]ince physicists do not yet have a fundamental theory, we can expect our current physics theories to be replaced by other theories in the future… Thus it seems that we expect our future science-informed metaphysics to be *radically different* from our current science-informed metaphysics' (2020:6).

At a first glance, the problem here might not be obvious. As we have just seen, while similar scepticism drove the likes of Kuhn and Laudan to embrace full-blown antirealism, from the ashes of Poincaré's "ruins piled upon ruins" rose the phoenix of realism reborn.[123] If the intuitively progressive nature of science can be secured in the face of the PMI, one might naturally expect the same to be true of metaphysics. *This*, however, is precisely what McKenzie denies.

In making her case, McKenzie adopts a conception of progress akin to the semantic account, arguing that 'it is through *better approximations to the truth* that we take science to make epistemic progress' (2020:8). For realists, the chief takeaway from debate surrounding the PMI is that a genuinely progressive theory ought to offer a more *refined* description than its predecessors: 'a theory which has been well corroborated can only be superseded by one

[123] Poincaré's famous statement of the argument is worth repeating in full: 'The ephemeral nature of scientific theories takes by surprise the man of the world. Their brief period of prosperity ended, he sees them abandoned one after another; he sees ruins piled upon ruins; he predicts that the theories in fashion to-day will in short time succumb in their turn, and he concludes that they are absolutely in vain. This is what he calls the *bankruptcy of science*' (1952:160).

[which] *contains* the old, well corroborated theory – or at least a good approximation of it' (Popper, 1959:276).[124] Such approximation must, according to McKenzie, reflect 'the continuity that exists between the central equations of those theories, relations that we call *correspondence*' (2020:10).

However, given that this notion of correspondence is defined in terms of the central equations of the theories involved, McKenzie argues that this necessary element of scientific progress has no obvious analogue outside of mathematicised disciplines. It is thus difficult to see how metaphysics can capture the idea that "retention through change" is a characteristic feature of progress. McKenzie concludes that, since the language of approximation cannot be meaningfully applied to metaphysical theses, 'it is very unclear how metaphysics could somehow inherit or participate in the progress enjoyed by science' (2020:8). In other words: metaphysics cannot make progress in the same way as science, so we have little reason to think that it makes progress *at all*.

For what it's worth, on this point McKenzie and I agree: when progress is defined in terms of better approximations to the truth, and continuity in terms of the central equations of scientific theories, the prospects for metaphysical progress do not look good. Where we *dis*agree, however, is with respect to the further claim that the *only* 'interpretation of metaphysical progress to invoke here is in terms of better approximations of the truth as well' (2020:19). As McKenzie freely admits, the nature of scientific progress is a topic of intense debate: 'there are disagreements over whether it is best conceived of in terms of truth, knowledge, representation, or understanding, or something else again' (2020:9, fn10).

[124] McKenzie highlights the transition from Galilean to special relativity as a paradigmatic example of what 'Post [1971] termed the *generalized correspondence principle*: the doctrine that old and new theories are virtually always retained as approximations of the new' (2020:11).

III. Progress, Knowledge and Understanding

According to Bird, given science's status as an epistemic activity, it seems 'almost tautologous to suggest that its success and so progress should be measured by epistemic standards' (2022:51). Nonetheless, it is precisely because of a failure to recognize such standards that Bird rejects both the semantic and functionalist-internalist accounts of scientific progress; arguing that they mischaracterize progress, and thus make it too easy to achieve.[125] On Bird's (2007, 2022) alternative *epistemic* account, science progresses through the accumulation of knowledge; true belief, justified by reliable scientific methodology.

The epistemic and semantic accounts agree that truth is a central component of progress; one which is abandoned entirely by *functionalist-internalist* analyses. While scientists are undoubtedly interested in solving puzzles, or problems, Bird argues that 'the notion of solving here is an objective externalist one. A scientist wants to uncover the *correct* solution to the problem, not merely an answer that meets the standards of her peers' (2022:48). Where the semantic and epistemic accounts diverge, is with respect to true beliefs with insufficient epistemic support to qualify for knowledge. For Bird, '[t]he acquisition of beliefs by an unreliable method cannot be genuine scientific progress, even if the beliefs so acquired are, by accident, true' (2022:52).

It appears, however, that one can recognize science as an epistemic activity, and nonetheless maintain that progress tracks a cognitive achievement *other* than knowledge. While Bird is explicit in locating his analysis of progress within a tradition which equates epistemic standards with justification, a growing number of epistemologists have become disillusioned with this knowledge-first project.[126] As Grimm elaborates: 'over the last several

[125] See Bird (2007:65-71; 2022:45-54). While there is not the space to discuss them here, nothing argued in this paper will turn on the success of Bird's arguments to this effect.
[126] As Bird explains: 'the view I found myself developing was one that rejected empiricism and embraced epistemological externalism, in particular Williamson's [1997] knowledge-first approach' (2022:viii).

years a number of leading epistemologists… have grown increasingly dissatisfied with the focus on knowledge and have attempted to "recover" the notion of understanding' (2012:103).[127]

Historically, there had been a relative consensus among philosophers of science that understanding is merely a species of knowledge. To understand *why p* occurs is thus to possess an explanation of *p*; a proposition equivalent to "*p* because *q*".[128] In contrast, Grimm notes that 'virtually every major epistemologist… has come to the conclusion that understanding is *not* a species of knowledge' (2006:516).[129] As a result, the landscape of debate surrounding scientific understanding has also begun to shift, with several leading philosophers of science having now broken ranks. One such example is Dellsén (2016, 2018, 2022, *forthcoming*), who makes use of this meta-epistemological trend in presenting a novel account of scientific progress.

According to Dellsén's *noetic* account, progress is made through increasing understanding, that is 'when scientists grasp how to correctly explain or predict more aspects of the natural world than they did before' (2016:75).[130] Of course, given that philosophers of science typically take understanding to be a *form* of knowledge, an understanding-based account of scientific progress is at serious risk of collapsing into a knowledge-based account.

---

[127] There is controversy concerning the exact translation of the Greek "*episteme*", from which the term "epistemology" derives. Moravscik (1979), Burnyeat (1984), Lear (1988) and Benson (2000) argue that *episteme* is more accurately translated as meaning "understanding", rather than the traditional translation of "knowledge". In this sense, Grimm (2012) argues that the apparent *shift* to an understanding-based conception of epistemology can actually be seen as a *return* to the original focus of the discipline.

[128] See e.g., Achinstein (1983); Salmon (1984, 1989); Lewis (1986); Miller (1987); Kim (1994); Kitcher (2002); Woodward (2003); Lipton (2004); Bird (2007); Stevens (2008).

[129] See e.g., Elgin (1996, 2004, 2007); Zagzebski (2001); and Kvanvig (2003, 2009); Pritchard (2008, 2010); Hills (2009, 2015); Gardiner (2012); and Mizrahi (2012). Grimm (2006, 2010) and Sliwa (2015) are notable exceptions to this rule.

[130] The idea that "grasping" is the psychological notion necessary in order to achieve understanding is commonplace. As Dellsén himself understands it, "grasping" involves the ability to 'infer, explain, or mentally manipulate, which extends not just to actual circumstances but also to various counterfactual circumstances' (2016:75). Also see e.g., Kvanvig (2003, 2009); Grimm (2006, 2010, 2014); Khalifa (2013); Wilkenfeld (2013); Hills (2015).

To avoid such a collapse, it must be shown that understanding and knowledge can be prized apart and that it is the former, *rather than* the latter, which tracks progress.

While, on Dellsén's (2016) view, both knowledge and understanding require truth, what distinguishes them is that the latter can be gained absent the *justification* required to possess the former.[131] For example, while Einstein's (1905/1956) explanation of Brownian motion in terms of the kinetic theory of heat was highly speculative at the time, lacking the justification required to constitute knowledge, Dellsén argues that this episode clearly constitutes progress.[132] The noetic account can make sense of this, since Einstein correctly explained the phenomenon in question, whether he *knew* it or not.[133]

While disagreement surrounding the status of understanding and its relation to knowledge continues, I expect such debate to ultimately be settled by meta-epistemological considerations, rather than considerations pertaining to the nature of scientific progress specifically. Fortunately, for the purposes of this paper, questions relating to the methodology of epistemology can be put to one side. This is because my primary interest in the epistemic and noetic accounts of progress doesn't concern their differences, but rather an important idea which they share: that understanding is the cognitive achievement towards which *explanation* is directed.

---

[131] While understanding is not an intrinsically realist notion (see e.g., Regt [2015, 2017]), Dellsén maintains that understanding must be at least *quasi*-factive, insofar as 'the explanatorily/predictively essential elements of a theory must be true in order for the theory to provide grounds for understanding' (2015:73, fn6). Most epistemologists seem to agree with Dellsén on this point (e.g., Kvanvig 2003, 2009; Mizrahi 2012; Wilkenfeld 2019), although some take understanding to require *fully* factivity (e.g., Grimm 2006; Hills 2009; Pritchard 2010).

[132] Indeed, Einstein himself admits that '[i]t is possible that the movements to be discussed here are identical with so-called "Brownian molecular movements"; however, the information available to me regarding the latter is so lacking in precision, that I can form no judgement in the matter' (1905/1956:1).

[133] In rejecting justification as a necessary component of progress, Dellsén admits that his view bears a striking resemblance to the semantic account; the chief difference being whether 'proposing a new explanation or making a new prediction could itself constitute progress, even when there is no change in the theories with which one would explain and predict' (2018:10). While Dellsén maintains that his own view can accommodate such progress, Rowbottom (2015) has argued that the semantic account cannot.

For Bird, understanding is merely a type of knowledge. To understand *why* something occurred, according to Bird, is 'to know what causes, processes, or laws brought it about' (2007:84). Consequently, on the epistemic account, science progresses via increasing knowledge, rather than understanding, because all genuine understanding *just is* knowledge. However, Bird also suggests that 'it is plausible to hold that those additions to knowledge that are also instances of understanding are, other things being equal, *more significant* than those that are not' (2007:84 – emphasis added).[134] Thus, on the epistemic account, the most significant episodes of scientific progress will be explanation-involving.

Dellsén appears to agree that explanation plays a particularly important role in promoting progress. The noetic account takes progress to be constituted by increasing understanding, which itself results from the ability to 'correctly explain and/or predict some aspects of the target phenomenon in the right circumstances' (Dellsén, 2018:7). However, Dellsén (2016) also suggests that a *complete* understanding of a phenomenon cannot be attained absent an explanation of that phenomenon. While an agent has *some* understanding if, for example, she realizes how to correctly predict changes in the weather based on barometer readings, a 'complete understanding of the weather would undoubtedly also require a grasp of how to *explain* the relevant changes' (2016:75, fn15).[135]

Consequently, on both the epistemic and noetic accounts of progress, explanation appears to hold a privileged epistemic position; explanation contributes *more* to progress than mere descriptive knowledge or prediction. In this sense, Bird and Dellsén agree that the *most valuable* kind of progress results from scientists providing explanations. With this in mind, it

[134] More recently, Bird reiterates that '[p]rogress with understanding is generally more valuable than progress without it' (2022:62).
[135] The idea that prediction is *sufficient* for understanding is uncommon. The more typical view, even among those who reject the conflation of understanding with knowledge, is that understanding results from explanation. Here is Strevens, for example: '[a]n individual has scientific understanding of a phenomenon just in case they grasp a correct scientific explanation of that phenomenon' (2013:510).

ought to come as a surprise to discover that neither party provides a substantive *analysis* of explanation in the context of scientific progress. Such an analysis important, I believe, because the viability of both knowledge- and understanding-based accounts of progress depends upon the corresponding notion of explanation which is adopted. In the next section, I argue that neither inferential nor explicitly causal analyses of explanation are compatible with an explanation-first approach to scientific progress; rendering both the epistemic and noetic accounts untenable.

IV. Inference, Causation and Progress-Making

For much of the 20th century, the philosophy of science was dominated by *inferential* analyses of explanation. The two most popular accounts, Hempel's (1965) deductive-nomological (DN) model and Kitcher's (1981) unificationist analysis, were united in taking explanation to involve deductive argument and nomic subsumption. According to the DN model, for example, an explanation is an argument which consists of a set of ancillary statements $C_1 \ldots C_n$, asserting the occurrence of events, and a law L, stating an exceptionless generalization. To explain some phenomenon E, is thus to derive it as a consequence of $C_1 \ldots C_n$, given L.[136]

As Bromberger (1965) and Barnes (1992) highlight, however, both the DN and unificationist analyses of explanation face a fundamental problem insofar as they are 'unable to respect a well-known feature of empirical explanations: its asymmetric structure' (Barnes,

[136] Kitcher (1981) argues that behind this "official model" of explanation in terms of nomic expectation, there is an "unofficial model" of explanation in terms of *unification*. Hempel himself corroborates this interpretation in arguing that scientific explanation aims to provide the kind of understanding which can only be achieved 'by a systematic unification, by exhibiting the phenomena as manifestations of common, underlying structures and processes' (1966:83). However, while the unificationist framework is widely regarded to be the more sophisticated of the two views, it nonetheless preserves the central features of its predecessor. Since it is these features which give rise to the problems detailed below, I will avoid a detailed discussion of unificationism.

1992:558). Suppose, for example, that Theresa becomes the first person to discover that we can deduce the length of a flagpole's shadow from its height, given the angle of incidence of the light-source. Now imagine that Theresa's apprentice, Boris, continues her work and discovers that the converse is also true; given the angle of incidence of the light-source, we can deduce the height of a flagpole from the length of its shadow.

Regardless of whether we take scientific progress to be constituted by knowledge or understanding, intuition tells us that Theresa's "discovery" is much more significant than Boris's; that Theresa contributes *more* to scientific progress here. One obvious reason for this, at least as far as Bird's and Dellsén's analyses suggest, is that only Theresa's discovery constitutes an explanation. While the heights of objects help to explain the length of their shadows, lengths of shadows don't explain the heights of their castors. What this means, of course, is that only Theresa's discovery can impart *understanding*. On an inferential analysis, however, it appears that both discoveries contribute equally to understanding and thus progress, since all that is necessary for explanation is deduction.

The issues for inferential accounts of explanation do not end here, however. As Salmon (1971) highlights, the problem of explanatory asymmetry is closely connected to another: the problem of explanatory relevance. Suppose, for example, that Rishi regularly consumes birth control pills. Given that Rishi is a cisgender man, and that all such men who regularly consume birth control pills fail to become pregnant, we can deduce that Rishi will fail to become pregnant. Thus, on an inferential analysis, it looks like Rishi's consumption of birth control pills *explains* his failure to become pregnant. However, to suggest that the "discovery" of this fact should be considered a significant contribution to scientific progress seems palpably absurd. Nonetheless, on an inferential analysis of explanation, this conclusion is unavoidable.

While inferential accounts of explanation share a focus upon *epistemic* considerations, widespread acknowledgement of the issues highlighted above helped to usher in a new era of explicitly causal, or "ontic", theories. According to Lewis (1986), for example, an explanatory claim is a causal claim, and an explanation is informative precisely insofar as it is informative about the causal history of a phenomenon. Similarly, Salmon suggests that 'causal processes, causal interactions and causal laws provide the mechanisms by which the world works; to understanding why certain things happen, we need to see how they are produced by mechanisms' (1984:132).

By tying explanation to causation in this way, it was hoped that that the asymmetry of explanation could itself be explained in terms of the asymmetry of cause and effect. However, the dominance of causal accounts of explanation was comparatively short-lived, largely because of increasing recognition that many scientific explanations do not cite causes at all.[137] Examples of noncausal scientific explanation are pervasive, and yet, if explaining a phenomenon is exclusively a matter of tracing its causal history, then these cases cannot count as contributing to understanding, and thus cannot add much (if anything) to scientific progress.[138]

Bird (2007) discusses an analogous case, which he takes to be evidence of knowledge increasing *without* a corresponding increase in understanding. This example concerns a group of researchers who are engaged in counting, measuring and classifying, geologically, the

[137] As early as 1974, Kim highlights that there appear to be explanatory relations between events that are not causal and that, as a result, we ought to revise our commitment to the thesis of universal determinism ("every event has a cause") to include these relations 'if we are to have a clear picture of how events hang together in the world' (1974/1993:22).

[138] An inexhaustive list of examples includes: graph-theoretic explanation (van Fraassen 1989; Pincock 2012; Lange 2013a); topological explanation (Huneman 2010; Lange 2013a); geometric explanation (Lange 2013a); statistical explanation (Lipton 2004; Lange 2013b); explanation in terms of symmetry principles and conservation laws (Lange 2011); kinematic explanation (Saatsi 2018); renormalization group theory (Batterman 2000; Reutlinger 2016); dimensional analysis (Lange 2009a; Pexton 2014) and inter-theoretic relations (Batterman 2002; Weatherall 2011).

billions of grains of sand on a beach. Bird maintains that while this process *might* add to scientific knowledge, 'it does not add much to understanding. Correspondingly it adds little to scientific progress' (2007:84). I take it to be uncontroversial that noncausal explanations in terms of e.g., conservation laws, kinematics or symmetry principles, represent a much greater cognitive achievement than counting grains of sand. Nonetheless, when combined with a causal analysis of explanation, the epistemic account appears to suggest that these activities are on par insofar as they add little to understanding and thus little (if anything) to scientific progress.

This issue appears to be even worse for the noetic account. According to Dellsén, episodes involving increasing knowledge without increasing understanding provide evidence for his own view: 'where science sees an accumulation of knowledge without also seeing an increase in scientific understanding, there does not seem to be *any* scientific progress' (Dellsén 2016:78 – emphasis added). Consequently, when combined with a causal analysis of explanation, it seems that the noetic account suggests that noncausal explanations involving e.g., statistical analysis, topology or geometry, contribute *nothing* to understanding and thus nothing to scientific progress.

Of course, this is not to say that either Bird or Dellsén actively propose that explanation be characterized purely in inferential or causal terms. Rather, what this discussion shows, is that explanation is more important to our conception of scientific progress than is typically recognized. This is the central motivation behind my proposed "explanation-first" methodology: by putting an analysis of explanation centre-stage in a theory of scientific progress, we can avoid committing to the counterintuitive results discussed above. With inferential and causal accounts of explanation off the table, however, we require an alternative analysis of explanation.

Fortunately, I believe that there is a theory of explanation available which can meet the demands of an analysis of scientific progress. *Interventionism*, popularized by Hitchcock and Woodward (Hitchcock & Woodward 2003a, 2003b; Woodward 2003), is widely recognized to provide a methodology which avoids the pitfalls of both inferential and causal analyses of explanation. In the next section, I shall demonstrate how an interventionist account of explanation can be used in order to characterise scientific progress. On this account, scientists make progress by providing *increasingly invariant* explanations of target phenomenon, and correspondence is achieved when the range of invariance displayed by a progressive theory strictly contains the range displayed by a superseded theory.

V. Interventionism, Invariance and Scientific Progress

According to Hitchcock and Woodward, explanation isn't merely a matter of nomic subsumption or tracing causal history. Rather, explaining involves elucidating patterns of counterfactual dependence, patterns which describe how the behaviour of the explanandum would change under a range of different conditions. An explanation can only support this kind of counterfactual dependence if it is *invariant* under testing interventions. To be invariant, in this sense, an explanatory generalization must 'describe a relationship which holds for certain *hypothetical* values of *X* and *Y possessed by the very object o…* where the value of *X* is changed by an intervention' (Hitchcock & Woodward, 2003a:20).

One of the central motivations behind this interventionist analysis of explanation is discomfort with the traditional conception of laws of nature as *exceptionless* generalizations. According to Hitchcock & Woodward, the demand that laws be exceptionless makes explanation an all-or-nothing affair; a true generalization is either a law in which case it can be used in explanatory derivations, or it is accidental in which case it can't; there are no other

options. In contrast, by cashing out explanatoriness in terms of invariance, we see that '[a]mong those generalizations that are invariant, some will be more invariant than others, and they will correspondingly provide deeper explanations' (Hitchcock & Woodward, 2003b:183-184. Italics added).

According to Woodward (2003), laws of nature are merely those generalizations which are *most* invariant. As a result, law-hood permits of both a "threshold" and (above this) a "continuum". Some generalizations will not be invariant at all (or invariant only under a *very* narrow range of interventions), and fall below the threshold for explanatoriness. However, other generalizations will be invariant under a larger or more significant range of changes, 'so that there will be a continuum of extent of invariance above the threshold' (Woodward, 2013:64). Understanding, on this interventionist picture, can thus be defined in terms of the possession of either knowledge of (in Bird's case), or merely true beliefs about (in Dellsén's case), patterns of counterfactual dependence concerning that phenomenon.

I maintain, however, that the *extent* to which a given explanation increases our understanding ought to be measured in terms of the *range* of a generalization's invariance.[139] The greater this range, the greater the contribution to understanding. This, in turn, allows us to characterise the *rate* of scientific progress; to show 'which additions to knowledge are significant and which are not' (Bird, 2007:85). The greater the range of invariance displayed by a generalization, the more significant the corresponding explanation's contribution to understanding and, as a result, scientific progress.

As an example, let's consider an uncontroversial case of scientific progress: the transition from Newtonian to relativistic mechanics. The laws of Newtonian mechanics are

[139] The *gradability* of understanding, the sense in which explanation and understanding comes in degrees, is emphasized by many, including Kim (1994); Kvanvig (2003, 2009); Khalifa (2013); Wilkenfeld (2013); Kelp (2015).

highly accurate for objects moving at relatively low velocities. When applied to some object with a velocity that is small compared to that of light, such generalizations will remain invariant under a range of interventions *R* on that velocity. As Hitchcock & Woodward note, however, '[t]he special relativistic correction to these laws has two related effects' (2003b:186). First, even though Newton's laws are 'approximately true' with respect to *R*, the corrected generalizations will be *more accurate* within *R*. Second, the latter will be invariant under a *wider range* of interventions *R*,* where *R** strictly contains *R*, but also contains interventions upon objects with velocities closer to that of light.

On an interventionist analysis, the laws of Newtonian mechanics *are* explanatory; they remain invariant under the range *R*, involving interventions on velocities that are relatively small compared to that of light. However, the laws of relativistic mechanics are *more* explanatory because they remain invariant under the wider range of interventions *R**, which includes the range *R*, in addition to interventions involving velocities closer to that of light. Thus, relativistic mechanics represents *significant* progress precisely because it not only increases our knowledge (or true beliefs), but also our *understanding*, insofar as the explanations it provides are *more invariant* than those provided by its predecessor.

In section II., I noted that much of the debate surrounding scientific progress was motivated (initially, at least) by responses to the PMI. As McKenzie characterises it, the problem raised by the PMI is this: 'for us to be able to interpret the temporal sequence of science as an arc of epistemic progress, we need to be able to interpret each change as us learning *more* about the world, not just something *different*' (2020:10). Learning more, in this sense, requires correspondence, since 'where there is correspondence, we have a picture that is, at least in some very important respects, one of retention and refinement – that is, a picture of *progress*' (McKenzie, 2020:12).

I believe that the IEF analysis allows us to characterize this retention and refinement without needing to appeal to the central equations of scientific theories, that is, to the *mathematicised* nature of the language involved. Above, I suggested that relativistic mechanics can naturally be seen as progressive with respect to Newtonian mechanics because the generalizations provided by the former are invariant under a wider range of testing interventions than the latter. Correspondence between these theories can thus be captured in terms of the range of invariance displayed by a superseded theory being *strictly contained* within the range of invariance displayed by a progressive theory.

While the generalizations supplied by Newtonian mechanics are invariant with respect to some range of interventions R, the relativistic corrections to these generalizations provide for explanations that are invariant under a wider range of interventions R^*. Importantly, however, R^* strictly contains R. That is, R^* includes those interventions contained in R (interventions involving velocities that are relatively small compared to that of light) *as well as* additional interventions not contained in R (those involving velocities that are closer to that of light).

This provides us with a natural way of parsing the idea that Newtonian mechanics 'still applies, at least with a high degree of approximation, in those cases in which is was successful' (Popper, 1959:250). That Newtonian mechanics has been superseded by relativistic mechanics, insofar as the latter is more invariant than the former, does nothing to undermine the fact that Newton's laws display a range of invariance which puts them above Woodward's *explanatory threshold*. In this sense, an IEF analysis of progress is able to characterise the important sense in which the explanatory power of Newtonian mechanics is both retained, and expanded upon, by its relativistic successor.

Of course, accounting for correspondence is not the only hurdle facing an explanation-first analysis of scientific progress. In section IV., I argued that when combined with either inferential or causal accounts of explanation, both knowledge and understanding-based accounts of progress produce some highly counterintuitive conclusions concerning what counts as a progressive episode in science. Clearly then, if the interventionist analysis of explanation is to be successfully utilized in characterizing scientific progress, it will need to avoid the issues facing these rivals.

In the next section, I begin by demonstrating how the IEF analysis of scientific progress avoids the issues raised by inferential accounts of explanation: the problem of explanatory asymmetry and the problem of explanatory relevance. I then move on to discuss the core issue facing explicitly causal accounts of explanation; that they fail to do justice to the role of noncausal explanations in facilitating progress. As I admitted at the outset, however, my aim in this paper is not merely to motivate a novel methodology for the analysis of scientific progress, but also to apply this methodology to metaphysics. This presents a particular challenge, insofar as a unifying explanation-first analysis of progress must not only account for both causal and noncausal explanation in science, but also the distinctive noncausal explanation characteristic of metaphysics. Nonetheless, I shall argue that interventionism is up to the task.

VI. Asymmetry, Irrelevance and Noncausal Explanation

Let's start by returning to the problem of explanatory irrelevance. As was highlighted in section IV., on an inferential analysis of explanation, it looks like the consumption of birth control pills *explains* Rishi's failure to become pregnant. This is because we can deduce that Rishi will fail to become pregnant from the fact that he regularly consumes birth controls

pills and the fact that all cisgender men who regularly consume birth control pills fail to become to become pregnant. As a result, when combined with an inferential analysis of explanation, both the epistemic and noetic accounts will mischaracterize such scenarios as making significant contributions to scientific progress.

By adopting an interventionist analysis of explanation, however, we can avoid this counterintuitive commitment, since intervening to prevent Rishi from consuming birth control pills will have no effect upon his chances of becoming pregnant. Consequently, the fact that Rishi regularly consumes birth control pills does not qualify as an explanation of his failure to become pregnant; the corresponding generalization is not invariant under testing interventions. Thus, on an invariantist analysis, the "discovery" that the consumption of birth control pills by cisgender men is correlated with a failure to become pregnant contributes little (if anything) to understanding, and thus little (if anything) to scientific progress.

Equally troubling for inferential accounts of explanation is the problem of explanatory asymmetry. Theresa's discovery that the length of a flagpole's shadow can be deduced from its height appears to represent a much more significant contribution to scientific progress than Boris' discovery of the converse relationship. Despite this, when combined with an inferential analysis, both the epistemic and noetic accounts will mischaracterize Boris' contribution as being equal to that of Theresa, since all that is necessary for explanation, and thus understanding, is nomological-deduction. However, an interventionist analysis of explanation preserves explanatory asymmetry and thus accurately tracks our intuitions about rates of progress in such cases.

This is because, while it is possible to intervene upon the height of a flagpole in order to manipulate the length of its shadow, the converse is not possible. One cannot intervene upon the length of a flagpole's shadow in order to manipulate the height of the flagpole. As a

result, the interventionist can argue that Theresa contributes more to scientific progress precisely because her discovery will remain invariant under a significant range of testing interventions and thus increase our understanding. Boris' discovery, by contrast, is not invariant at all; it contributes nothing to understanding, and thus little (if anything) to progress.

As we also saw in section IV., while causal accounts of explanation are typically thought to avoid the issues raised by inferential analyses, they bring their own. Increasing recognition that many scientific explanations do not cite causes at all, suggests that tracing causal histories cannot be the full story when it comes to the connection between explanation and understanding. As a result, when combined with explicitly causal accounts of explanation, both the epistemic and noetic accounts of progress provide counterintuitive judgements regarding the progressive character of noncausal explanations. On Bird's account, episodes involving noncausal explanation don't contribute as much to scientific progress as they intuitively should and, on Dellsén's account, they contribute nothing at all.

Aside from its success in preserving explanatory asymmetry and rooting out explanatorily irrelevant factors, it is the inherent *flexibility* of interventionism which stands as one of its greatest assets. Indeed, recent years have seen a proliferation of attempts to adapt interventionism in order to characterise noncausal explanation in science. To take just one example, Reutlinger argues that 'causal and noncausal explanations are explanatory by virtue of exhibiting how the explanandum counterfactually depends upon the explanans' (2017:223).[140] More importantly for our purposes, however, several authors have utilized

[140] Also see e.g., Craver (2007a; 2007b); Saatsi & Pexton (2013); Jansson (2015); Saatsi (2016); Reutlinger (2016); Baron *et al* (2017); Baron *et al* (2020); Woodward (2018b); Baron & Colyvan (2021); Emmerson (2021).

Woodward's (2003) interventionist methodology in order to elucidate noncausal explanation in *metaphysics.*[141]

According to Schaffer, for instance, both scientific and metaphysical explanations are characterized by three central roles, each of which points towards the existence of distinctive "laws of metaphysics", minimally understood in analogy with laws of nature as 'counterfactual-supporting general principal[s]' (2017:305).[142] The first such role is to reveal patterns and unify phenomena. Kim, for example, argues that dependence relations 'reduce the number of independent events, states, facts, and properties we need to recognize… Unity and structure go hand in hand; dependence enhances unity by generating structure' (1994:68). As such, the unificatory role of explanation clearly calls for generalizations which are counterfactually robust, in so far as they 'serve to subsume a given case under a more general pattern' (Schaffer, 2017:306). [143]

The second role of scientific and metaphysical explanation can 'be seen as connected to Woodward's (2003) guiding conception of explanations as serving to answer "what if things had been different questions"' (Schaffer, 2017:306). As I have already noted, this "manipulationist" role of explanation also requires explanatory generalizations to be counterfactually robust or, in the parlance of Hitchcock and Woodward, *invariant under testing interventions*. Metaphysical explanations must similarly support an appropriate pattern of counterfactuals, according to Schaffer, since 'it is through counterfactual-supporting

[141] See e.g., Kment (2014); Schaffer (2016); Reutlinger (2017); Wilson (2018a, 2018b); Emmerson (2022b); and Miller & Norton (2022a, 2022b). Interventionism is not the only analysis of scientific explanation which has been repurposed for this task, however. A brief review of the literature reveals accounts of metaphysical explanation which utilize: Hempel's (1965) deductive-nomological framework (e.g., Wilsch 2015, 2016); Kitcher's (1981, 1989) unificationist framework (e.g., Kovacs 2020; Baron & Norton 2021); and the causal-mechanical framework developed by Salmon (1984) (e.g., Trogdon 2018).
[142] Also see Kment (2014) and Wilson (2020).
[143] It is important to note that although Schaffer suggests that such accounts 'have a plausible motivation', he also argues that they are unlikely to be correct, owing to their failure to 'capture the asymmetry of explanation' (2017:306).

generalizations that one can calculate the impact of potential interventions' (Schaffer, 2017:306).[144]

The final role of explanation is that of providing 'a basis for *understanding* the phenomena and so dispel wonderment and offer illumination' (Schaffer, 2017:306 – italics added). Here Schaffer argues, once again, that counterfactual supporting general principles are necessary to make sense of the role of explanation in understanding. According to Baumberger, Beisbart & Brum (2017) for example, knowing that "*p* because *q"* requires both a grasp of the underlying explanatory principal involved, and of counterfactual variations. Similarly, Grimm (2010) argues that understanding requires the ability to apply a law to both actual and counterfactual cases, and explicitly connects explanatory understanding with Woodward's (2003) interventionist methodology.

I believe that Shaffer's interventionist characterisation of metaphysical explanation provides us with all the tools necessary to motivate an analogous explanation-first analysis of progress in metaphysics. In the final section of this paper, I shall argue that metaphysics makes progress in precisely the same way as science: when metaphysicians provide increasingly invariant explanations of target phenomenon. What's more, this unifying analysis of progress allows us to characterise an analogous notion of *correspondence* which applies across scientific and metaphysical instances. Correspondence, in this sense, is achieved when the range of invariance displayed by a progressive theory, be it scientific or metaphysical, strictly contains the range of invariance displayed by a superseded theory.

[144] It is worth highting that some, recently labelled "intervention-puritans" by Emmerson (2021), are unconvinced that interventions can play a substantive role in characterizing noncausal explanations (see e.g., Saatsi & Pexton 2013; Jansson 2015; Reutlinger 2016, 2017; French and Saatsi 2018; Lange 2019; Khalifa *et al* 2020). For the purposes of this paper, however, I shall adopt the opposing position "intervention-liberalism" (Emmerson 2021) and assume that at least *some* forms of noncausal explanation (including metaphysical explanation) *can* be successfully analysed in terms of interventions. Having said this, I expect that much of what follows could be cashed out in terms of non-interventionist counterfactuals without any significant loss of content.

Contra McKenzie (2020), the IEF account is thus able to capture the important sense in which a metaphysical theory can both retain, and expand upon, the explanatory power of its predecessor(s). In order show this, however, we will need to locate a test case of metaphysical progress. Of course, given scepticism regarding the very *possibility* of progress in metaphysics, this might seem like a fool's errand. However, McKenzie herself welcomes this challenge: 'I invite anyone to find counterexamples to the way I say metaphysics typically works' (2020:8). My chosen counterexample, concerning competing metaphysical explanations of the identity and distinctness of concrete objects is, I believe, as close to uncontroversial as it is possible to come.

VII. Interventionism, Invariance and Metaphysical Progress

In "Explaining Identity and Distinctness", Erica Shumener (2020) attempts to provide a novel metaphysical explanation of the identity and distinctness of concrete objects. Before presenting her own account in terms of "quantitative properties" Shumener highlights two prior proposals which are now widely regarded as inadequate.[145] The first, the *qualitative properties proposal* (*QPP*) suggests that identity facts of the form [$x = y$] are explained by the fact that x and y share all of their qualitative properties (e.g., Black 1952; Rocca 2005).[146] The second, the *weak discernibility proposal* (WDP) suggests that such identity facts are explained by the fact that x and y stand in only *reflexive* relations to one another (e.g., Saunders 2006).

[145] Unfortunately, a critical analysis of Shumener's *quantitative properties proposal* is beyond the scope of this paper. My interest in Shumener (2020) is not in the novelty of their theory, but rather in the dialectical progression of debate surrounding the identity and distinctness of concrete objects; an intuitive interpretation of which can be given in terms of increasingly invariant metaphysical explanations.

[146] Shumener suggests that qualitative properties are those that 'do not involve the identity relation or involve specific relations. So, for example, *5km mass, adjacent to, same colors as* are qualitative features' (2020:2079).

(a) **The qualitative properties proposal**: *for any objects x and y, if x and y share all of their qualitative features, then x is identical to y; and if x has some qualitative feature that y lacks, then x and y are distinct.*

(b) **The weak discernibility proposal**: *for any objects x and y, if x and y only stand in reflexive relations to one another, then x is identical to y; and if x stands in an irreflexive relation to y, then x and y are distinct.*

By adopting an interventionist analysis of explanation, we can provide a satisfying account of why the *WDP* ought to be considered progressive with respect to the *QPP*. Like Newtonian mechanics, the *QPP* is explanatory insofar as it remains invariant under a certain range of interventions, *R*. Nonetheless, like relativistic mechanics, the *WDP* is invariant under a much wider range of interventions *R**. In this sense, both proposals contribute to understanding. However, since the *WDP* is *more invariant*, it provides greater understanding of the identity and distinctness of concrete objects than the *QPP*. As a result, I maintain that the transition from the *QPP* to the *WDP* ought to be considered a significant progressive step in the history of metaphysics.

Our first step is to establish the range of invariance for the *QPP*; this range will include counterfactual scenarios in which objects sharing all of their qualitative properties are identical, and objects possessing different qualitative properties are distinct. However, in order to show that the *WDP* is progressive with respect to the *QPP*, we then need to establish that the former has a wider range of invariance than the former. To do this requires a counterfactual scenario which demonstrates that it's possible for qualitatively identical objects to be numerically distinct. Additionally, however, this scenario must be one in which the objects involved stand in only reflexive relations to one another, thus securing the invariance of the generalization specified by the *WDP*.

Fortunately, Black (1952) has popularized just such a scenario. First, imagine a possible world containing only two spatially separated objects, *A* and *B*, which possess *different* qualitative properties. Let us suppose, for the sake of argument, that *A* is spherical, while *B* is cuboid. It appears that the *QPP* does provide an explanation of the distinctness of *A* and *B* here. (a) suggests that if *A* and *B* share all of their qualitative properties, then they are identical; and that if *A* has some quality which *B* lacks, then they are distinct. Since *A* possesses the quality 'being spherical', which *B* lacks, (a) will be invariant under interventions resulting in qualitatively discernible, spatially separated objects.

Now imagine that we intervene upon *A* or *B* (or both), altering them to ensure that they now share all of their qualitative properties. They are, in other words, indistinguishable in terms of their qualitative properties (size, shape, mass etc). Under such an intervention, (a) is violated. Since *A* no longer possess any qualitative property which *B* lacks, according to the *QPP*, *A* and *B* are identical. Yet, given that the objects in his case remain spatially separated, we know them to be distinct. As such, (a) is *not* invariant under testing interventions resulting in qualitatively indiscernible, spatially separated objects.

However, as Saunders (2006) argues, the generalization specified by the *WDP* will remain invariant in such cases. According to (b), *A* and *B* would be identical if they only stood in *reflexive* relations to one another, and distinct if *A* stood in at least one *irreflexive* relation to *B*. While all of the qualitative relations which *A* and *B* stand in are reflexive, 'the spheres [also] stand in irreflexive relations like *five meters away from* to one another' (Shumener, 2020:2080). Consequently, it appears that (b) remains invariant under testing

interventions which violate (a); those resulting in qualitatively *in*discernible, spatially separated objects.[147]

In section V., I argued that interventionism provides us with a natural way of analysing the connection between explanation and understanding. To understand a phenomenon, on this picture, is to have knowledge of, or merely true beliefs about, patterns of counterfactual dependence concerning that phenomenon. This analysis allows us to characterise the *rate* of scientific progress, since the extent to which a given explanation increases our understanding can be defined in terms of the extent to which it is more invariant than its predecessor(s). The greater the range of *additional* invariance an explanation displays, the greater its novel contribution understanding and thus the greater its contribution to scientific progress.

What's more, in the previous section, I noted that Schaffer's (2017) interventionist account of metaphysical explanation provides us with all of the tools necessary in order to define an analogous notion of progress in metaphysics; something which McKenzie (2020) argues is impossible. I agree with McKenzie on the importance of lessons learned from the PMI; that progressive theories must be able to capture the explanatory potential of their predecessors, while also adding novel explanatory power of their own. Nonetheless, I have argued that this necessary component of scientific progress need not be cashed out in terms of "approximation" or the "central equations" of the theories involved. Rather, by adopting either a knowledge- or understanding-based conception of scientific progress, and an

[147] It is important to note that I do not take the *WDP* to be the most invariant explanation of the identity and distinctness of concrete objects available. My claim is only that the WDP is more invariant than the QPP. Indeed, as Shumener herself notes, the WDP will not be invariant under counterfactual scenarios involving 'co-located, qualitatively indiscernible objects' (2020:2081). Shumener's own *quantitative properties proposal* is supposedly able to cope with such cases. If so, *it* will possess a wider range of invariance than both the *QPP* and the *WDP*. On my IEF analysis of progress, this would make Shumener's proposal progressive with respect to both prior proposals. However, this does not undermine my core argument that the transition from the *QPP* to the *WDP* is itself a progressive step.

interventionist account of explanation, we can make sense of correspondence in terms of *invariance*.

According to both Bird's (2007) epistemic account and Dellsén's (2016) noetic account of scientific progress, explanation-involving episodes constitute the most significant examples of scientific progress. In this sense, relativistic mechanics can naturally be seen as significant progress with respect to Newtonian mechanics because the generalizations provided by the former are invariant under a wider range of testing interventions than the latter. What's more, the *continuity* between these theories can be captured by appealing to the ranges of invariance they display. While Newtonian mechanics is invariant under the range *R*, relativistic mechanics is invariant under the significantly wider range *R**. Since *R* strictly contains R*, invariance provides us with a natural way of capturing the sense in which the former theory tells us *more* about the world than the latter; not merely something *different*.

In this section, I have attempted to apply this methodology to a test-case of progress in the history of metaphysics. I argued that the *QPP* is invariant with respect to a range of interventions *R*, where *R* includes interventions resulting in qualitatively discernible, spatially separated objects. Like Newtonian mechanics, the extent of this range places the *QPP* above Woodward's (2003) *threshold* for explanatoriness. Nonetheless, I have also argued that the *WDP* is invariant under the wider range of interventions *R**, where *R* additionally* includes interventions resulting in qualitatively indiscernible, spatially separated objects.

As a result, on the unifying IEF account of progress, the *WDP* ought to be considered progressive with respect to the *QPP* for precisely the same reason that relativistic mechanics ought to be considered progressive with respect to Newtonian mechanics. The *WDP* is invariant under a wider range of interventions, and thus provides greater understanding of the identity and distinctness of concrete objects than the *QPP*. What's more, however, since the

range of interventions under which the *WDP* remains invariant *strictly contains* the range of interventions under which the *QPP* remains invariant, these theories *correspond*; the explanatory power of the *QPP* is both retained and expanded upon by the *WDP*.

Concluding Remarks:

At the outset of this paper, I highlighted three interrelated goals:

1. To motivate a novel *explanation-first* approach to analysing scientific progress.
2. To defend a specific *interventionist* variant of this approach.
3. To show that this approach can be utilized to characterize progress in *metaphysics.*

In sections III. and IV., I tackled the first of these goals. In section III., I argued that despite agreeing on the importance of the role of explanation in facilitating progress, recent accounts of scientific progress fail to offer a substantive analysis of how or why explanation contributes to progress. In section IV., I demonstrated the importance of such an analysis, by showing that, when combined with either inferential or causal analyses of explanation, both knowledge- and understanding-based conceptions of progress are rendered untenable.

In sections V. and VI., I moved on to the second goal. In section V., I defended an *interventionist* explanation-first approach to characterizing progress. On this account, scientists make progress by providing increasingly invariant explanations of target phenomena. Further, I argued that correspondence between theories can be cashed out in terms of the range of invariance displayed by a suspended theory being strictly contained by the range of invariance displayed by a progressive theory. In section VI., I demonstrated that this interventionist explanation-first analysis of progress avoids the problems facing rival inferential and causal accounts.

Finally, in the previous section, I addressed the third goal, by applying the interventionist explanation-first account of progress to a case-study concerning two rival explanations of the identity and distinctness of concrete objects. In so doing, I argued that progress in this case can be characterized in precisely the same terms as scientific progress. On this *unifying* explanation-first analysis, both scientists and metaphysicians make progress by providing increasingly invariant explanations of target phenomena, and correspondence is achieved when the range of invariance displayed by a superseded theory is *strictly* contained within the range of invariance displayed by a progressive theory.

Chapter 6

Moral Invariantism

This paper provides a novel account of moral principles. I argue that morality can be principled, even if there are no exceptionless generalizations governing moral inquiry. In this respect, my analysis goes against the grain of current thinking in metaethics. Traditional debate concerning the existence of moral principles has been dominated by two opposing positions. On the one hand, moral generalists maintain that moral principles, characterized as exceptionless generalizations, play an integral role in normative explanation. On the other, moral particularists argue that exceptionlessness is an unattainable goal and, as a result, that such principles can play no part in normative inquiry. I shall argue, however, that moral principles do have an important role to play in normative explanation (contra particularism) even though such principles are not best understood as being exceptionless (contra generalism). Rather, I propose that moral principles are those explanatory generalizations which are most invariant. I call the resulting view "moral invariantism".

I. Introduction

Is morality principled? Debate surrounding this question had been dominated by two opposing positions. On the one hand, moral generalists maintain that morality *is* principled, that exceptionless generalizations provide the truth conditions for the application of a moral concept. As McKeever & Ridge highlight, 'the history of moral philosophy is in large part a history of attempts to map the moral landscape with a set of principles' (2006:4). On the other, moral particularists maintain that morality *isn't* principled, that normative

considerations are too complex for there to be principles, 'even very complicated ones… capable of codifying the moral landscape' (Little, 2000:277).[148]

In this paper, however, I shall argue that this debate is founded upon a false equivocation, and that morality can be principled (*contra* particularism) even if there are no exceptionless generalizations governing moral inquiry (*contra* generalism). In order to do so, I draw upon an analogous debate concerning the role of *laws of nature* in science. For most of the last century, laws of nature were understood in generalist terms, as entirely exceptionless generalizations. However, it is now widely recognized that exceptionlessness is neither necessary nor sufficient for law-hood; not all explanatory generalizations are exceptionless, and not all exceptionless generalizations are explanatory.

In response, some have argued that it is *invariance*, rather than exceptionlessness, which ought to be considered the mark of law-hood.[149] According to Woodward, for example, there is 'nothing more to lawfulness than, so to speak, *de facto* invariance under some appropriately large range of changes in initial/boundary conditions, including changes involving interventions' (2013:65). Laws of nature, on this invariantist analysis, are merely those explanatory generalizations that fall at the upper end of this range. It is my belief that similar considerations ought to motivate an analogous position with respect moral principles; a view I call "moral invariantism".

---

[148] Ewing (1929) and Ross (1930) are sometimes identified as the fathers of moral particularism, however Dancy (1981, 1993, 2004, 2009) is the name most closely associated with the position. For more recent articulations (although not necessarily *defences*) of particularism see e.g., Hooker & Little (2000); Lance & Little (2004, 2008); Kompa (2004); Strangl (2006); Raz (2006); Väyrynen (2006, 2008); Gleeson (2007); Hooker (2008); Salay (2008); Strahovnik (2008). Moral generalism has typically been *assumed* rather than explicitly defended. Having said this, in his first attempt at articulating particularism, several figures are highlighted by Dancy (1981) as proponents generalism, including: Hare (1952); Singer (1961); Kovesi (1967); and Swinburne (1976).

[149] See e.g., Mitchell (1997, 2000); Nozick (2001); Hitchcock & Woodward (2003a, 2003b); Woodward (2003, 2013, 2018a); and Lange (2009b); Sher (2021).

The structure of this paper is as follows. In the next section I outline the motivation behind generalism with respect to laws of nature and, in section III, I show that remarkably similar considerations have motivated moral generalism. In both cases, generalists have been forced to concede that counterfactual robustness and *ceteris paribus* conditions are required if laws or principles are to be understood in terms of exceptionless generalizations. In section IV., I show that a series of problems, originally raised by Hitchcock & Woodward (2003a, 2003b) and Woodward (2003, 2013, 2018) and targeted at generalism with respect to laws of nature, also present intractable difficulties for moral generalism.

In section V., I motivate the invariantist analysis of laws of nature and argue that, when applied to the normative domain, it allows us to avoid the pitfalls of moral generalism. On my account, a generalization explains the rightness or wrongness of an action to the extent that it is invariant, and moral principles are merely the *most* invariant of these explanatory generalizations. In the final section, I discuss a natural objection to my moral invariantist proposal: that it makes identifying moral principles a matter of pragmatics. However, I argue that this is a feature, rather than a bug of my account and provide two useful heuristics to help us delineate nonexplanatory generalizations, explanatory generalizations, and moral principles. I also briefly discuss the *taxonomy* of moral invariantism and how the view is importantly distinct from both moral generalism and moral particularism.

II. Scientific Generalism

During the 20th century, debate surrounding the nature of scientific explanation was dominated by a collection of views committed to scientific generalism. The most popular such accounts, Hempel's (1965) deductive-nomological model and unificationist accounts of Friedman (1974) and Kitcher (1981), provide analyses of explanation as *inference*, carrying

little in the way of serious ontological commitment.[150] While the unificationist framework is widely regarded to be the more sophisticated of the two, it nonetheless preserves the central features of its predecessor: deductive argument and subsumption under laws.[151]

On both accounts, laws are taken to be exceptionless generalizations; quantified conditional claims of the form "All *A*s are *B*s". However, even generalists typically concede that exceptionlessness is not a sufficient condition of law-hood. Not just any old exceptionless generalization can qualify as a law of nature because not all exceptionless generalizations can be used in explanatory derivations. These "accidental" generalizations must be prevented from infiltrating our stock of laws. Take the following, for example:

CYLING: all members of the Eritrean National Cycling Team have black hair.

CYCLING is clearly not a law of nature. While it is true that all members of the Eritrean National Cycling Team have back hair, this fact cannot be used to explain why an individual member, say, Biniam Girmay, has black hair. Girmay does not have black hair *because* he is a member of the Eritrean National Cycling Team. According to the generalist, what separates explanatory generalizations (the laws) from nonexplanatory generalizations, is that the former will apply across a wide range of different cases. In other words, to qualify as a law of nature, a generalization must be both exceptionless *and* counterfactually robust.

Quantified conditionals, of the sort taken by the generalist to be indicative of laws, support what Woodward (2003) calls "other object counterfactuals" (OOCFs): 'If some object o*, different from *o* and that does not possess property *A,* were to be an *A*, then it

[150] Explanation enables us to *understand why* some phenomenon occurs, according to Hempel, because 'given the particular circumstances and the laws in question, the occurrence of the phenomenon *was to be expected*' (1965:337). Kitcher (1981), however, argues that behind this "official model" of explanation in terms of nomic expectation, there is an "unofficial model" of explanation in terms of *unification*. See Hempel (1966:83) for evidence in support of Kitcher's reading.

[151] See e.g., de Regt (2017:49-58); Woodward (2003:265-268).

would be a *B'* (2003:281). **CYCLING** is *not* counterfactually robust in this sense.[152] Suppose, for example, that Mathieu van de Poel, a blonde-haired member of the Dutch National Cycling Team, were to join the Eritrean National Cycling Team. All else being equal, would van de Poel now have black hair? The intuitive answer is "no". What this tells us, according to the scientific generalist, is that **CYCLING** is merely accidental, and can thus be eliminated as a candidate law.

However, the existence of nonexplanatory exceptionless generalizations is not the only difficulty facing the scientific generalist. Indeed, it has long been observed that many explanatory generalizations used in branches of the life and social sciences are *not* exceptionless. Consider the following example concerning Mendel's "law" of segregation:

> **MENDEL:** 'with respect to each pair of genes in a sexual organism, 50% of the organism's gametes will carry one representative of that pair and 50% will carry the other representative of that pair' (Beatty, 1995:50-51).

Somewhat problematically for the generalist, not all sexually reproducing organisms conform to this generalization. This is because of *meiotic drive*, a phenomenon whereby an allele influences meiosis in such a way that it has a greater that 50% chance of ending up in one gamete rather than the other. If we understand laws in the generalist sense, then **MENDEL** does not qualify and thus cannot be used to explain.

To make matters worse, this problem extends beyond the confines of the "special" sciences. As Woodward highlights, for example, 'Maxwell's equations break down under conditions under which quantum mechanical effects become important, general relativity is widely believed to require correction at very small length scales (Plank length), and so on'

[152] As Woodward notes 'Exactly what support means and exactly how the counterfactuals… are to be interpreted is typically left unclear, but in practice, the requirement is often interpreted in such a way that [a generalization] counts as a law if we can find some true counterfactual (or perhaps some small set of true counterfactuals) to associate with it' (2003:279).

(2013:62). If generalism requires that laws be exceptionless, a condition which cannot be met in many instances of scientific explanation, then we appear to have two options: we must either abandon generalism; or accept that a much of our best science isn't explanatory after all.

However, in response to this dilemma, some scientific generalists have argued that we can characterise the importance of exceptionlessness generalizations, as well as the explanatory function of generalizations that permit of exceptions, by appealing to *ceteris paribus* laws. The basic idea here is that where a generalization of the form "All *A*s are *B*s" has exceptions, it can nonetheless be regarded as explanatory, as long as there is some further set of conditions, a 'completer', under which it *is* exceptionless: i.e., "All *A*s in *C* are *B*s" (Woodward, 2003:307).

In the case of **MENDEL**, *C* will be a list of additional conditions which guarantee the exceptionlessness of Mendel's law: that the system in question is capable of sexual reproduction, that meiotic drive doesn't occur, and other factors don't disrupt gamete production *etc*. While **MENDEL** itself isn't exceptionless, generalists argue that it still counts as a law, and can thus still be used to explain, because it's closely associated with this "*ceteris paribus* law of segregation", which is genuinely exceptionless.

Consequently, the scientific generalist appears to be able to have their cake and eat it too, insofar as they can capture both: the importance of exceptionlessness for law-hood; and the explanatory role of generalizations that permit of exceptions. While some generalizations are not exceptionless in the strict sense, they will nonetheless qualify as laws of nature, so long they are associated with, or "backed" by, genuinely exceptionless *ceteris paribus* laws. In the next section, I show that remarkably similar considerations have motivated *moral*

generalism, with many having adopted an analogous *ceteris paribus* strategy in response to particularist's concerns regarding exceptionable moral principles.

III. Generalism About Moral Principles

While moral generalism has typically been assumed, rather than explicitly defended, it can nonetheless be motivated by analogous considerations to scientific generalism, concerning the role of moral principles in normative explanation.[153] DePaul (1987), for example, notes that a key motivation for constructing moral theories is a desire to know why certain actions are right, and others wrong; to *explain* the normative characteristics of action. We want to know 'what *makes* an act obligatory or a person evil, and for this the connection between moral and non-moral properties must be strong indeed' (DePaul, 1987:427). The strength of this connection was also historically thought to derive from a combination of exceptionlessness and counterfactual robustness.

Suppose that David fails to declare his income to the tax office. Assuming that we can agree that this act is wrong, how can we explain this wrongness? According to Lance and Little, the core thesis of moral generalism states that 'generalizations must be exceptionless if they are to do genuine and fundamental explanatory work' (2008:54). Supposing then, that David's action is wrong because it is deceitful, the moral generalist maintains that we require an exceptionless generalization to allow us to infer the normative claim that David's action is wrong from the non-normative fact that David failed to declare his income. Something like:

DECEPTION: All acts of deception are wrong.

[153] As McKeever & Ridge attest, 'the basic presupposition that morality can and should (indeed, on some accounts, must) be captured by a set of principles has not so much been argued for as assumed' (2006:4).

However, moral generalists are also cognizant of the fact that exceptionlessness *alone* appears to be insufficient to secure the explanatory character of moral principles. Indeed, it has long been maintained that, in order to qualify as a moral principle, a generalization must also be counterfactually robust:

> 'Hence to give a reason in support of the judgement that a certain individual, A, ought or has the right to do some act, presupposes that anyone with the characteristics specified in the statement of the reason ought or has the right to do the same kind of act in a situation of the kind specified' (Singer, 1961:24).[154]

In terms of OOCFs, what makes **DECEPTION** explanatory, on the generalist picture, is that we can use it to infer that it would be wrong for *anyone* to perform the same act. The explanatory power of **DECEPTION** thus derives from the fact that it will continue to hold, no matter who we replace David with. From the fact that Theresa, Boris or Liz failed to declare their own income to the tax office, for example, we can use **DECEPTION** to infer that any of these actions would also be wrong.[155]

According to moral *particularists*, however, moral principles can play no role in moral theorizing. As Kirchin characterizes it, the core claim of particularism is this: 'what can be a reason that helps make one action right [or wrong] need not be a reason that always helps to make actions right [or wrong]' (2007:9). Particularists tell us that the *valency*, or the various ways in which non-moral features of a situation combine to make something morally

[154] Also see e.g., Hare (1962); Kovesi (1967); and Swinburne (1976).

[155] It is interesting to note, however, that the *reason* that such generalizations are thought to require counterfactual robustness varies across scientific and moral generalism. In the case of scientific generalism, counterfactual robustness is thought to rule out the possibility of accidental generalizations making it into our stock of laws. In the case of moral generalism, the counterfactual robustness condition is thought to ensure that moral principles apply *universally*; if **DECEPTION** is a moral principle, then it ought to apply to you just as much as it applies to me (See Dancy, 1981:378). Nothing argued in this paper will hang on this distinction.

relevant, is 'too complex and sensitive to context to be captured even in principles concerning how morality works' (Väyrynen 2011:6).[156]

Suppose, for example, that Theresa has a legitimate concern that the state will use her tax contribution to suppress dissent or otherwise mistreat her fellow citizens. David, on the other hand, has no such concerns, but wants to use the money he would save to finance a second home. In this case, it appears that the very same action can be both right and wrong, depending on *circumstance*. As such, while **DECEPTION** appears to explain the wrongness of David's action, it does not capture our converse judgement regarding Theresa's action. In other words, it appears that **DECEPTION** is *not* exceptionless and thus cannot qualify as a moral principle on the generalist account.

This argument from the complexity of moral reasoning is typically referred to as the "holism of reasons" and is widely regarded by particularists to be the key motivation for the rejection of generalism.[157] Despite this, several authors have argued that this move is unsound, and that generalism is entirely compatible with the context sensitivity of moral reasoning. According to McKeever & Ridge (2006), for example, while the holism of reasons certainly makes moral theorizing more complicated, it doesn't undermine exceptionlessness as a *regulative ideal*. This is because, as Väyrynen explains, 'the examples in support of holism are ineffective because they specify reasons incompletely. Full reasons for action include background conditions which holism classifies as defeaters and enablers' (2011:11).[158]

---

[156] For similar explications see e.g., Väyrynen (2006, 2011); Hooker (2008); Dancy (2009); Gleeson (2007); and Flynn (2009).

[157] Dancy, for example, suggests that '[a] principle-based approach to ethics is inconsistent with the holism of reasons' (2000:135); and Little maintains that 'if reason-giving considerations function holistically in the moral realm then we simply shouldn't expect to find rules that mark out in nonmoral terms the sufficiency conditions for applying moral concepts' (2000:284).

[158] I take it that Woodward's notion of a "completer" is equivalent to the combination of "defeaters" and "enablers". In short, these are clauses which specify the conditions under which a given generalization *would be*

There are two different ways in which moral principles can be brought into line with the holism of reasons in this way. The first strategy is to pursue *unhedged* principles, which specify 'a complete list of the requisite qualifications and exceptions, and thus to give at least contributory principles which hold without exception' (Väyrynen, 2011:14). The second is to allow that such a list might be open-ended, and so to simply quantify over possible complications. This method results in what Väyrynen calls "hedged" principles, and what McKeever & Ridge call "default" principles. As the latter elaborate, '[r]ather than trying to list all of the possible defeaters and countervailing reasons, the antecedent… quantifies over them and claims that none is present' (McKeever & Ridge, 2006:119).

In this way, the moral generalist also appears to be able to both have, and eat, their cake. Like the scientific generalist, by including *ceteris paribus* conditions in their characterization of moral principles, the moral generalist hopes to capture the importance of exceptionlessness while, at the same time, allowing generalizations which are not exceptionless to play a central role in normative explanation. Indeed, McKeever & Ridge draw an explicit comparison between their own thesis and that of Pietroski & Ray (1995), who 'have independently developed an account of *ceteris paribus* laws in science… that is in many ways similar to our account of default principles in ethics' (McKeever & Ridge, 2006:123, fn. 6).

However, as I noted at the outset, not everyone is convinced that use of *ceteris paribus* conditions can save the generalist conception of laws. Indeed, Hitchcock & Woodward (2003a, 2003b) and Woodward (2003, 2013, 2018) argue that the very idea that exceptionlessness functions as a guide to explanatoriness is fundamentally mistaken. Despite generalists' attempts to use counterfactual robustness and *ceteris paribus* conditions to

---

exceptionless. I do not think that anything important hangs on this distinction, so I will use "completers" from here on out to avoid unnecessary complications.

strengthen the connection between exceptionlessness and law-hood, Hitchcock and Woodward ultimately conclude that we must abandon the idea that laws of nature are exceptionless generalization *altogether*. In the next section, I outline their motivation for this claim, and argue that remarkably similar considerations can be applied to generalism about moral principles.

IV. Exceptionlessness as a Guide to Explanatoriness

According to Woodward, there is a fundamental and unassailable problem with the idea that generalizations permitting of exceptions can nonetheless qualify for law-hood provided there exists a closely associated *ceteris paribus* law/principle which *is* exceptionless. The issue is this: we can easily identify *ceteris paribus* conditions for generalizations which are *not* explanatory. As such, the distinction between those generalizations that have "completers", "defeater" or "enablers", and those that don't, fails to track the distinction between explanatory and nonexplanatory generalizations.

Consider the following example of Woodward's (2003:309)

> **ACCENT**: All human beings with normal neurophysiological equipment speak English with a southern U.S. accent.

Clearly, this generalization has exceptions. I, for one, am a human being with normal neurophysiological equipment, but I don't speak English with a southern U.S. accent. However, Woodward notes that there will be a very complicated set of environmental conditions "*K*" which, when combined with being a human with the appropriate neurophysiological structures, 'are nomologically sufficient to ensure that one will learn to speak English with a southern accent.' (2003:309-310).

> **ACCENT$_K$**: All human beings with normal neurophysiological equipment, in *K*, speak English with a southern U.S. accent.

Consequently, it appears that the generalist is committed to accepting **ACCENT** as a genuine law of nature, since it's appropriately related to a *ceteris paribus* law, **ACCENT$_K$**, which is exceptionless. What's more, the same considerations suggest that the generalist will also be so committed with respect to other nonexplanatory generalizations like "All human beings speak Chinese" and "All human beings speak Urdu", *etc*. Such generalizations are, as Woodward argues, not plausible candidates for laws 'of any sort, ceteris paribus or otherwise' (2003:310). The moral generalist faces a similar problem.

On the face of it, the idea that it is always wrong to donate money to not-for-profit organizations seems absurd, as does idea that an organization's *being* not-for-profit can be a reason not to donate money to them. While there may well be not-for-profit organizations whose work is morally dubious, it is undeniable that many others perform an invaluable service, which justifies their patrons' support. However, the moral generalist appears to be committed to accepting the following generalization as a moral principle:

> **DONATION**: all acts which constitute donating money to a not-for-profit organization are wrong.

Suppose there exists a not-for-profit organization whose sole purpose is to promote bear baiting. Let's call this organization "the World Bear Baiting Alliance" (WBBA). I would hope that we could all agree that anyone who donates money to such an organization acts immorally.[159] As a result, the following *ceteris paribus* principle appears to be genuinely exceptionless:

[159] This is presuming that the act was not coerced in any way or was not performed in pursuit of some higher moral ideal (etc). However, such factors could, of course, be included in our list of *ceteris paribus* conditions, or simply quantified over and supposed not to have been the case.

> **DONATION$_{WBBA}$**: all (uncoerced *etc*) acts which constitute donating money to a not-for-profit organization, where the organization in question is the WBBA, are wrong.

All that is required for **DONATION** to qualify as a moral principle, according to the generalist, is that there be a nearby *ceteris paribus* principle, which *is* genuinely exceptionless. **DONATION$_{WBBA}$** appears to meet this requirement. The only obvious response available to the generalist here, is to appeal to counterfactual robustness.

As we saw in section II., exceptionlessness is an insufficient condition for law-hood; not all exceptionless generalizations are explanatory. According to the generalist, however, we can use counterfactual robustness to sort exceptionless generalizations that are not explanatory (the merely accidental), from those exceptionless generalization that are explanatory (the laws). Thus, if we can show that **ACCENT$_K$** and **DONATION$_{WBBA}$** fail to hold under an appropriate range of OOCFs, then the generalist can discount them, and thus **ACCENT** and **DONATION**, as candidate laws/principles.

The problem with this line of argument is that OOCFs do not track explanatoriness; they track exceptionlessness. Where a generalization fails to hold under OOCFs, this is precisely *because* it permits of exceptions. Consider **CYCLING** once again. On a generalist analysis, this generalisation does not qualify as a law because it fails to support an appropriate range of OOCFs: were van de Poel to become a member of the Eritrean National Cycling Team, he would not, as a result, have black hair. However, to say that van de Poel would not have black hair as a result of joining the Eritrean National Cycling Team, is just to say that **CYCLING** is not genuinely exceptionless.

The full force of this problem might not be immediately apparent. After all, **CYCLING** is clearly not a law of nature. Whether **CYCLING** fails to qualify as a law because it has exceptions, or because it isn't counterfactually robust, seems irrelevant. Either

way, the generalist gets *this* call right. However, the scale of the problem presented by the fact that OOCFs do not track explanatoriness can be seen by considering *genuinely* exceptionless generalizations which nonetheless fail to be explanatory. Here is an example adapted from Salmon (1971):

> **PREGNANT:** All cisgender men who regularly consume birth control pills fail to become pregnant.

Clearly, **PREGNANT** is not an explanatory generalization. The fact that Jacob, a cisgender man, fails to become pregnant is entirely unconnected to the fact that he regularly consumes birth control pills. Nonetheless, **PREGNANT** *is* counterfactually robust in the requisite sense. Indeed, were *any* cisgender man to regularly consume birth control pills, then they would fail to become pregnant. As a result, it appears that the scientific generalist is committed to accepting **PREGNANT** as a law of nature, even though it is not explanatory.

The same appears to be true of both $\textbf{ACCENT}_{\textbf{K}}$ and $\textbf{DONATION}_{\textbf{WBBA}}$. As noted above, despite the fact that I am a human being with the appropriate neurophysiological structures, I do not speak English with a southern U.S. accent. However, *were* it the case that the very complicated set of environmental conditions "*K*" applied to me, then I would speak English with a southern U.S. accent. Indeed, by stipulation, were *any* human being with normal neurophysiological equipment to meet these conditions, then they would speak English with a southern U.S. accent.

Similarly, were *anyone* to donate money to a not-for-profit organization, where the organization in question is the WBBA, their action would be wrong. As a result, $\textbf{ACCENT}_{\textbf{K}}$ and $\textbf{DONATION}_{\textbf{WBBA}}$ appear to be counterfactually robust in just the sense required to qualify as a law/principle on a generalist account. While OOCFs can help establish whether a generalization permits of exceptions, this begs the question when it comes to the connection

between exceptionlessness and explanatoriness; all genuinely exceptionless generalizations are also robust under OOCFs, regardless of their explanatory content.

The key issue with OOCFs, according to Woodward (2003), is that, to the extent that we can understand such counterfactuals at all, they concern changes in the identity of the explananda, and this simply isn't a factor upon which the explanandum *depends*. To see this, consider an alternative candidate explanation for Jacob's failure to become pregnant:

PREGNANT$_U$: All cisgender men who lack a uterus fail to become pregnant.

PREGNANT$_U$ provides a better explanation than **PREGNANT**, because it successfully identifies *a* factor upon which pregnancy depends: the possession of a uterus.[160] Were it the case, as the generalists appear to suggest, that OOCFs track explanatoriness, then we ought to be able to draw a meaningful distinction between these generalizations in terms of their robustness under OOCFs. However, **PREGNANT** and **PREGNANT$_U$** are equally counterfactually robust in this sense. *Any* cisgender man who either, regularly consumes birth control pills, or lacks a uterus, will invariably fail to become pregnant.

As Hitchcock & Woodward describe it, the problem here is that 'the traditional distinction involves an exhaustive dichotomy of true generalizations–a true generalization is either a law, in which case it is explanatory, or it is accidental, in which case it is not explanatory. There are no other options' (Hitchcock & Woodward, 2003b:183).[161] Consequently, not only must the scientific generalist accept both **PREGNANT** and **PREGNANT$_U$** as laws of nature, but they must also concede that these generalizations do an

[160] Which is not to say that **PREGNANT$_U$** provides the *best* possible explanation by any means.
[161] Kim (1994) makes a similar point.

equally good job of explaining why Jacob, or any other cisgender man, fails to become pregnant.

Once again, the same appears to be true of moral generalism. As I noted above, that the WBBA is not-for-profit is *not* a reason that donating money to them would be wrong. While there is clearly *something* wrong with donating money to such an organization, there are several seemingly viable candidate explanations of this fact. Here are two examples:

> **DONATION$_{\text{BEARS}}$**: All acts which constitute donating money to an organization, whose sole purpose is to encourage the mistreatment of bears, are wrong.
>
> **DONATION$_{\text{SUFFERING}}$**: All acts which constitute donating money to an organization, whose sole purpose is to encourage the mistreatment of beings capable of suffering, are wrong.

Assuming, as is necessary on a generalist picture, that both explanatory generalizations are exceptionless, **DONATION$_{\text{BEARS}}$** and **DONATION$_{\text{SUFFERING}}$** are equally counterfactually robust. Were *anyone* to donate money to either an organization whose sole purpose is to encourage the mistreatment of bears, or an organization whole sole purpose is to encourage the mistreatment of beings capable of suffering, they would be acting immorally. In other words, there is no *other object* with which we can place "*o*", that will allow us to draw a meaningful distinction between these generalizations.

Ultimately, these problems derive directly from the generalist's demand that explanatory generalizations be exceptionless. In attempting to allow certain explanatory generalizations, like **MENDEL** and **DECEPTION**, to qualify as laws/principles, generalism allows other nonexplanatory generalizations, like **ACCENT** and **DONATION**, in through the back door. This situation arises precisely because *ceteris paribus* laws, like **ACCENT$_{\text{K}}$** and **DONATION$_{\text{WBBA}}$**, are required to be genuinely exceptionless. Similarly, appealing to OOCFs cannot help the generalist here, precisely *because* such counterfactual robustness is a

precondition of exceptionlessness. What's more, it is the binary nature of the exceptionless/exceptionable distinction which makes generalism unable to characterize the sense in which certain generalizations appear to provide *better* explanations than others.

These considerations have driven Hitchcock and Woodward to abandon scientific generalism entirely, and with it, the idea that exceptionlessness and explanatoriness are connected *at all*. This is not to say, however, that Hitchcock and Woodward see no role for laws in scientific explanation; far from it. In place of generalism, they motivate an altogether different picture of law-hood, one which is based upon the idea that explanatoriness is a matter of *invariance*, rather than exceptionlessness.

In what remains of this paper, I shall argue that we ought to similarly abandon *moral* generalism, and with it, the idea that moral principles are exceptionless generalizations. I do not, however, propose that we concede the particularists' claim that such principles can play no role in normative theorizing. Rather, I shall argue for an invariantist analysis of moral principles, analogous to that advocated by Hitchcock and Woodward with respect to laws of nature. On this account, a generalization explains the rightness or wrongness of an action to the extent that it is invariant, and moral principles are nothing more that the *most* invariant of these explanatory generalizations.

V. Moral Invariantism

According to Hitchcock and Woodward, what makes a generalization explanatory is not whether it's exceptionless, but rather, the *extent* to which it is invariant under testing interventions. To be invariant in this respect, is to 'describe a relationship which holds for certain *hypothetical* values of *X* and *Y possessed by the very object o*… where *X* is changed by an intervention' (Hitchcock & Woodward, 2003a:20). As such, while the invariantist

agrees with the generalist that explanatory generalizations must be counterfactually robust, they disagree as to the *form* that such counterfactuals must take.

An intervention represents 'a hypothetical or counterfactual experiment that shows us that and how manipulation of the factors mentioned in the explanation… would be a way of manipulating or altering the phenomenon to be explained' (Woodward, 2003:11). What this means, is that to qualify as invariant, and hence explanatory, with respect to some object *o*, 'a generalization must support "same object" counterfactuals [SOCFs] that describe how the *very object o* would behave under an intervention' (2003:281).

As we have seen, by insisting that laws must be entirely exceptionless, the generalist is committed to the use of OOCFs in order to discern explanatory generalizations from merely accidental generalizations. Unlike OOCFs, however, SOCFs make explicit how the explanandum variable *depends* on the explanans: 'it is only if a generalization is invariant under testing interventions that it conveys information about how one variable depends upon another' (Hitchcock & Woodward, 2003a:19). While OOCFs often 'lack a clear interpretation… and seem incoherent or lack any scientific basis', Woodward argues that SOCFs are 'clear enough in meaning and we often have or can obtain scientific evidence that is relevant to their truth' (2003:282).

Laws of nature, on this invariantist analysis, are merely those generalizations which are *most* invariant. As a result, law-hood permits of both a *threshold* and (above this) a *continuum*. Some generalizations will not be invariant at all (or invariant only under a *very* narrow range of interventions) and will fall below the threshold for explanatoriness.[162] However, other generalizations will be invariant under a larger or more significant range of

[162] The generalization **ACCENT** would fall into this category. Even if we can think up some hypothetical scenario in which it is true (perhaps in some post-apocalypse future), its truth would nonetheless 'depend on a great many very specific contingencies, and if these were to change [the generalization] would be disrupted' (Woodward, 2003:310).

changes, involving both testing interventions and initial conditions, 'so that there will be a continuum of extent of invariance above the threshold' (Woodward, 2013:64).[163]

By adopting this invariantist analysis of explanation, and applying it to normative explanation, I believe that we can provide an account of moral principles which avoids the pitfalls of generalism. For starters, the moral invariantist can avoid accepting generalizations like **DONATION** as moral principles, by arguing that they fall *below* the required threshold for explanatoriness. By utilizing SOCFs, the invariantist is able to show that the wrongness of donating money to the WBBA isn't dependent (or is, at least, only peripherally dependent) upon the organization's being not-for-profit.

Suppose that Boris is intending to donate $100k to the WBBA, but we intervene and convince him to donate his money to the Philosophy Foundation instead. Such an intervention is clearly one under which **DONATION** *does not* remain invariant. Unlike the WBBA, there is nothing intuitively wrong with donating money to the Philosophy Foundation. Indeed, **DONATION** will not be invariant under any intervention resulting in an agent donating money to a not-for-profit organization where, in so doing, they act morally. Given the extensive range of such scenarios, this tells us that the wrongness of donating money to the WBBA is dependent upon something *other* than its being not-for-profit.

We have already considered two additional generalizations which could be used to explain the wrongness of Boris donating money to the WBBA: **DONATION**$_{\textbf{BEARS}}$ and **DONATION**$_{\textbf{SUFFERING}}$. I have argued, however, that the generalist has no way of comparing these candidate explanations, that is, no way of assessing what the wrongness of Boris' action actually *depends* upon. In contrast, the moral invariantist can provide a satisfying account of

[163] Newtonian mechanics, for example, is invariant under a considerable range of changes in terms of both interventions and initial conditions. Even though variable values can be assigned to interventions (concerning velocities close to that of light), under which such generalizations will not be invariant, Newtonian mechanics clearly meets the threshold for explanatoriness nonetheless.

why the latter provides a better, or "deeper", explanation than the former. The first step is to acknowledge that we must 'abandon the law/accident dichotomy, and replace it with an alternative framework' (Hitchcock & Woodward, 2003b:183-184).

On an invariantist analysis, among those generalizations that are invariant, some will be more invariant than others, they will hold under a *wider range* of changes and will correspondingly provide deeper explanations of target phenomenon. As such, for **DONATION$_{SUFFERING}$** to provide a better explanation of the wrongness of Boris' donating to the WBBA than **DONATION$_{BEARS}$**, it ought to be the case that the former is *more* invariant than the latter. What we should expect, then, is for **DONATION$_{BEARS}$** to be invariant under some range of interventions *R*, while **DONATION$_{SUFFERING}$** is invariant under the wider range *R**, which strictly contains *R,* as well as additional interventions under which **DONATION$_{BEARS}$** is not invariant. This is precisely what we find.

Suppose, for example, that we are able to convince Boris not to donate money to the WBBA and he instead donates his $100k to the "Association of Bear Baiting Promoters" (ABBP) another organization whose sole purpose is to encourage the mistreatment of bears. In this scenario, both **DONATION$_{BEARS}$** and **DONATION$_{SUFFERING}$** remain invariant; they can both be used to explain the wrongness of Boris' action. Now consider the range of interventions under which only **DONATION$_{SUFFERING}$** remains invariant. Imagine that, in response to our intervention, Boris instead donates $100k to the "American Cock Fighting Federation" (ACFF). In this instance, since male chickens are capable of suffering, **DONATION$_{SUFFERING}$** explains the obvious wrongness of Boris' donating to the ACFF, and thus remains invariant. **DONATION$_{BEARS}$**, however, does not.

Consequently, **DONATION$_{BEARS}$** is invariant under the range of interventions *R*, which includes scenarios in which Boris donates money to any organization whose sole

purpose is to promote the mistreatment of bears. In contrast, **DONATION**$_{\textbf{SUFFERING}}$ is invariant under the wider range *R**, which strictly contains *R*, in addition to interventions resulting in scenarios in which Boris donates money to an organization whose sole purpose is to promote the mistreatment of *any* being capable of suffering, including bears. As a result, on an invariantist picture, we can see that the wrongness of donating money to the WBBA dependents more upon the fact that they promote the mistreatment of beings capable of suffering, than on the fact that they encourage the mistreatment of bears specifically.

Clearly, an invariantist conception of moral principles is not plagued by the issues facing moral generalism. On this analysis, moral principles are understood as being those generalizations which are *most* invariant, rather than those generalizations which are exceptionless. In this sense, the invariantist can account for the explanatory character of generalizations like **DECEPTION** but need not accept generalizations like **DONATION** as moral principles. While the former is invariant under a wide enough range of interventions to meet the threshold for explanatoriness, the latter is not. What's more, this conception of explanation, in terms of range of invariance, provides a natural *measure* of explanatoriness: the more invariant a generalization, the "deeper" the explanations it provides.

Admittedly, however, this picture is, in some ways, less clear-cut than that provided by the generalist. For example, where we locate the threshold at which a generalization qualifies as a law/principle is a difficult question and will be dictated, to at least some extent, by pragmatic considerations. According to Woodward (2013), while there is nothing more to lawfulness than *de facto* invariance under some appropriately large range of testing interventions, there are no hard-and-fast rules by which we can identify the point where an explanatory generalization becomes *invariant enough* to qualify as a genuine law of nature.

However, I take this apparent lack of clarity to be a feature, rather than a bug, of my account. After all, it is precisely in drawing a hard border between nonexplanatory generalisations, and explanatory laws, which causes much of the difficulty facing the generalist analysis. Having said this, I do think that we can identify some general rules of thumb which can at least help us to *eliminate* candidate generalizations as laws or principles, if not confirm them outright. In the final section of this paper, I attempt to assuage concerns on this point by providing some useful heuristics which can be applied when attempting to assess the explanatory character of a generalization. I also briefly discuss the *taxonomy* of moral invariantism, and why I take it to be importantly distinct from both moral generalism and moral particularism.

VI. Some Remarks on Heuristics and Taxonomy

Given that, on an invariantist pictures, laws and principles are our *most* explanatory generalizations, eliminating those which are not invariant enough to qualify as explanatory at all seems like the obvious first step is assessing a candidate. To do this, I believe that we ought to utilize a distinction between two different *ways* in which a generalization can fail to be invariant. When Boris donates money to the Philosophy Foundation, for example, **DONATIO$_{\text{BEARS}}$** and **DONATION$_{\text{SUFFERING}}$** fail to be invariant because they simply don't *apply*. While both generalizations remain true, this truth is trivial; a result of the conditions specified in their antecedents not being met. However, under the very same testing intervention **DONATION** will be *violated* because any act which involves donating money to a not-for-profit organization, where that act isn't wrong, makes the antecedent of **DONATION** true and its conclusion false.

This provides us with our first rule of thumb:

> **RULE1**: a generalization will fall below the threshold to qualify as explanatory, if it can be easily violated.

Consider the generalization **ACCENT** once again. As I noted in section IV., I consider myself to have normal neurophysiological equipment, yet I do not speak English with a southern U.S. accent. As a result, I appear to *violate* **ACCENT**. Indeed, this generalization will be violated under a very wide range of testing interventions; *any* intervention which results in a scenario in which I speak English with an accent other than southern U.S. I believe that this is, in large part, what motivates the intuition that generalizations like **DONATION** and **ACCENT** fail to reach the threshold for explanatoriness. That such generalizations can be so easily violated suggests that, even in those instances where they hold, they do so only by accident, rather than by virtue of having latched onto a factor upon which the explanandum variable *depends*.

Another reason that **RULE1** represents a natural place to start assessing a generalization for invariance, is that it allows us to eliminate candidate laws/principles *in isolation*. When considering the range of interventions under which **DONATION** is violated, we are not required to consider how this generalization stacks up against other candidate explanations of the same phenomena. In contrast, **DONATION$_{\text{BEARS}}$** appears to be violated only under very select circumstances, which are hard to imagine; those where donating money to an organization whose sole purpose is to promote the mistreatment of bears *isn't* wrong. As a result, it would be very difficult to establish whether **DONATION$_{\text{BEARS}}$** qualifies as a principle without considering *other* generalizations which purport to explain the same phenomena.

This brings us to our second rule of thumb:

> **RULE2**: a generalization will fall below the threshold to qualify as a law/principle, if there is a generalization which provides a deeper explanation of the same phenomenon.

While we cannot specify precisely how invariant an explanation must be in order to qualify as a principle, the fact that **DONATION**$_{\textbf{SUFFERING}}$ provides a deeper explanation than **DONATION**$_{\textbf{BEARS}}$ tells us that the latter generalization *cannot* be a principle. Why? Because, where principles are understood as the set of generalizations which are most invariant with respect to a given phenomenon, the fact that **DONATION**$_{\textbf{SUFFERING}}$ is more invariant than **DONATION**$_{\textbf{BEARS}}$ means that the latter cannot be among this set. In this sense, **DONATION**$_{\textbf{SUFFERING}}$ remains, at least a *candidate* for, a moral principle.

In light of my claim that we ought to define explanatoriness in terms of invariance, rather than exceptionlessness, one might think that we can simply do away with the notion of a "moral principle" altogether. For the generalist, moral principles are *required* in order to provide normative explanations; to explain why an act is right or wrong. However, I have argued that a generalization needn't be a moral principle in order to be explanatory; "moral principle" is merely the mark of our *most invariant* normatively explanatory generalizations. As a result, it could be argued that my position leaves little work for moral principles, and that invariantism is therefore best understood as a form of moral *particularism.*

Interestingly, Woodward considers whether his own invariantist stance on laws of nature ought to similarly commit the scientific invariantist to rejecting the utility of laws of nature *tout court*. Such a position would, for all intents and purposes, be an analogue of moral particularism; *scientific* particularism.[164] However, Woodward (2003) is reluctant to abandon laws completely and argues that, while their role is both less crucial and less widespread than generalist approaches suppose, they nonetheless play an important part in (at least some areas of) scientific theorizing. In my view, the invariantist regarding moral principles ought to adopt a similar stance with respect to moral principles.

[164] Something like this position has been adopted by the likes of Cartwright (1983); Giere (1988); and van Fraassen (1989), all of whom are sceptical of the role standardly assigned to laws of nature in science.

On the one hand, the moral invariantist rejects the generalist's claim that moral principles play a *fundamental* role in normative explanation; that it is *all and only* those generalizations that qualify as principles which are explanatory. On the other, the moral invariantist maintains that there is an important distinction between explanatory generalizations and moral principles. While this distinction is, as Woodward puts it, 'fuzzy and contentious', a distinction in *degree* rather than *kind*, it is a distinction nonetheless (2003:286).

Consider for example the following plausible candidate for a moral principle:

UTILITY: All acts which fail to maximize utility are wrong.

It seems uncontentious to suggest that **UTILITY** is simply of a different *magnitude* in terms of explanatory potential than the generalizations discussed so far. When Boris donates money to the WBBA, for example, it is arguable that the wrongness of this action boils down to a failure to maximize utility. Indeed, **UTILITY** appears to provide a significantly *deeper* explanation of the wrongness of Boris action than **DONATION**$_{\textbf{SUFFERING}}$.

To see this, observe that the latter will not be invariant in scenarios in which Boris acts wrongly in donating money to any organization whose sole purpose isn't to encourage the mistreatment of beings capable of suffering. Perhaps Boris donates money to the Global Warming Policy Foundation, for example, a lobbying group known for their staunch anti-climate change agenda. A strong case could certainly be made that **UTILITY** remains invariant in this case; it explains the wrongness of Boris' action here, where **DONATION**$_{\textbf{SUFFERING}}$ does not.

To accept that invariantism is merely a form of particularism would be to concede that there is no meaningful distinction between generalizations like **DONATION**$_{\textbf{SUFFERING}}$, and generalizations like **UTILITY**; the latter of which appear to have a depth of explanatory

power that genuinely warrants their being singled out as of a higher standard, i.e., *as* principles. Indeed, somewhat ironically, generalists have accused particularists of "flattening" the moral landscape, of failing to capture the way in which certain considerations have greater moral significance than others.[165] In this respect, moral invariantism clearly has the upper hand.

Of course, I have shown that moral generalism suffers from a very similar issue. However, I have also argued that this issue stems from confusion surrounding whether we should take exceptionlessness as indicative of explanatoriness. It is, therefore, unsurprising that both generalism and particularism can be accused of "flattening" the moral landscape, since they both agree that exceptionlessness is the mark explanatoriness. Where they disagree is with respect to further question "is such exceptionlessness *attainable*?" The generalist says "yes" (with the help of *ceteris paribus* conditions at least), while the particularist says "no".

Nonetheless, the invariantist *is* sympathetic to the traditional focus of moral generalism, 'the basic presupposition that morality can and should (indeed, on some accounts, must) be captured by a set of principles' (Mckeever & Ridge, 2006:4). By jettisoning the connection between exceptionlessness and explanatoriness, however, the invariantist can paint a much clearer picture of the merits of this goal. For the invariantist, traditional candidates for principle-hood, like the "principle of utility", or "the categorical imperative", warrant this status precisely *because* of their wide range of invariance; their explanatory power. The role which such "principles" have played in the history of normative inquiry remains a mystery unless we can provide an account of how they differ from other, exceptionable, but nonetheless explanatory, generalizations. In my view, only moral invariantism can provide such an account.

---

[165] See e.g., Dancy (1993); Little (2000); Cullity (2002).

References

Achinstein, P. (1983). *The Nature of Explanation*. Oxford: Oxford University Press.

Audi, P. (2012). A clarification and defences of the notion of grounding. In Fabrice Correia & Benjamin Schneider (eds) *Metaphysical Grounding: Understanding the Structure of Reality* 101-121. Cambridge: Cambridge University Press. 101-121.

Barnes, E. (1992). Explanatory Unification and the Problem of Asymmetry. *Philosophy of Science* 59 (4):558-571.

Baron, S. (2022). Counterfactuals of Ontological Dependence. *Journal of the American Philosophical Association* 8 (2):278-299.

Baron, S. & Colyvan, M. (2021). Explanation Impossible. *Philosophical Studies* 178 (2):559-576.

Baron, S. & Norton, J. (2021). Metaphysical Explanation: The Kitcher Picture. *Erkenntnis* 86 (1):187-207.

Baron, S., Colyvan, M. & Ripley, D. (2017). How mathematics can make a difference. *Philosopher's Imprint* 17 (3):1-9.

Baron, S., Colyvan, M. & Ripley, D. (2020). A counterfactual approach to explanation in mathematics. *Philosophia Mathematica* 28 (1):1-34.

Batterman, R. (2000). Multiple realizability and universality. *British Journal for the Philosophy of Science* 51 (1):115-145.

Batterman, R. (2002). Asymptotics and the role of minimal models. *British Journal for the Philosophy of Science* 53 (1):21-38.

Baumberger, C., Beisbart, C. & Brun, G. (2017). What is Understanding? An Overview of Recent Debates in Epistemology and Philosophy of Science. In Grimm, S., Baumberger, C. & Ammon, S. (eds.), *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science* 1-34. London: Routledge.

Baumgartner, M. & Casini, L. (2017). An Abductive Theory of Constitution. *Philosophy of Science* 84 (2):214-233.

Baumgartner, M. & Gebharter, A. (2016). Constitutive Relevance, Mutual Manipulability, and Fat-Handedness. *British Journal for the Philosophy of Science* 67 (3):731-756.

Beatty, J. (1995). The Evolutionary Contingency Thesis. In Lennox & Wolters (eds.), *Concepts, Theories and Rationality in the Biological Sciences*, 45-81. Pittsburgh: University of Pittsburgh Press.

Bechtel, W. (2008). Mechanisms in cognitive psychology: What are the operations. *Philosophy of Science* 75 (5):983-994.

Bechtel, W. & Abrahamson, A. (2005). Explanation: a mechanist alternative. *Studies in History and Philosophy of Science of Biological and Biomedical Sciences* 36 (2):421-441.

Bengson, J., Cuneo, T., & Shafer-Landau, R. (2019). Method in the service of progress. *Analytic Philosophy* 60:179-205.

Benson, H. (2000). *Socrative Wisdom: The Model of Knowledge in Plato's Early Dialogues.* New York: Oxford.

Berker, S. (2018). The Unity of Grounding. *Mind* 127 (507):729-777.

Berto, F. & Jago, M. (2013). *Impossible Worlds*. Oxford: Oxford University Press.

Bird, A. (2007). What is scientific progress? *Noûs 41 (1):64-89.*

Bird, A. (2008). Scientific progress as accumulation of knowledge: a reply to Rowbottom. *Studies in the History and Philosophy of Science Part A* 39:279-281.

Bird, A. (2019). The aim of belief and the aim of science. *Theoria. An International Journal for Theory, History and Foundations of Science* 34 (2):171-193.

Bird, A. (2022). *Knowing Science.* Oxford: Oxford University Press.

Black, M. (1952). The Identity of Indiscernibles. *Mind* 61 (242):153-164.

Bokulich, A. (2011). How scientific models can explain. *Synthese* 180 (1):33-45.

Bokulich, A. (2018). Searching for Noncausal Explanations in a Sea of Causes. In Alexander Reutlinger & Juha Saatsi (eds) *Explanation Beyond Causation: Philosophical Perspectives on Non-Causal Explanations*. Oxford: Oxford University Press.

Bourget, D. (2017). The Role of Consciousness in Grasping and Understanding. *Philosophy and Phenomenological Research* 95 (2):285-318.

Bourget, D. & Chalmers, D. (2014). What do philosophers believe? *Philosophical Studies*, *170* (3):465-500.

Bratman, M. (1992). Practical Reasoning and Acceptance in Context. *Mind* 101 (401):1-16.

Briggs, R. (2012). Interventionist Counterfactuals. *Philosophical Studies* 160 (1): 139-166.

Brock, S. (2017). Is philosophy progressing fast enough? In Blackford, R. & Broderick, D. (eds.), *Philosophy's Future. The Problem of Philosophical Progress* 119-131. Hoboken: Wiley Blackwell.

Bromberger, S. (1965). An approach to explanation. In Butler, R. (ed.), *Studies in analytic philosophy, Volume 2* 72-105. Oxford: Blackwell.

Bryant, A. (2020). Keep the chickens cooped: the epistemic inadequacy of free range metaphysics. *Synthese* 197 (5):1867-1887.

Bunge, M. (1973). Philosophy of Physics. *Journal for General Philosophy of Science/ Zeitschrift für Allgemeine Wissenschaftstheorie* 4 (2):407-409.

Burnyeat, M. (1984). Aristotle on understanding knowledge. In Berti, E. (ed.), *Aristotle on Science: the Posterior Analytics* 97-139. Padua: Editrice Antenore.

Callender, C. (2011). Philosophy of Science and Metaphysics. In French, S. & Saatsi, J. (eds.), *The Continuum Companion to the Philosophy of Science* 33-54. London: Continuum.

Cappelen, H. (2017). Disagreement in philosophy: An optimistic perspective. In D'oro, G. & Overgaard, S. (eds.), *The Cambridge Companion to Philosophical Methodology* 56-75. Cambridge: Cambridge University Press.

Carnap, R. (1932). The Elimination of Metaphysics Through Logical Analysis. In Ayer, A. J. (ed.), *Logical Positivism* (1959) 60-81. New York: The Free Press.

Cartwright, N. (1979). Causal laws and effective strategies. *Noûs* 13 (4):419-437.

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

Cevolani, G. & Tambolo, L. (2013). Progress as Approximation to the Truth: A defence of the Verisimilitudinarian Approach. *Erkenntnis* 78 (4):921-935.

Chakravarty, A. (2017). *Scientific Ontology: Integrating Naturalized Metaphysics and Voluntarist Epistemology.* Oxford: Oxford University Press.

Chalmers, D. (2015). Why isn't there more progress in philosophy? *Philosophy*, *90* (1):3-31.

Cohen, L., J. (1992). *An Essay on Believe and Acceptance*. New York: Clarendon Press.

Collingwood, R., G. (1940). *An Essay on Metaphysics.* Oxford: Oxford University Press.

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

Craver, C. (2007b). Constitutive explanatory relevance. *Journal of Philosophical Research* 32:3-20.

Craver, C. (2014). The Ontic Account of Scientific Explanation. In Kaiser, M., Scholz, O., Plenge, D., & Hüttemann, A. (eds.), *Explanation in the Special Sciences: The Case of Biology and History*. Berlin: Springer Verlag.

Craver, C. & Darden, L. (2001). Discovering mechanisms in neurobiology: The case of spatial memory. In Machamer, P., Grush, R., & McLaughlin, P. (eds.) *Theory and Method in Neuroscience*. Pittsburgh: University of Pittsburgh Press.

Craver, C. & Darden, L. (2002). Strategies in the interfiled discovery of the mechanism of protein synthesis. *Studies in History and Philosophy of Science Part C: Studies in the History and Philosophy of Biological and Biomedical Sciences* 33 (1):1-28.

Craver, F. & Darden, L. (2013). *In Search of Mechanisms. Discoveries across the Life Sciences*. Chicago: University of Chicago Press

Crisp, R. (2000). Particularizing particularism. In Hooker & Little (eds.), *Moral Particularism*, 23-47. Oxford: Oxford University Press.

Cullity, G. (2002). Particularism and presumptive reasons. *Aristotelian Society Supplementary Volume* 76 (1):169-190.

Dancy, J. (1981). On moral properties. *Mind* 90 (359):367-385.

Dancy, J. (1993). *Moral Reasons*. Oxford: Blackwell.

Dancy, J. (2004). *Ethics Without Principles*. Oxford: Oxford University Press.

Dancy, J. (2009). Moral Particularism. In Zalta (ed.) *Standford Encyclopedia of Philosophy*. Center for the Study of Language and Information, Stanford University.

Darden, L. (2002). Rethinking Mechanistic Explanation. *Philosophy of Science* 69 (S3):342-352.

Darden, L., & Craver, C. (2002). Strategies in the interfiled discovery of the mechanism of protein synthesis. *Studies in History and Philosophy of Science Part C: Studies in the History and Philosophy of Biological and Biomedical Sciences* 33 (1):1-28.

Dasgupta, S. (2014). The Possibility of Physicalism. *Journal of Philosophy* 111 (9-10):557-592.

Dasgupta, S. (2017). Constitutive Explanation. *Philosophical Issues* 27 (1):74-97.

de Regt, H. W. (2015). Scientific understanding: truth or dare? *Synthese* 192 (12):3781-3797.

de Regt, H. W. (2017). *Understanding Scientific Understanding.* Oxford: Oxford University Press.

de Regt, H. W. (2020). Understanding, Values, and the Aims of Science. *Philosophy of Science* 87 (5):921-932.

de Regt, H. W. (2022). Can scientific understanding be reduced to knowledge? In Lawler. L., Khalifa, K. & Shech, E. (eds.), *Scientific Understanding and Representation: Modeling in the Physical Sciences* 17-32. New York: Routledge

Dellsén, F. (2016). Scientific progress: Knowledge versus understanding. *Studies in History and Philosophy of Science Part A* 56:72-83.

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

Dellsén, F. (2018). Scientific Progress: Four Accounts. *Philosophy Compass* 13 (11):e12525.

Dellsén, F. (2021). Understanding scientific progress: the noetic account. *Sytnhese* 199 (3-4):11249-11278.

Dellsén, F. (2022). Scientific Progress: By-Whom or For-Whom? *Studies in the History and philosophy of science Part A* 97 (c):2028.

Dellsén, F. (forthcoming). Scientific Progress Without Justification. In Khalifa, K., Lawler, I. & Shech, E. (eds.), *Scientific Understanding and Representation: Modeling in the Physical Sciences* 370-386. New York: Routledge.

Dellsén, F., Lawler, I. & Norton, J. (2022). Thinking about Progress: From Science to Philosophy. *Noûs* 56 (4):814-840.

Dellsén, F., Lawler, I. & Norton, J. (*forthcoming*). Would Disagreement Undermine Progress? *Journal of Philosophy*.

Depaul, M. R. (1987). Supervenience and moral dependence. *Philosophical Studies* 51 (3):425-439.

Dietrich, E. (2011). There is no progress in philosophy. *Essays in Philosophy* 12 (2):330-345.

Dowe, P. (1992). Wesley Salmon's Process Theory of Causation and the Conserved Quantity Theory. *Philosophy of Science* 59 (2):195-216).

Dowe, P. (2000). *Physical Causation.* Cambridge: Cambridge University Press.

Duhem, P. (1904-5). *The Aim and Structure of Physical Theory.* Wiener, P. (trans.). Ney York: Atheneum.

Einstein, A. (1905/1956). *Investigations on the Theory of Brownian Movement*. Cooper (trans.). New York: Dover.

Elgin, C. (1996). *Considered Judgement.* Princeton: Princeton University Press.

Elgin, C. (2004). True Enough. *Philosophical Issu*es 14 (1):113-131.

Elgin, C. (2007). Understanding and the facts. *Philosophical Studies* 132 (1):33-42.

Emmerson, N. (2021). A Defence of Manipulationist Noncausal Explanation: The Case for Intervention Liberalism. *Erkenntnis* https://doi.org/10.1007/s10670-021-00497-4.

Emmerson, N. (2022a). Understanding and scientific progress: Lessons from epistemology. *Synthese* 200 (1):1-18.

Emmerson, N. (2022b). Plumbing metaphysical explanatory depth. *Philosophical Studies* https://doi.org/10.1007/s11098-022-01886-3.

Esau, K. (1965). *Plant anatomy.* New York: John Wiley.

Ewing, A. (1929). *The Morality of Punishment: With Some Suggestions for a General Theory of Ethics.* Oxford: Routledge.

Flynn, J. (2010). Recent Work: Moral Particularism. *Analysis* 70 (1):140-148.

Frances, B. (2017). Extensive philosophical agreement and progress. *Metaphilosophy* 48:47-57.

French, S. (1989). Identity and Individuality in Classical and Quantum Physics. *Australasian Journal of Philosophy* 67 (4):432-446.

French, S. & McKenzie, K. (2015). Rethinking Outside the Toolbox: Reflecting Again on the Relationship between Philosophy of Science and Metaphysics. In Bigaj, T. & Wüthrich, C. (eds.), *Metaphysics in Contemporary Physics* 25-54. Leiden: Brill.

French, S. & Saatsi, J. (2018). Symmetries and explanatory dependencies in physics. In Reutlinger, A. & Saatsi, J. (eds.), *Explanation Beyond Causation: Philosophical Perspectives on Non-Causal Explanations* 185-205. Oxford: Oxford University Press.

Friedman, M. (1974). Explanation and Scientific Understanding. *Journal of Philosophy* 71 (1):5-19.

Frigg, R. (2010). Models and Fiction. *Synthese* 172 (2):251-168.

Gardiner, G. (2012). Understanding, Integration, and Epistemic Value. *Acta Analytica* 27 (2):163-181.

Gasking, D. (1955). Causation and recipes. *Mind* 64 (256):479-487.

Giere, R. (1988). *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.

Gillies, D. (1993). *Philosophy of Science in the Twentieth Century*. Oxford: Blackwell.

Gleeson, A. (2007). Moral particularism reconfigured. *Philosophical Investigations* 30 (4):363-380.

Glennan, S. (2010). Mechanisms, Causes, and the Layered Model of the World. *Philosophy and Phenomenological Research* 81 (2):362-381.

Glennan, S. (2017). The New Mechanical Philosophy. Oxford: Oxford University Press.

Goodman, J. (2004). An Extended Lewis/Stalnaker Semantics and the New Problem of Counterpossibles. *Philosophical Papers* 33 (1):35-66.

Grimm, S. (2006). Is understanding a species of knowledge? *British Journal for the Philosophy of Science* 57 (3):515-535.

Grimm, S. (2010). The Goal of Explanation. *Studies in History and Philosophy of Science Part A* 41 (4):377-344.

Grimm, S. (2012). The Value of Understanding. *Philosophy Compass* 7 (2):103-117.

Grimm S. (2014) Understanding as Knowledge of Causes. In Fairweather A. (eds) *Virtue Epistemology Naturalized: Bridges Between Virtue Epistemology and Philosophy of Science.* Synthese Library (Studies in Epistemology, Logic, Methodology, and Philosophy of Science) (366):329-345. Berlin: Springer.

Gubbins, S., Hunter, P., Nairsmith, J., Wood, J., & Woolhouse, M. (2020). Expert reaction to unpublished paper modelling what percentage of the UK population have been exposed to covid-19. *Science Media Centre*. https://www.sciencemediacentre.org/expert-reaction-to-unpublished-paper-modelling-what-percentage-of-the-uk-population-may-have-been-exposed-to-covid-19/

Gutting, G. (2016). Philosophical progress. In Cappelen, H. Gendler, T. & Hawthorne, J. (eds.), *The Oxford Handbook of Philosophical Methodology* 309-325. Oxford: Oxford University Press.

Haack, S. (1993). Evidence and Inquiry: Towards Reconstruction in Epistemology. Oxford: Blackwell.

Handfield, T. (2004). Counterlegals and Necessary Laws. *Philosophical Quarterly* 54 (216):402-419.

Hare, R. M. (1952). *The Language of Morals*. Oxford: University of Oxford Press.

Harinen, T. (2014). Mutual manipulability and causal inbetweenness. *Synthese, 195*(1), 35–54.

Hausman, D. M. (1982). Causal and Explanatory Asymmetry. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1982:43-54.

Hempel, C. (1959). "The Logic of Functional Analysis." In *Symposium on Sociological Theory*, ed. L. Gross, 271–87. New York: Harper & Row.

Hempel, C. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science.* New York: The Free Press.

Hempel, C. (1966). *Philosophy of Natural Science*. Prentice-Hall Foundations of Philosophy Series. New Jersey: Prentice-Hall.

Hempel, C. (1983/2001). Valuation and objectivity in Science. In Fetzer, J., H. (ed.), *The Philosophy of Carl G. Hempel*. New York: Oxford University Press. 372-396.

Hills, A. (2009). Moral Testimony and Moral Epistemology. *Ethics* 120 (1):94-127.

Hills, A. (2015). Understanding Why. *Noûs* 49 (2):661-688.

Hitchcock, C. (1995). Salmon on Explanatory Relevance. *Philosophy of Science* 62 (2):304-320.

Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *Journal of Philosophy* 98 (6):273-299.

Hitchcock, C. & Woodward, J. (2003a). Explanatory Generalizations, Part I: A Counterfactual Account. *Noûs* 37 (1):1-24.

Hitchcock, C. & Woodward, J. (2003b). Explanatory Generalizations, Part II: Plumbing Explanatory Depth. *Noûs* 37 (2):181–199.

Hooker, B. (2008). Moral Particularism in the real world. In Lance, Potrč & Strahovnik (eds.), *Challenging Moral Particularism,* 12-30. New York: Routledge.

Hooker, B. & Little, M. (2000). *Moral Particularism*. Oxford: Oxford University Press.

Horwich, P. (2012). *Wittgenstein's Metaphilosophy*. Oxford: Oxford University Press.

Humphreys, P. W. (1989). Scientific explanation - the causes, some of the causes and nothing but the causes. *Minnesota Studies in the Philosophy of Science* 13:283-306.

Huneman, P. (2010). Topological explanations and robustness in biological sciences. *Synthese* 177 (2):213-245.

Illari, P. & Williamson, J. (2012). What is a Mechanism? Thinking about mechanisms across sciences. *European Journal for Philosophy of Science* 2 (1):119-135.

Jaffe, J., Leopold, A. & Staples, R. (2002). Thigmo Responses in Plants and Fungi. *American Journal of Botany* 89: 375–82

Jago, M. (2015). Hyperintensional propositions. *Synthese* 192 (3):585-601.

Jansson, L. (2015). Explanatory Asymmetries: Laws of Nature Rehabilitated. *Journal of Philosophy* 112 (11):577-599.

Jansson, L. & Saatsi, J. (2019). Explanatory Abstractness. *British Journal for the Philosophy of Science* 70 (3):817-844.

Jones, W. (2017). Philosophy, progress, and identity. In Blackford, R. & Broderick, D. (Eds.), *Philosophy's Future. The Problem of Philosophical Progress* 227-239. Hoboken: Wiley Blackwell.

Kamber, R. (2017). Does philosophical progress matter? In Blackford, R. & Broderick, D. (eds.), *Philosophy's Future. The Problem of Philosophical Progress* 133-143. Hoboken: Wiley Blackwell.

Kelp, C. (2015). Understanding Phenomena. *Synthese* 192 (12):3799-3816.

Khalifa, K. (2013). The Role of Explanation in Understanding. *British Journal for the Philosophy of Science* 64 (1):161-187.

Khalifa, K. (2017). *Understanding, Explanation, and Scientific Knowledge*. Cambridge: Cambridge University Press.

Khalifa, K., Doble, G. & Millson, J. (2020). Counterfactuals and explanatory pluralism. *British Journal for the Philosophy of Science* 71 (4):1439-1460.

Khalifa, K., Millson, J. & Risjord, M. (2018). Inference, explanation, and asymmetry. *Synthese* (4):929-953.

Kim, J. (1973). Causes and Counterfactuals. *Journal of Philosophy* 70 (17):570-572.

Kim, J. (1974). Noncausal connections. *Noûs* 8 (1):41-52.

Kim, J. (1974/1993). Noncausal Connections. In Kim, J. (ed.), *Supervenience and Mind* 22-32. Cambridge: Cambridge University Press.

Kim, J. (1984/1993). Concepts of Supervenience. In Kim, J. (ed.), *Supervenience and Mind* 53-78. Cambridge: Cambridge University Press.

Kim, J. (1990). Supervenience as a Philosophical Concept. *Metaphilosophy* 21 (1-2):1-27.

Kim, J. (1994). Explanatory knowledge and metaphysical dependence. *Philosophical Issues* 5:51-69.

Kimpton-Nye, S. (2020). Necessary Laws and the Problem of Counterlegals. *Philosophy of Science* 87 (3):518-535.

Kirchin, S. (2007). Moral Particularism: an introduction. *Journal of Moral Philosophy* 4 (1):8-15.

Kitcher, P. (1981). Explanatory Unification. *Philosophy of Science* 48 (4):507-531.

Kitcher, P. (1989). Explanatory Unification and the Causal Structure of the World. In Philip Kitcher & Wesley Salmon (eds.), *Scientific Explanation* 410-505. Minneapolis: University of Minnesota Press.

Kitcher, P. (2002). Scientific Knowledge. In Moser (ed.), *The Oxford Handbook of Epistemology* 385-408. Oxford: Oxford University Press.

Kment, B. (2014). *Modality and Explanatory Reasoning.* New York: Oxford University Press.

Knudson, M., Desjarlais, D., & Dolan, D. (2008). Shock-Wave Exploration of the High-Pressure Phases of Carbon. *Science* 322 (5909):1822-1825.

Kompa, N. (2004). Moral particularism and epistemic contextualism: comments on Lance and Little. *Erkenntnis* 61 (2-3):457-467.

Kovacs, D. (2017). Grounding and the argument from explanatoriness. *Philosophical Studies* 174 (12):2927-2952.

Kovacs, D. (2019). The myth of the myth of supervenience. *Philosophical Studies* 176 (8):1967-1989.

Kovacs, D. (2020). Metaphysical Explanatory Unification. *Philosophical Studies* 177 (6):1659-1683.

Kovacs, D. M. (2020). Metaphysically Explanatory Unification. *Philosophical Studies* 177 (6):1659-1683.

Kovesi, J. (1967). *Moral Notions*. London: Routledge & Keegan Paul

Krickel, B. (2018). Saving the mutual manipulability account of constitutive relevance. *Studies in History and Philosophy of Science Part A* 68:58-67.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions.* Chicago: University of Chicago Press.

Kuhn, T. S. (1991). The road since structure. In Fine, Forbes & Wessels (eds.), *PSA 1990: Proceedings of the Biennial Meeting of the Philosophy of Science Association.* Chicago: University of Chicago Press.

Kuipers, T. (2009). Empirical progress and truth approximation by the 'hypothetico-probabilistic method'. *Erkenntnis* 70 (3):313-330.

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

Kvanvig, J., L. (2009). The value of understanding. In Pritchard, P., Haddock, A. & Millar, A. (eds.), *Epistemic Value* 95-112. Oxford: Oxford University Press.

Ladyman, J. (2014). Structural Realism. *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab: Stanford University.

Ladyman, J. & Ross, D. (2007). *Every Thing Must Go: Metaphysics Naturalized*. Oxford: Oxford University Press.

Lance, M. & Little, M. (2004). Defeasibility and The Normative Grasp of Context. *Erkenntnis* 61 (2-3):435-455.

Lance, M. & Little, M. (2008). From particularism to defeasibility in ethics. Lance, Potrč & Strahovnik (eds.), *Challenging Moral Particularism,* 53-74. New York: Routledge.

Lange, M. (2009a). Dimensional explanations. *Noûs* 43 (4):742-755.

Lange, M. (2009b). *Laws and Lawmakers: Science, Metaphysics, and Laws of Nature*. Oxford: Oxford University Press.

Lange, M. (2013a). What makes a scientific explanation distinctively mathematical? *British Journal for the Philosophy of Science* 64:485-511.

Lange, M. (2013b). Really statistical explanations and genetic drift. *Philosophy of Science* 80:169-188.

Lange, M. (2019). Asymmetry as a challenge to counterfactual accounts of non-causal explanation. *Synthese* 198:3893-3918.

Laudan, L. (1977). *Progress and its Problems: Toward a Theory of Scientific Growth*. Berkeley: University of California Press.

Laudan, L. (1981). A confutation of convergent realism. *Philosophy of Science* 48 (1):19-49.

Laudan, L. (1984). *Science and Values: The Aims of Science and Their Role in Scientific Debate.* Berkeley: University of California Press.

Lawler, I. (2022). Scientific progress and idealization. In Shan, Y. (ed.), *New Philosophical Perspectives on Scientific Progress* 332-354. New York: Routledge.

Lear, J. (1988). *Aristotle: The Desire to Understanding*. New York: Cambridge University Press.

Leuridan, B. (2012). Three problems for the mutual manipulability account of constitutive relevance in mechanisms. *British Journal for the Philosophy of Science, 63*(2), 399–427.

Lewis, D. (1973). *Counterfactuals*. Oxford: Blackwell.

Lewis, D. (1986). *Philosophical Papers Vol. II.* Oxford: Oxford University Press.

Lipton, P. (2001). What Good Is an Explanation? In Hon, G. & Rakover, S. (eds.), *Explanation* 43-59. Berlin: Springer Verlag.

Lipton, P. (2003). *Inference to the Best Explanation*. New York & London: Routledge.

Litpon, P. (2004). Inference to the best explanation. In Curd, M. & Psillos, S. (eds.), *The Routledge Companion to Philosophy of Science* 225-234. New York: Routledge.

Little, M. (2000). Moral Generalities Revisited. In Hooker & Little (eds.), *Moral Particularism,* 276-304. Oxford: Clarendon Press.

Lourenço, J., Paton, R., Ghafari, M., Kraemer, M., Thompson, C., Simmonds, P., Klenerman, P. & Gupta, S. (ms). "Fundamental principles of epidemic spread highlight the immediate need for large-scale serological surveys to assess the stage of the SARS-CoV-2 epidemic". https://doi.org/10.1101/2020.03.24.20042291.

Machamer, P., Darden, L. & Craver, C. (2000). Thinking About Mechanisms. *Philosophy of Science* 67 (1):1-25.

Maher, P. (1993). *Betting on Theories*. Cambridge: Cambridge University Press.

Maudlin, T. (2007). *The Metaphysics in Physics.* New York: Oxford University Press.

Maurin, (2018). Grounding and metaphysical explanation: it's complicated. *Philosophical Studies* 176 (6):1573-1594.

McKeever, S. & Ridge, M. (2006). *Principled Ethics: Generalism as a Regulative Ideal*. Oxford: Oxford University Press.

McKenzie, K. (2020). A Curse on Both Houses: Naturalistic Versus A Priori Metaphysics and the Problem of Progress. *Res Philosophica* 97 (1):1-29.

Meek, C. & Glymour, C., (1994). Conditioning and Interventing. *British Journal for the Philosophy of Science* 45 (4):1001-1021.

Menzies, P. & Price, H. (1993). Causation as a secondary quality. *British Journal for the Philosophy of Science* 44 (2):187-203.

Miller K. & Norton, J. (2022a). *Everyday Metaphysical Explanation*. Oxford: Oxford University Press.

Miller K. & Norton, J. (2022b). Non-cognitivism About Metaphysical Explanation. *Analytic Philosophy* 64 (2):1-20.

Miller, R. (1987). *Fact and Method: Explanation, Confirmation and Reality in the Natural and the Social Sciences.* Princeton: Princeton University Press.

Mitchell, S. D. (1997). Pragmatic laws. *Philosophy of Science* 64 (4):468-479

Mitchell, S. D. (2000). Dimensions of Scientific Law. *Philosophy of Science* 67 (2):242-265.

Mizrahi, M. (2012). Idealization and scientific understanding. *Philosophical Studies* 160 (2):237-252.

Mizrahi, M. (2013). What is Scientific Progress? Lessons from Scientific Practice. *Journal for General Philosophy of Science / Zeitschrift für Allgemeine Wissenschaftstheorie* (2):375-390.

Mizrahi, M. (2017). Scientific Progress: Why Getting Closer to Truth is Not Enough. *International Studies in the Philosophy of Science* 31 (4):415-419 .

Mizrahi, M. (2021). Conceptions of scientific progress in scientific practice: An empirical study. *Synthese* 199 (1-2):2375-2394.

Mizrahi, M. (2022). What is the Basic Unity of Scientific Progress? A Quantitative, Corpus-Based Study. *Journal for General Philosophy of Science / Zeitschrift für Allgemeine Wissenschaftstheorie 53 (4):441-458.*

Moravscik, J. (1979). Understanding and Knowledge in Plato's Philosophy. *Neue Hefte für Philosophy* 15:53-69.

Morrison, M. (2000). *Unifying Scientific Theories: Physical Concepts and Mathematical Structures.* Cambridge: Cambridge University Press.

Ney, A. (2016). Grounding in the Philosophy of Mind: A Defense. In Aizawa, K., & Gillett, C. (eds.), *Scientific Composition and Metaphysical Grounding*. London: Palgrave-Macmillan. 271-300.

Niiniluoto, I. (1980). Scientific progress. *Synthese* 45 (3):427-462.

Niiniluoto, I. (1984). *Is Science Progressive?* New York: Springer.

Niiniluoto, I. (1987). Is Science Progressive? *British Journal for the Philosophy of Science* 38 (2):272-276.

Niiniluoto, I. (1999). *Critical Scientific Realism.* Oxford: Oxford University Press.

Niiniluoto, I. (2014). Scientific progress as increasing verisimilitude. *Studies in the History and Philosophy of Science Part A* 46:73-77.

Niiniluoto, I. (2017). Optimistic realism about scientific progress. *Synthese* 194 (9):3291-3309.

Niiniluoto, I. (2019). Scientific progress. In Zalta, E. (ed.). *Stanford Encyclopaedia of Philosophy*. Metaphysics Research Lab: Stanford University.

Nozick, R. (2001). *Invariances: the structure of the objective world.* Cambridge: Cambridge University Press.

Nussbaum, M. (2000). In Hooker & Little (eds.), *Moral Particularism*, 227-255. Oxford: Oxford University Press.

Oddie, G. (1986). *Likeness to Truth.* Dordrecht and Boston: Reidel.

Park, S. (2017). Does Scientific Progress Consist in Increasing Knowledge or Understanding? *Zeitschrift für Allgemeine Wissenschaftstheorie* 48 (4):569-579

Park, S. (2020). Scientific Understanding, Fictional Understanding, and Scientific Progress. *Journal for General Philosophy of Science / Zeitschrift für Allgemeine Wissenschaftstheorie* 51 (1):173-184.

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

Pearl, J. (2009). Causal inference in statistics. An overview. *Statistical Surveys* 3:96-146.

Peirce, C. S. (1900-). *The Writings of Charles S. Peirce: A Chronological Edition.* E. Moore (ed.) Bloomington: Indiana University Press.

Peirce, C. S. (1931-58). *Collected Papers of Charles Sanders Peirce*. C. Hartshorne & P. Weiss (vols. i-vi), A. Burks (vols. vii & viii) (eds.) Cambridge MA: Belknap Press.

Perrin, J. (1909). *Brownian Motion and Molecular Reality*. Soddy (trans.) Oxford: Bow Press.

Pexton, M. (2014). How dimensional analysis can explain. *Synthese* 191 (10):2333-2351.

Pietroski, P. & Rey, G. (1995). When Other Things Aren't Equal: Saving *Ceteris Paribus* Laws from Vacuity. *British Journal for Philosophy of Science* 46:81-110.

Pincock, C. (2012). *Mathematics and Scientific Representation*. Oxford: Oxford University Press.

Poincaré, H. (1952). *Science and Hypothesis.* Reprint of the first English translation originally published (Paris, 1902) as *La Science et L'Hypothèse*. New York: Dover.

Popper, K. (1959). *The Logic of Scientific Discovery*. London: Hutchinson.

Popper, K. (1962). *Conjectures and Refutations*. New York & London: Routledge.

Popper, K. (1972). *Objective Knowledge: An Evolutionary Approach.* Oxford: Oxford University Press.

Post, H. R. (1971). Correspondence, Invariance and Heuristics: In Prise of Conservative Induction. *Studies in the History and Philosophy of Science Part A* 2 (3):213-255.

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

Priest, G. (2005). *Towards Non-being: The Logic and Metaphysics of Intentionality.* Oxford: Oxford University Press.

Pritchard, D. (2008). Knowing the answer, understanding and epistemic value. *Grazer Philosophische Studien* 77 (1):325-339.

Pritchard, D. (2010). The Nature and Value of Knowledge: Three Investigations. Oxford: Oxford University Press.

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

Putnam, H. (1975). What is Mathematical Truth? In Putnam (ed.), *Mathematics, Matter and Method*. Cambridge: Cambridge University Press.

Quine, W. V. O. (1951). Two Dogmas of Empiricism. *Philosophical Review* 60 (1):20-43.

Railton, P. (1981). Probability, explanation, and information. *Synthese* 48 (2):233-256.

Rapaport, W. (1982). Unsolvable Problems and Philosophical Progress. *American Philosophical Quarterly* 19 (4):289-298.

Raz, J. (2000). The truth in particularism. In Hooker & Little (eds.), *Moral Particularism*, 48-78. Oxford: Oxford University Press.

Raz, J. (2006). The trouble with particularism (Dancy's Version). *Mind* 115 (457):99-119.

Rescher, N. (2014). *Philosophical Progress: And Other Philosophical Studies*. Berlin: De Gruyter

Restall, G. (1997). Ways Things Can't Be. *Notre Dame Journal of Formal Logic* 38 (4):583-596.

Reutlinger, A. (2016). Is There a Monist Theory of Causal and Non-Causal Explanations? The Counterfactual Theory of Scientific Explanation. *Philosophy of Science* 83 (5):733-745.

Reutlinger, A. (2017). Does the Counterfactual Theory of Explanation Apply to Non-Causal Explanation in Metaphysics? *European Journal for Philosophy of Science* 1-18.

Reutlinger, A. (2018). Extending the Counterfactual Theory of Explanation. In Alexander Reutlinger & Juha Saatsi (eds) *Explanation Beyond Causation: Philosophical Perspectives on Non-Causal Explanations*. Oxford: Oxford University Press.

Reutlinger, A., Colyvan, M. & Krzyżanowska, K. (2020). The Prospects for a Monist Theory of Non-Causal Explanation in Science and Mathematics. *Erkenntnis* https://doi.org/10.1007/s10670-020-00273-w.

Rice, C. (2015). Moving Beyond Causes: Optimality Models and Scientific Explanation. *Noûs* 49 (3):589-615.

Rocca, M. (2005). Two Spheres, Twenty Spheres, and the Identity of Indiscernibles. *Pacific Philosophical Quarterly* 86 (4):480-492.

Romer, F. (2015). Why there isn't inter-level causation in mechanisms. *Synthese* 192 (11):3731-3755.

Rosen, G. (2010). Metaphysical Dependence: Grounding and Reduction. In Hale, B. & Hoffmann, A. (eds.), *Modality: Metaphysics, Logic, and Epistemology*. Oxford: Oxford University Press. 109-135.

Roski, S. (2020). Metaphysical Explanations and the Counterfactual Theory of Explanation. *Philosophical Studies*. https://doi.org/10.1007/s11098-020-01518-8.

Ross, W. D. (1930). *The Right and the Good: Some Problems in Ethics.* Oxford: Clarendon Press.

Rowbottom, D. P. (2008). N-rays and the semantic view of scientific progress. *Studies in the History and Philosophy of Science Part A* 39 (2):277-278.

Rowbottom, D. P. (2015). Scientific progress without increasing verisimilitude: In response to Niiniluoto. *Studies in History and Philosophy of Science Part A* 51:100-104.

Ruben, D. (1990). *Explaining Explanation*. Abingdon: Routledge.

Russell, B. (1912). *The Problems of Philosophy*. Portland: Barnes & Noble.

Saatsi, J. (2018). On Explanations from Geometry of Motion. *British Journal for the Philosophy of Science* 69 (1):253-273.

Saatsi, J. & Pexton, M. (2013). Reassessing Woodward's Account of Explanation: Regularities, Counterfactuals, and Noncausal Explanations. *Philosophy of Science* 80 (5):613-624.

Salay, N. (2008). Thinking without global generalizations: a cognatic defence of moral particularism. *Inquiry: An Interdisciplinary Journal of Philosophy* 51 (4):390-411.

Salmon, W. (1971). *Statistical Explanation and Statistical Relevance.* Pittsburgh: University of Pittsburgh Press

Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World.* Princeton: Princeton University Press.

Salmon, W. (1989). *Scientific Explanation*. Minneapolis: University of Minnesota Press.

Saunders, S. (2006). Are Quantum Particles Objects? *Analysis* 66 (1):52-63.

Schaffer, J. (2009). On what grounds what. In Manley, D., Chalmers, D., & Wasserman, R. (eds.), *Metametaphysics: New Essays on the Foundations of Ontology* 347-383. Oxford: Oxford University Press.

Schaffer, J. (2012). Grounding, transitivity, and contrastivity. In Correia, F., & Schneider, B. (eds.), *Metaphysical Grounding: Understanding the Structure of Reality* 122-138. Cambridge: Cambridge University Press.

Schaffer, J. (2016). Grounding in the image of causation. *Philosophical Studies* 173 (1):49-100.

Schaffer, J. (2017). Laws for Metaphysical Explanation. *Philosophical Issues* 27 (1):302-321.

Schindler, S. (2013). Mechanistic Explanation: asymmetry lost. In Karakostas & Dieks (eds.) *Recents Progress in Philosophy of Science: Perspectives and Foundational Problems.* Berlin: Springer.

Shan, Y. (2019). A New Functional Approach to Scientific Progress. *Philosophy of Science* 86 (4):739-758.

Shan, Y. (2022a). The Functional Approach: Scientific Progress as Increased Usefulness. In Shan, Y. (ed.), *New Philosophical Perspectives on Scientific Progress* 46-61. New York: Routledge.

Shan, Y. (2022b). Philosophy doesn't need a concept of progress. *Metaphilosophy* 53 (2-3):176-184.

Shand, J. (2017). Philosophy makes no progress, so what is the point of it? *Metaphilosophy* 48 (3):284-295.

Sher, G. (2021). Invariance as a basis for necessity and laws. *Philosophical Studies* 178 (12):3945-2974.

Shumener, E. (2020). Explaining Identity and Distinctness. *Philosophical Studies* 177 (7):2073-2096.

Singer, M. G. (1961). *Generalization in Ethics: An Essay in the Logical of Ethics, with the Rudements of a System of Moral Philosophy*. New York: Atheneum.

Skow, B. (2014). Are There Non-causal Explanations (of Particular Events)? *British Journal for the Philosophy of Science* 63 (3):445-467.

Skyrm, B. (1980). Causal Necessity: A Pragmatic Investigation of the Necessity of Laws. New Haven: Yale University Press.

Slezak, P. (2018). Is there progress in philosophy? The case for taking history seriously. *Philosophy* 93 (4):529-555.

Sliwa, P. (2015). Understanding and Knowing. *Proceedings of the Aristotelian Society* 115 (1):57-74.

Smart, J. J. C. (1963). *Philosophy and Scientific Realism.* London: Routledge.

Stalnaker, R. (1968). A Theory of Conditionals. In Nicholas Rescher (ed.), *Studies in Logical Theory (American Philosophical Quarterly Monographs 2)* 19-112. Oxford: Blackwell.

Stangl, R. (2006). Particularism and the point of moral principles. *Ethical Theory and Moral Practice* 9 (2):201-29.

Sterba, J. (2004). *The Triumph of Practice Over Theory in Ethics*. Oxford: Oxford University Press.

Sterpetti, F. (2018). The Noetic Account of Scientific Progress and the Factivity of Understanding. In Danks, D. & Ippoliti, E. (eds.), *Building Theories: Heuristics and Hypotheses in Science* 213-244. Cham: Springer Verlag.

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

Strahovnik, V. (2008). Introduction: challenging moral particularism. In Lance, M., Potrč, M. & Strahovnik, V. (eds.), *Challenging Moral Particularism,* 1-11. New York: Routledge.

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

Strevens, M. (2013). No understanding without explanation. *Studies in the History and Philosophy of Science Part A* 44 (3):510-515.

Swinburne, R. G. (1976). The Objectivity of Morality. *Philosophy* 51 (195):5-20.

Tan, P. (2019). Counterpossible Non-vacuity in Scientific Practice. *Journal of Philosophy* 116 (1):32-60.

Taylor, E. (2018). Against explanatory realism. *Philosophical Studies* 175 (1):197-219.

Thompson, N. (2016). Metaphysical Interdependence. In Jago, M. (ed.), *Reality Making*. Oxford: Oxford University Press.

Thompson, N. (2018). Irrealism about Grounding. *Royal Institute of Philosophy Supplement* 82:23-44.

Trogdon, K. (2013). An Introduction to Grounding. In Hoeltje, M., Schnieder, B., & Steinberg, A. (eds.), *Varieties of Dependence* 97-122. Munich: Philosophia Verlag.

Trogdon, K. (2018). Grounding Mechanical Explanation. *Philosophical Studies* 175 (6):1289-1309.

Trogdon, K. & Skiles (2021). Should explanation be a guide to ground? *Philosophical Studies* 178 (12):4083-4098.

Trout, J. D. (2002). Scientific explanation and the sense of understanding. *Philosophy of Science* 69 (2):212-233.

van Fraassen, B. (1980). *The Scientific Image.* Oxford: Oxford University Press.

van Fraassen, B. (1989). *Laws and Symmetry*. Oxford: Oxford University Press.

van Inwagen, P. (2004). The freedom to break laws. *Midwest Studies in Philosophy* 28:334-350.

Väyrynen, P. (2006). Moral generalism: enjoy in moderation. *Ethics* 116 (4):707-41.

Väyrynen, P. (2008). Usable moral principles. In Lance, M., Potrč, M. & Strahovnik, V. (eds.), *Challenging Moral Particularism,* 75-106. New York: Routledge.

Väyrynen, P. (2011). Moral Particularism. In Miller, C. (ed.), *The Continuum Companion to Ethics*, 247-260. London: Continuum.

von Wright, G., H. (1975). *Causality and Determinism*. New York: Columbia University Press.

Walton, K. (1990). *Mimesis as Make-Believe: On the Foundations of the Representational Art.* Cambridge MA: Harvard University Press.

Weatherall, J. O. (2011). On (Some) Explanations in Physics. *Philosophy of Science* 78 (3):421-477.

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

Wilkenfeld, D. (2013). Understanding as representation manipulability. *Synthese* 190 (6):997-1016.

Wilkenfeld, D. (2017). MUDdy understanding. *Synthese* 194 (4):1273-1293.

Wilkenfeld, D. (2019). Understanding as compression. *Philosophical Studies* 176 (10):2807-2831.

Williamson, T. (1997). Knowledge as evidence. *Mind* 106 (424):1-25.

Williamson, T. (2013). What Is Naturalism? In Haug, M. (ed.), *Philosophical Methodology: The Armchair or the Laboratory*? Abingdon and New York: Routledge.

Wilsch, T. (2015). The Nomological Account of Ground. *Philosophical Studies* 172 (12): 3293-3312.

Wilsch, T. (2016). The Deductive-Nomological Account of Metaphysical Explanation. *Australasian Journal of Philosophy* 94 (1):1-23.

Wilson, A. (2018a). Grounding Entails Counterpossible Non-Triviality. *Philosophy and Phenomenological Research* 92 (3):716-728.

Wilson, A. (2018b). Metaphysical causation. *Noûs* 50 (4):1-29.

Wilson, A. (2020). Classifying Dependencies. In David Glick, George Darby & Anna Marmodoro (eds) *The Foundation of Reality: Fundamentality, Space, and Time* 46-68. Oxford: Oxford University Press.

Wilson, A. (2021). Counterpossible Reasoning in Physics. *Philosophy of Science* 88 (5).

Wittgenstein, L. (1921). *Tractatus Logico-Philosophicus*. Anscombe, E., Hacker, P. & Schulte, J. (trans.). Hoboken: Wiley Blackwell.

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

Woodward, J. (2013). Laws, Causes, and Invariance. In Mumford & Tugby (eds.), *Metaphysics and Science*, 48-72. Oxford: Oxford University Press.

Woodward, J. (2015). Interventionism and Causal Exclusion. *Philosophy and Phenomenological Research* 91 (2):303-347.

Woodward, J. (2018a). Laws: An Invariance-Based Account. In Ott & Patton (eds.), *Laws of Nature*, 158-180. Oxford: Oxford University Press.

Woodward, J. (2018b). Some varieties of non-causal explanation. In Reutlinger, A. & Saatsi, J. (eds.), *Explanation Beyond Causation: Philosophical Perspectives on Non-Causal Explanation* 117-141. Oxford: Oxford University Press.

Woodward, J. (2021). *Causation with a Human Face: Normative Theory and Descriptive Psychology*. Oxford: Oxford University Press.

Worrall, J. (1989). Structural realism: The best of both worlds? *Dialectica* 43 (1-2):99-124.

Ylikoski, P. (2013). Causal and Constitutive Explanation Compared. *Erkenntnis* 78 (2):277-297.

Zagzebski, L. (2001). Recovering Understanding. In Steup (ed.), *Knowledge, Truth, and Duty: Essays on Epistemic Justification, Responsibility, and Virtue* 235-252. Oxford: Oxford University Press.