Mental Causation
The Happy Autonomy of Psychology and Physics

Jakeway, Ian

Awarding institution:
King's College London

Mental Causation: The Happy Autonomy of Psychology & Physics

Ian Jakeway

A Thesis submitted for the degree of Doctor of Philosophy

King's College London

April 2022

Abstract of the Thesis

In this thesis, I defend the genuine efficacy of mental causation, given both mental-physical supervenience and a causally complete physics. In so doing, I defend the mutual autonomy of psychology and physics.

This is principally achieved through a proposed dissolution of the Exclusion Problem. By focusing upon the role of supervenience in supposedly generating 'diagonal' causal relations between levels, I show that non-reductive physicalists need not accept the crux of the problem: the alleged entailment of systematic overdetermination. I argue that such overdetermination depends upon the obtaining of diagonal causation, and such causation is only entailed given further, auxiliary assumptions that might plausibly be resisted.

I also offer arguments to show that the above should *not* be taken to suggest a form of causal parallelism, on which there is no causal interaction between the mental and physical. There are plausible cases of mental-physical causation that are consistent with my arguments against the Exclusion Problem, and with a prohibition on systematic overdetermination, because they are not cases of supervenience-based causation. I therefore claim we have good grounds for thinking there can be mental to physical causation.

Beyond this, the thesis also develops a diagnosis of Exclusion worries through engaging with a recent Interventionist solution to the Exclusion Problem. I argue that the shortcomings of this solution are rooted in the same oversights from which more general Exclusion worries emerge.

Finally, I offer a speculative solution to two related concerns for Interventionists: intuitions of mental causal redundancy given the completeness of physics, and tensions internal to Interventionism with respect to causal relations within physics. I argue that both might be relieved by adopting causal pluralism, by which psychological causal relations are conceived in terms of correlations under intervention, and physical causation in terms of nomological regularity.

Acknowledgements

Only very few, if any, could ever describe writing their thesis as a sprint. In this regard, I unequivocally belong with the majority. Whilst working as a teacher alongside my research has been vital, it has not always been optimally conducive. All the more crucial, then, that I have been fortunate enough to have the enduring support of family, supervisors, friends and colleagues. Without it, my mental states could not have caused the physical effect you now see here.

I owe an enormous debt of gratitude to my parents, who have demonstrated infinite patience and perhaps irrational levels of confidence in me. They have never questioned the wisdom of embarking on a PhD, even when I might have, and have provided constant moral and practical support. This would be remarkable under any circumstances, but all the more so given they have not the slightest clue as to what I'm going on about. Thank you, both.

On the academic front, I must first thank Matthew Soteriou, my supervisor for the majority of the PhD. It is commonplace to say that a thesis would not possess its merits were it not for the input of a supervisor. This is undoubtedly true in the present case. But it is insufficient to express the impact that Matt has had on this work. I could perhaps write a chapter analysing this impact, but instead, I shall just say that Matt's command of the relevant areas, tactically deployed insight and general air of quiet encouragement have all been invaluable. Many, many thanks to Matt.

I also want to thank another Matt: Matthew Parrott. He supervised me during my first year at King's, and was hugely helpful in coaxing me through the 'upgrade' from MPhil to official PhD.

I am also grateful to another member of the current King's department, but who was in fact my Master's supervisor at Reading University, James Stazicker. James is a philosopher who, like Matt Soteriou and Matt Parrott, bucks the perceived notion that philosophers are combative, bullish and difficult. I benefitted enormously from both his guidance when writing my MA dissertation and his support when writing my PhD application.

Around the same time, I also had Gabriel Segal, emeritus professor at King's, as a supervisor for one of my MA topics: philosophy of mind. At that point, I had not formally studied the

area before, having chosen different papers during my undergraduate degree. Gabriel's relaxed, down-to-earth engagement with the issues and generous feedback were influential in my choosing to work on mind-related questions in the thesis.

Going right back to the beginning, I should thank Bill Mander, professor at Oxford. I might never have written a thesis without the good fortune of having Bill as my first and main tutor in Philosophy as an undergraduate. Tutorials with Bill were always challenging, and always in a particular, essentially egalitarian register: I had the sense that, whilst I was still largely ignorant, it was not entirely impossible to perhaps say something worthwhile. My impression of the subject, from Bill, was that it was endlessly intriguing, obstinately difficult and absolutely worth doing.

I also thank my friends, particularly Philip Clark, Jon Kershaw and Ben Child, for their patience and understanding. There have been uncountable occasions on which I have been unable to meet or virtually beyond contact. They have always been gracious, and have continued to signal approval of the project, despite its deleterious impact on the serious business of having fun. My thanks also go to Philip Clark for discussions on artificial intelligence and mind that served to both divert and inspire.

Due to working whilst researching, I have not always been able to participate in the life of the department at King's as much as I would have liked. However, I am grateful to participants in the Advanced Research Seminar, particularly between 2016 and 2018, for stimulating research talks and discussion, and for questions and feedback in response to my own presentations. I particularly benefitted from having spent time with Vanessa Brassey, Will Sharpe, and Athamos Stradis, and from discussions with David Papineau about physicalism and causation. More generally, I found the atmosphere and attitude amongst members of the department to be positive and welcoming.

Finally, I owe a quite literally unpayable debt of gratitude to Emma. The only way in which I might reciprocate her endless patience, academic advice and practical support would be through offering her the same whilst she completes a PhD thesis. But I can't; she's already written one. Despite her disappointment that I did not call this thesis, 'The Adventures of M and P', I hope she knows how grateful I am.

Table of contents

4. Pluralism & Exclusion

5. Towards Causal Pluralism: Interventionism & Incompleteness

Introduction

The title of this thesis can be read as a statement: psychology and physics are autonomous with respect to the causal relations pertinent to each. As such, the causal efficacy of mental properties is not degraded or precluded by a world in which the subjects of those properties – human persons – are also vectors of physical causation.

There are several senses in which the statement might seem trivial. From the point of view of practising scientists, be they psychologists or physicists, the mutual autonomy of these disciplines is taken largely for granted. Psychologists need no more consult, say, the equations of quantum dynamics, than physicists need concern themselves with whether a spatiotemporal region is occupied by minded creatures and what those creatures might be thinking or feeling.

Equally, the majority of non-scientists may well be disinclined to any suspicion that the theories and laws of physics might substantially interfere with or contribute to their own conception of human psychology. In responding to each other or musing on each other's behaviour, they will not typically worry about the subterranean goings-on of particles, forces and fields.

In what sense, then, is the statement substantive? It is substantive from the point of view of philosophers worried about the implications of physics for psychology when the ambitions and principles of the former are brought to bear on the question of causal efficacy. In other words, it is substantive as a response to the Exclusion Problem. In broad outline, the problem is this: mental-behavioural effects of ostensible mental causes supervene upon – and so are determined by – physical effects; but physical effects are guaranteed to have sufficient physical causes, so it looks like all that is needed to bring about the mental effect is a physical cause. This appears to be so for all mental-behavioural effects, and so it appears that no candidate mental causes have any causal work to do. Physical causation excludes mental causation.

Given the terms of the problem, the autonomy of psychology is threatened insofar as psychology might seem to merely limn the patterns of epiphenomena generated by the

underlying physical causation. Psychology is downgraded from dealing with genuine causal relations to tracing out shapes made at the mental level by the *real* causal relations obtaining at the physical.

This result is not of merely theoretical concern. It is not only scientific psychologists and philosophers of mind and psychology who need be moved by the threat to mental causation. The aforementioned non-scientists – as well as scientists and philosophers themselves in their everyday lives – are arguably wed, both theoretically and practically, to a conception of the human person that plausibly cannot withstand the loss of genuine mental efficacy. Notions of human motive, practical and theoretical rationality, agency and moral responsibility, to name just a few, all seem to presuppose the reality of mental causation. To lose it is to perhaps lose ourselves as intelligible creatures.

In this sense, then, the central argument of this thesis is far from trivial. Yet I draw attention to the prosaic perspectives above because my argument is intended as a recommendation that philosophers grant those perspectives philosophical significance. That is to say, the problem of mental causation is not a problem borne of how the world is, or how it is pre-theoretically taken to be, but by certain assumptions overlaid by philosophy. A guiding task of this thesis is therefore to show how those assumptions have generated and sustained the Exclusion Problem for mental causation. In so doing, we will see that, given the problematic nature of those assumptions, non-reductive physicalists can resist the problem from the outset. The mutual autonomy of psychology and physics need not be taken as jeopardized.

Metaphysical and Terminological Clarifications

Philosophical Orientation

As suggested by the above, the arguments of the thesis are directed at philosophers of a certain kind: non-reductive physicalists. Part of the reason for this is simply sociological: it is uncontroversial to think that many philosophers interested in the Exclusion Problem are sympathetic to this form of physicalism. On that basis, arguments amenable to non-reductive physicalists are arguments amenable to a significant portion of those party to the debate. But it is not the only reason. I also take it that non-reductive physicalism is reasonably well

motivated. I think it plausible that all concrete properties supervene upon physical properties, and that physics is causally complete. Equally, I do not rule out multiple realisability – and distinctness – of mental properties. To the extent that these commitments characterize non-reductive physicalism, I take non-reductive physicalism to be a position worth keeping on the table. Indeed, nothing that follows depends upon the meaning of 'non-reductive physicalism' beyond its entailing commitment to the core principles of the Exclusion Problem. So questions of definition are not here to the point.

In making further clarifications below, it will help to have in view the formulation of Exclusion that I adopt throughout. It comprises the following:

> **Supervenience:** Mental properties supervene upon physical properties.
>
> **Distinctness:** Mental properties are distinct from physical properties.
>
> **Causation:** Mental properties are, qua mental, causally efficacious.
>
> **Completeness:** All physical effects have sufficient physical causes.
>
> **Exclusion:** No effect has more than one sufficient cause unless it is genuinely overdetermined, i.e. unless overdetermined by independently sufficient causes.

Ontological Levels

When discussing the Causation and Completeness principles, I will sometimes refer to mental or physical properties as though they occupied different levels. It is commonplace for philosophers to talk of 'levels', roughly meaning domains corresponding to those of different sciences, or of different degrees of constitutive complexity (Searle, 1983; Post, 1991). Such talk, when interpreted as literal, is open to ontological objections (Heil, 2003). As a result, even loose metaphorical talk of levels is sometimes decried as potentially misleading. I do not enter these debates, and do not commit to any substantive notion of ontological levels. But I do adopt this way of talking as a convention, simply as a means of expressing the potential competition or other forms of interaction between classes of property. That is to say that if I use terms such as, 'the level of physics', I only do so because a picture of levels is of illustrative value. It is simplistic because, for instance, properties 'at the level of physics' are not of uniform complexity and do not all enter into the same laws or descriptions. But it does, I think, serve a purpose. Yet I stress, such talk is not intended to signify any substantive commitment to a genuinely *layered conception* of reality (whatever that might entail).

Physics

Given the importance of the Completeness principle, I will refer often to physics, and some of the arguments in later chapters will turn upon differences between the properties individuated in theories and laws of physics and those dealt with by special sciences. However, despite the importance of physical properties and laws in thinking about the Exclusion Problem, I do not commit to any substantive view about which entities properly belong to physics, and which to adjacent fields, such as atomic chemistry. When I refer to properties as physical, I mean *those properties covered by the Completeness principle.* This will suffice to pick out those properties that are alleged to compete with mental properties as causes.

Causation

One might expect a thesis about mental causation to be committed to a view about causation. This one is not. When we look at the Exclusion Problem in more detail (Chapter 2), there is a model of the problem that shares much of Lewis's early theory of causation (1973). There is also a model which assumes a broadly Interventionist conception of causation. But there is also another model which does not presuppose any particular conception, and more importantly, my arguments concerning those models are in principle applicable for other notions of causation.

However, any formulation of the issues relating to the Exclusion Problem must opt for some way or other of framing them. In my formulation of the principles above, e.g. Causation, I speak of 'properties'. So I here take the relata of causal relations to be properties. There are several reasons for this, none of them principled. One reason is just that many philosophers engaged in the debate would take the relata to be such. Another is that this fits well with the assumption, again common, that supervenience is a relation between properties; given that supervenience is closely tied to causation in the Exclusion Problem, it is useful to specify a kind of entity common to both. Additionally, opting for properties as relata is compatible with at least some views on which the relata are events, e.g. a Kimian notion of events as property instantiations by objects at times (Kim, 1976). Given that the causal relata are properties, where singular causation is at issue, the relata are property instantiations. At some points, I talk of properties, and at others, instantiations. Sometimes, when I use the term 'property', I

mean an instantiation of that property; the context of use should suffice to make clear the intended meaning.

I often use the term, 'diagonal' causation. This refers to causal relations between mental and physical levels, i.e. mental causes of physical effects or physical causes of mental effects. Similarly, 'horizontal' causation is causation between properties at the same level, mental or physical.

Overdetermination

The Exclusion principle above does important work. It serves to rule out causal overdetermination that is not the result of two or more causes that are independently sufficient for the effect. If the other principles imply diagonal causal relations (i.e. causal relations between mental and physical properties), and Completeness applies, then it looks like mental causation cases are cases of overdetermination. If that overdetermination is not independently sufficient, then the Exclusion principle rules it out. That is why the mental cause might be excluded.

But this begs the question, why is independently sufficient overdetermination ok, and non-independent overdetermination problematic? It is important for us because in Chapter 3, I propose a model of diagonal mental causation, from the mental to the physical, that I claim does not entail *problematic* overdetermination. So if we want to distinguish between benign and problematic forms of overdetermination, we had better have some notion of what makes the difference (see e.g. Funkhouser, 2002; Bernstein, 2016).

Accordingly, I will be using a stipulative notion of problematic overdetermination throughout. The idea is that *overdetermination is problematic only if systematic*. And *overdetermination is systematic if and only if entailed, in conjunction with Completeness, by a principle that is both general and modally strong.*

So overdetermination is here taken to be ruled out if it is entailed, in conjunction with Completeness, by a principle that is both general and modally strong. The candidate principle is Supervenience. It is general in the sense of covering all mental properties; it is modally strong because holding with metaphysical necessity.[1]

[1] Not all formulations of supervenience hold with metaphysical necessity. Some, for example, involve only physical necessity. But we will be assuming supervenience with this modal strength throughout.

The motivation for this notion of problematic overdetermination is the following. Standard examples of benign overdetermination are cases such as the firing squad, where multiple bullets each suffice to kill the target. By tweaking the case, we can test our intuitions about the limits of acceptability. It is sometimes suggested that it is *widespread* overdetermination that is the problem. If proponents of the Exclusion Problem are right, then mental causation – as something that is rife – entails an indefinitely large number of overdetermined effects; the sheer number of instances render overdetermination problematic. But if we were to assemble eager firing squads and an indefinitely large number of unfortunate victims, and schedule them to ritually overdetermine the latter's deaths, would this be metaphysically objectionable? I suggest not. If so, then perhaps being widespread is not the problem with problematic overdetermination.

It might be natural to then suppose that it is not the sheer number of cases, but the systematicity of the cases. I think this is on the right lines. But again, we should be careful about what this means. Perhaps we mean that the overdetermination occurs in accordance with a general principle or convention. But this seems vulnerable to similar considerations as above: imagine that, over time, our ritualized firing squad practice became part and parcel of social convention. This might include firing squads meeting at particular, regular times and in specific places. It might be a rite of passage for people to participate in a firing at least once before adulthood. But even if so, I do not find this metaphysically objectionable. I do not think that we would want to say that this simply *cannot happen*.

I suggest that it is systematicity of overdetermination that accounts for its being problematic, but systematicity has to amount to more than occurring in accordance with a general principle. It amounts to overdetermination being entailed by a general principle that is modally strong (in conjunction with Completeness). It is overdetermination that *must* occur, given Completeness, where 'must' indicates both the entailment and the modal strength. And it is overdetermination that must occur, if entailed, for *every* mental cause. If overdetermination results from the conjunction of Supervenience and Completeness, then it is of this kind.

Thesis Plan

The primary claim of the thesis is that our common, everyday conviction in the efficacious mental causation is not mis-placed; psychology need not yield to physics. In order to support that claim, I argue that non-reductive physicalists need not accept the Exclusion Problem, and my arguments in this regard show that the Exclusion principles (Supervenience, Distinctness, Causation and Completeness) do not entail diagonal causal relations between the mental and the physical. But this argument sets the stage for other important claims. I here outline the content, context and motivations for the key claims of the thesis.

In Chapter 1, I motivate the question of causal efficacy for mental properties by introducing the Exclusion Problem. I suggest that mental-physical supervenience and the causal completeness of physics are central to the problem, and yet not typically the focus of responses to it. I review three broad strategies for dealing with Exclusion: autonomy responses, compatibilist responses and those based upon an Interventionist conception of causation. Although intended to secure genuine mental efficacy in the face of the Exclusion Problem, these solutions are dogged by the threat of epiphenomenalism. I draw two lessons from the review: first, there is room for greater scrutiny of the role of supervenience in the problem, and second, a prospective dissolution of the problem holds considerable appeal.

In Chapter 2, I set about providing that dissolution. My method is to propose reconstructions of the Exclusion Problem and then show that diagonal causal relations are only hereby implied if further substantive and controversial assumptions are adopted. The purpose of the reconstructions is to explicate the role of supervenience in the standard formulation of the Exclusion Problem. To that end, the models build upon the standard formulation by adding auxiliary principles by which to potentially move from supervenience to diagonal causation.

But each of the models requires the adoption of assumptions that are problematic for the non-reductive physicalist. In particular, they require that mental properties are not multiply realized at close possible worlds and that supervenience is radically local, i.e. that specific mental properties depend upon and are necessitated by specific physical properties. I marshal arguments to show that neither of these assumptions are comfortably taken up by non-reductive physicalists.

Given that each of the models require at least one of the problematic assumptions, it follows that none of them is straightforwardly amenable to non-reductive physicalists. Since these models require those assumptions to imply diagonal causation, and hence systematic overdetermination, non-reductive physicalists can resist the threat of overdetermination. For the same reason, the conjunction of core Exclusion principles does not entail systematic overdetermination. Furthermore, the results of the chapter plausibly generalize, for the requisite assumptions are needed by any model on which supervenience is central.

In Chapter 3, I urge caution over the results of the preceding chapter. At this point two related claims have been made: the core Exclusion principles do not entail diagonal causation and the extended models need not – for non-reductive physicalists – entail diagonal causation. But we should not thereby conclude that diagonal causation does not, or cannot, occur. So we should not take the arguments of Chapter 2 to suggest a kind of parallelism.

The motive for this caution is an intuitive case of diagonal mental causation: Hadron Collider. This is a case that one would judge, pre-theoretically, as a case in which a mental cause brings about a physical effect. I intentionally flick the switch of the Hadron Collider and the subatomic activity commences. I argue that this judgement would be right, consistent with the arguments of Chapter 2, and coherent with the core Exclusion principles: Supervenience, Distinctness, Causation, Completeness and Exclusion. The non-reductive physicalist, sympathetic to those principles, is at liberty to endorse the picture suggested by the Hadron Collider case: direct mental causation.

The core proposal here is that the Hadron case is plausibly a case of *direct* mental causation. Such causation obtains not in virtue of any causal relation between the relevant subvening physical property and the relevant effect. So it is not supervenience-based causation. But the alleged diagonal causation of the Exclusion models in Chapter 2 *was* supervenience-based causation. It follows that one can endorse direct mental causation consistently with denying the entailment of diagonal relations by the Exclusion principles.

In Chapter 4, we assess a different response to the Exclusion Problem from John Campbell (2020). I argue that the response fails, and then diagnose the mistakes that might lead one to think it promising. That diagnosis then extends to a diagnosis of the Exclusion Problem more generally.

Campbell works within an Interventionist framework and dispels exclusion worries in the following way. First, all causation is explicated relative to an appropriate variable set. Second, when causal variables are related by supervenience, each variable must belong to mutually exclusive variable sets. So if a mental cause M supervenes upon physical cause P, M is a cause relative to one set, and P a cause relative to another. This principle is the Leanness principle. But then, third, causal overdetermination only occurs when the candidate causes are both causes relative to the *same* set. It seems to follow that, if M and P are related by supervenience, they cannot overdetermine any common effect.

I argue that this solution fails. In the first place, it depends upon application of the Leanness principle, and that principle is methodologically redundant. In short, we could only be in a position to separate variables, related by supervenience, into distinct variable sets if we had *already* distinguished causal relations between specific variables from supervenience relations between them. But to do that, if Campbell is right, we would have had to already apply the principle. So its grounds of application presuppose its redundancy.

In the second place, I argue that the Leanness principle is unwarranted because unmotivated. I elaborate the supposed motives for the principle by examining the potential for misleading correlations between variables that have not been sorted into different sets. I claim that in Exclusion contexts – in cases implied by the core Exclusion principles – there is no reason for thinking that such misleading correlations will arise. That is for reasons familiar from the arguments of Chapter 2: the prospects of those correlations rest upon denying close multiple realisability of mental properties and upon adopting a radically local formulation of supervenience. This suggests a diagnosis of the position on which the Leanness principle might be thought appealing: failure to take seriously the role of supervenience in Exclusion contexts. I then argue that the same diagnosis applies to the Exclusion Problem.

In Chapter 5, I address a potential worry that might linger, even in cases of direct mental causation, i.e. cases of non-supervenience-based causation such as Hadron. In cases like Hadron, we have a direct mental cause of a physical effect. The worry is not a worry about systematic overdetermination, but rather, about the dominance of the physical in the single case: if every physical effect is guaranteed a sufficient physical cause by Completeness, then how can the mental cause be genuinely efficacious? What is there for it to do? How is the mental cause not redundant?

I address this concern insofar as it might arise for the Interventionist. My solution to the worry is by way of a schematic, speculative proposal: the Interventionist can assuage their concern by endorsing causal pluralism. That is, they can advocate an Interventionist conception of causation for mental properties and a nomological regularity view for physical. In that way, they might avoid the intuition of competing causation, plausibly at the root of the redundancy concern.

The causal pluralism proposal is not only intended as a cure for that worry. It is also offered as a prospective means of alleviating the tension within an Interventionist conception of physical causation.

The argument for causal pluralism proceeds by way of the Incompleteness Argument, reconstructed from Campbell's argument in his (2020). The argument claims that, within an Interventionist framework, there can be no complete set of explicit causal relations. In essence, the thought is that any explicit causal relation must be between variables within a set. But any such relation requires another causal relation that is, so to speak, outside the set. This is because to render explicit a causal relation between variables within a set, one needs to intervene upon the cause from outside the set, and that intervention is itself causal. So there can be no set that includes all the explicit causal relations.

Such a conclusion suggests that no complete set of physical causal principles will be forthcoming under Interventionism (Campbell, 2020). But I argue that our non-reductive physicalist, sympathetic to the Exclusion principles, cannot comfortably tolerate this result. For they will endorse causal completeness of the physical, which I argue entails a commitment to exceptionless physical laws. And if the physical laws are exceptionless, then we should expect a complete set of explicit physical causal relations. There is therefore a potential tension within the Interventionist position.

There is another option for the Interventionist. They might try to evade the Incompleteness Problem by deploying a strategy adapted from Woodward (2003). On this strategy, explicit causal relations within the complete set do not require implicit causal relations outside of the set. To render the causal relations explicit, we do not need to intervene upon causal variables within the set; rather, we use the laws of physics to assign values to the causal variables and to calculate the correlative values in the effect variables.

I argue that this strategy would be another source of tension for the Interventionist. If they held onto their Interventionist conception of causation for physics, but deployed the strategy

above, their position would be unstable. If the laws of physics are exceptionless, then this suggests a nomological regularity conception of causation. But this is not an Interventionist conception. An Interventionist notion of causation does not require nomological regularity between correlated variables, only correlations under intervention sufficient to support interventionist counterfactuals. For that reason, Interventionist causation does not require exceptionless correlations, either. The Interventionist who attempts to evade the Incompleteness Problem by appealing to the laws of physics is in danger of trying to hold together incompatible conceptions of causation.

On my picture, then, it looks like the Interventionist has a dilemma regarding the Incompleteness Problem. If they accept that the set of explicit causal relations is incomplete, they appear at odds with the causal completeness principle. If they appeal to Woodward's strategy, they appear committed to incompatible conceptions of causation.

My speculative proposal is that the Interventionist avoid the dilemma by dropping Interventionism for causal relations in physics. The suggestion is that they retain Interventionism for mental causation, and accept nomological regularity for physical.

I finish the chapter by extending the proposal as a diagnosis and cure for lingering worries about causal completeness in cases of direct mental causation. On a pluralistic picture, the physical causation entailed by Completeness is a distinct kind of relation from the mental relation. If so, then perhaps we need not worry about the former precluding the latter. Cleaving to monistic causation is the root of our worry; letting it go might be the cure.

Chapter One - The Threat of Physics for Psychology

Introduction

The primary aim of this thesis is to argue that non-reductive physicalists need not fear for the mutual autonomy of psychology and physics. They need not worry that their commitments, particularly to mental-physical supervenience and causal completeness of the physical, might threaten mental causation. In this chapter, I set the scene for my arguments by introducing the philosophical framework of that worry: the Exclusion Problem. I then proceed to show the central preoccupations and difficulties of broad strategies for dealing with it. The main claims of the chapter are these:

(i) Mental causation matters and is prima facie autonomous of physics.

(ii) Supervenience and causal completeness are central to the Exclusion Problem.

(iii) Non-reductive physicalist solutions have typically taken for granted the alleged role of supervenience in the problem and been beset by worries of epiphenomenalism.

(iv) To better defend genuine causal efficacy of the mental, a dissolution of the problem holds considerable appeal.

By focusing on the role of supervenience, I will argue for that dissolution in forthcoming chapters.

In Section 1, I give a brief overview of the significance of mental causation for our folk psychology and scientific psychology. We will also see that both forms of explanatory practice operate autonomously of physics.

In Section 2, I consider how physics stands in relation to mental causation. Whilst physicists do not generally regard mental causation as significant in their practice, we will see that features of a worldview informed by physics – supervenience and causal completeness – give rise to the Exclusion Problem.

In Section 3, I introduce the Exclusion Problem in more detail.

In Section 4, we review some of the broad non-reductive strategies for solving the problem. None manages to unequivocally ward off the threat of epiphenomenalism, and their focus

has not typically been upon the role of supervenience. There is therefore room for greater critical attention upon supervenience.

Section 1: Mental Causation & Psychology

In this section, I begin to motivate the central concerns of this thesis by drawing attention to the multivarious ways in which mental causation is vital to our self-conception as human beings. This self-conception is in the first place manifest in our folk psychology and, beyond this, in our scientific psychology.[2] Our brief overview of each can only be, at best, partial. But I hope that it suffices to highlight what all reflective people already know: mental causation matters.

I break into a run on the way to the train station; I greet an acquaintance in the street; I check that my cloud storage has sufficient space for the rest of the thesis. In every case, there is mentality at work. I want to be on time, and believe that walking will not get me there quickly enough; I believe the person in front of me to be someone I once met and do not want them to think me unfriendly; I want the thesis to be safely stored and know my cloud capacity to be limited. For even these relatively simple behaviours, there is more that might be profitably said about the relation between my mental states and my actions. But when making sense of our own behaviour or that of others, we typically do so through attribution of common-sense mental states such as desires, beliefs, fears, hopes and more. All of this is commonplace; part and parcel of our folk psychology. Central, and seemingly indispensable, to our common, everyday explanatory practice is the notion not only that we are each minded creatures, but that our mindedness is *causally efficacious*. My desire to be on time does not merely happen to magically co-exist with my starting to run for the train, the two states related only by brute coincidence; rather, my desire is causally efficacious with respect to my running. It is the putative causal efficacy of the mental that grounds our explanations of behaviour in mental terms. On this picture we behave as we do, in no small part, because we think, feel, remember, imagine, and so on.

[2] It is also of significant importance for a number of philosophical debates in addition to those at the heart of this thesis. But we confine ourselves here to its importance for psychology, since it is psychology that is our concern throughout.

Such mentalistic explanation would be significant even if purely theoretical. But of course, it is not. We do not engage in folk psychology as dispassionate by-standers, but as creatures already embedded in a social world. Our explanatory practice is often subconscious, subtly informing our cognitive, emotional, and behavioural responses to others. We relate to others with the presupposition that they are reasonably similar to others; similar enough to warrant unconscious application of loose generalisations about how our mental life causally relates to our behaviour. Furthermore, we are not only in the business of explaining – facing, so to speak, backwards – but also of predicting, if only vaguely. Again, such predictions are not theoretical, but enmeshed in our interactions with others. We are, for instance, disappointed in our friend because we expected more; we expected more because we perhaps unconsciously predicted that they would, given certain desires and beliefs, have behaved differently. It is plausible to think that our commonsense notions of mental causation, expressed in our common concepts such as those of beliefs, desires or fears, profoundly impact what P. F. Strawson called our 'reactive attitudes': our attitudes in reaction to others in the midst of our interaction with them.[3]

Implicit in our reactions and responses to others is something fundamental that again speaks of mental causation. We do not regard ourselves or others as machines, our bodily behaviours either randomly generated or causally governed by physical laws. Rather, we regard ourselves as, in some sense, the genesis of our actions. That is, we consider ourselves to be willed beings, capable of deliberation and of acting accordingly. This conception of agency is apparently essential for making sense of the reactive attitudes outlined above (indeed, those attitudes presuppose that conception), but also arguably for making sense of ourselves as moral creatures more generally. Without agency, we appear to lack responsibility (Kim, 2005). Without this self-conception, our moral and legal practice would be much impoverished and perhaps lacking in justification. But agency itself seems to require the causal efficacy of the mental, for unless what I will is causally connected with what I do, then my agency is inert and perhaps does not qualify as agency at all (Kane, 1996). In this way, mental causation is plausibly constitutively related to our personhood.

[3] See Strawson, 1962.

All of the above is to say that mental causation is of utmost importance. It is integral to our commonsense notions of mentality and our conditioned stances toward each other as social beings. Even a rough, cursory review of its place in our self-conception and social behaviour suggests that mental causation is a condition of intelligibility for us. Without the supposition that our mental lives are causally efficacious – and efficacious not only with respect to mental states but also to physical, behavioural states – it is plausible that we would lose much, if not all, of what we take for granted in understanding ourselves and each other.

Mental causation is also of central importance in scientific psychology. It is difficult to know how one might begin to characterize the aims of scientific psychology without invoking the notion of mental causation. If psychologists are interested in, say, the relationship between trauma and depression, they are interested in the *causal* relationship between them (Campbell, 2020). If they are concerned with investigating the statistical relationship between, say, material deprivation and cognitive performance in young children, their concern is with the statistics insofar as they support causal inference. Broadly speaking, psychological explanation is causal explanation.

Furthermore, the methodology of psychology presupposes the causal efficacy of the mental. When instructing a control group in an experiment, the psychologist assumes that their verbal instructions will be acted upon. But this assumption requires that the members of the control group can act upon their beliefs; that is, it requires that their beliefs can cause their actions. The same is true for psychology conducted at the level of neuroscience: the minimally conscious patient in an MRI scanner is given instructions, and again, the presumption must be that hearing those instructions will be causally related to the patient's subsequent behaviour.

Given the above, mental causation is plausibly paramount in making sense of ourselves, both via our everyday commonsense conception and via our scientific picture. Without it, it is not clear that we could intelligibly live as we do. Indeed, our conception of our minds as causal may be so crucial that we would perhaps have to go on living *as if* it obtained even if intellectually debunked as myth.[4]

[4] This is similar to, but distinct from, Strawson's claim in his 'Freedom and Resentment' (1962) that a belief in determinism (as counter to our having free will) could have no material impact on our form of life. His point is that our reactive attitudes are entrenched in our forms of interaction, and could not be discarded on the basis

There is one more point of significance: all attributions and explanations of mental causation in folk psychology, and many in scientific psychology, proceed without reference to what goes on at the level of physics.[5] That is to say, mental causation is in the main considered as *autonomous* from causation at the level of physics. Causal explanations in psychology do not typically need contributions from the theories and laws of physics; nor are they taken to in any way threaten those theories and laws. Where mental causes are efficacious, they are efficacious qua mental causes, i.e. not in virtue of any relation between the mental and the physical. Psychology proceeds as if in a state of happy autonomy from physics.

Section 2: Mental Causation & Physics

Matters are similar for physics with respect to psychology. Physicists do not typically regard their projects as requiring consideration of mental causes, either as positive factors or as potential disruptors.[6] They approach their domain as causally isolated from that of psychology. This is not to say that physicists do not also, in non-professional contexts, assume that mental states are causally efficacious. Indeed, in some experimental contexts, too, mental causation is plausibly presupposed, as when physicists activate the Large Hadron Collider. But mental causes do not figure in the theories and laws of physics.

From what we have thus far said, one might wonder how the mutual autonomy of psychology and physics could be threatened. If we and scientific psychologists carry on with our attribution of causal efficacy to the mind unconcerned with physical goings-on, and physicists proceed with their investigations unperturbed by psychological causation, why is

---

of an intellectual stance. Similarly, perhaps our conception of mental causation is so integral to our rationalization of ourselves, to our relationships and our sense of meaning, that we would have to go on as if it were accurate even if intellectually persuaded otherwise.

[5] I say 'many' for scientific psychology because, depending upon what we include as physical and what we include as psychological, there may be some overlap. In areas of neuroscience, for example, investigation will be concerned with electrochemical reactions which might, depending on one's boundaries, qualify as physical. But such cases do not detract from our point, since much of what goes on in psychology broadly conceived does not involve reference to anything sensibly considered physical. In any case, a dispute here would be largely verbal. The point is that psychologists are not physicists, and do not aspire to be.

[6] The von Neumann-Wigner Interpretation of quantum mechanics is a *possible* exception. On this view, as on that of the standard Copenhagen Interpretation, measurement of a system causes collapse of the wavefunction where the superposition of states collapses to one of the possible states. But the von Neumann-Wigner reading of the problem posits the consciousness of the observer as the cause of the collapse. However, this is a controversial view and, in any case, it is not clear that it would amount to mental causation as understood by psychology (see Chalmers, 1996).

there anything more to be said? The problems start not strictly within either field, but rather, from philosophical intuitions provoked by principles of physics. That is, the problems arise in philosophy.

I take the genesis of the problem to be in two principles informally derived from reflection upon physics: causal completeness and supervenience. The former, though not strictly a principle within physics, is a working assumption of physics and desideratum of the commitment to exceptionless laws.[7] The latter is a philosophical thesis rendered plausible, at least in part, by causal completeness together with the assumed generality of physics.

Completeness: Every physical effect has a sufficient physical cause.

Supervenience: Mental properties supervene upon physical properties.

Completeness connects to physicists' ambition to formulate exceptionless laws; if physical laws were *ceteris paribus*, then they would in principle be open to instances of physical effects having causes from outside the physical domain.[8] Supervenience, read here as entailing dependence of mental properties upon physical properties, and necessitation of mental properties by physical properties, can be linked to Completeness and the generality of physics. Physics is, of all the sciences, maximally general in the sense of dealing with the basic constituents of the higher sciences.[9] Because these constituents, e.g. particles, forces, fields, are taken to be distributed across the space of entities investigated by the special sciences and taken to, in some sense, compose or constitute those entities[10], the physical constituents are naturally taken as the base upon which higher entities depend. Furthermore, if Completeness holds, then it seems natural to suppose that no higher-level entity will depend upon anything non-physical, or be itself independent of the physical, in virtue of its *causing* something physical; every base property in physics has a sufficient cause from within its domain.

[7] See e.g. Lewis (1966). Papineau (2000) also claims the principle finds inductive support from the progress of neurophysiology, made without the positing of any extra, non-physical forces.
[8] A physical law's being exceptionless is compatible with its holding relative to a background context.
[9] See e.g. Armstrong (1980, p. 20): "It seems increasingly likely that the body and the brain of man are constituted and work according to exactly the same principles as those physical principles that govern other, non-organic, matter".
[10] This is very much a simplified account of the picture, but here we only want a rough view of how Supervenience might be motivated by physics.

The broad picture that we now have is one on which the properties of physics are causally sufficient unto themselves, and those same properties are metaphysically sufficient for properties at higher levels. It can then start to look like the properties at higher levels, such as mental properties, are squeezed out of the running where causation is concerned. For if mental properties synchronically depend upon physical properties, and the latter are guaranteed as causally sufficient by Completeness, then the physical properties might look to be causally sufficient for the mental (including the behavioural). My arm is ultimately constituted of physical entities. If reaching for my glass therefore supervenes upon physical properties, and Completeness holds, then the properties that metaphysically determine my arms movement have sufficient physical causes. Given the sufficiency of the physical cause, what room is left for a mental cause of my reaching? In the above, we have a picture that might suggest a tension between the causal claims of psychology and those of physics, and hence the beginnings of a threat to their autonomy. We have a picture that conforms to the Exclusion Problem.

Section 3: Mental Causation & Exclusion

In the remainder of this thesis, I will argue that the above picture only serves to threaten the mutual autonomy of psychology and physics if conjoined with a host of other assumptions. To set the scene, most of the remainder of this chapter will seek to illustrate two main points:

1) Supervenience and Completeness are central to Exclusion worries.

2) Non-reductive physicalists have typically felt pressed by the Exclusion Problem to offer ostensible solutions that are in various ways costly.

To start with (1), let's consider the Exclusion Problem in more detail. The Exclusion Problem can be presented as a problem of inconsistency, alleging that the following conjunction of propositions entails a contradiction:

Supervenience: Mental properties supervene upon physical properties.

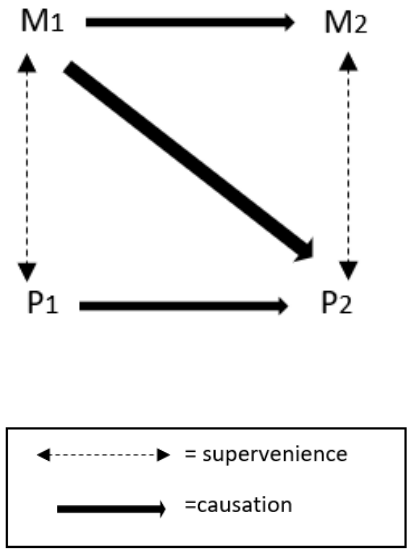
Distinctness: Mental properties are distinct from physical properties.

Causation: Mental properties are, qua mental, causally efficacious.

Completeness: All physical effects have sufficient physical causes.

Exclusion: No effect has more than one sufficient cause unless it is genuinely overdetermined, i.e. unless overdetermined by independently sufficient causes.

(Fig.1)



Here is one way that the above conjunction might play out (Figure 1).[11] In accordance with Causation, let us suppose that M1 causes some mental effect, M2. Given Supervenience, M2 will supervene upon some physical property, P2. Given Distinctness, M2 and P2 are type-distinct. Now, we take Supervenience to entail that P2 necessitates M2. In which case, it looks like M1 causes M2 via bringing about P2. Hence, it looks like M1 causes P2.

But this where the trouble starts, because given Completeness, P2 must have a *sufficient physical cause*, P1. It therefore looks like P2 has two sufficient causes: P1 and M1. We have already said that P1 and M1 are distinct properties, which suggests that P2 has two sufficient causes. The Exclusion principle tells us that no effect has two sufficient causes unless they are independently so. But because M1 supervenes upon P1, it looks like M1 does not cause P2 independently of P1. So if M1 and P1 each cause P2, we appear to have a contradiction of the Exclusion principle. It seems that the conjunction of Supervenience, Distinctness, Causation

[11] Broadly, this is the line of thought that Kim (2005) proposes in his presentation of the problem. This is only one route to potential Exclusion, though. In Chapter Two, we will see more.

and Completeness entails the negation of Exclusion; the conjunction of all the principles is internally inconsistent. The suggestion, then, is that those who wish to insist upon the efficacy of the mental are obliged to drop one of the other principles.[12] Conversely, those committed to those other principles must bite the bullet and accept epiphenomenalism of mental properties.

We can now see more clearly how tension between psychology and physics might arise. And we can begin to see the centrality of the Supervenience and Completeness principles for the Exclusion Problem. On the present account, Supervenience implies necessitation of m-properties by p-properties, and that is taken to imply that if a mental cause is to bring about a mental effect, it will do so via bringing about the subvening p-property. It is Supervenience that sets up the vertical determination from physical to mental, and so sets up the inferred causal relation from the mental cause to the physical effect.[13] But that causal relation is only problematic because apparently in competition with a horizontal relation between the physical effect and its sufficient physical cause; a relation guaranteed by Completeness. Where Supervenience appears to guarantee a diagonal causal relation, Completeness appears to guarantee a competing horizontal relation. The combination is at the heart of the apparent tension between psychology and physics. Whilst there are other routes to the alleged contradiction - and so potentially to the exclusion of mental causation - all such routes turn upon Supervenience and Completeness in broadly this way. And whilst there is more said in the forthcoming chapters about both principles, we can see in outline the challenge that they pose for mental causation.

Section 4: Supervenience Displaced – Some Existing Solutions to Exclusion

In this section, I want to draw attention to some main strategies for dealing with the Exclusion Problem. My purpose in so doing is to show that supervenience has often been overlooked by responses to the problem.[14] The nature and function of the relation has been

[12] Kim argues that physicalists in favour of mental causation must drop Distinctness, thereby committing to type-identity of mental and physical properties. His original Exclusion argument (1993a) was in the service of showing that non-reductive physicalism, insofar as it is opposed to mental epiphenomenalism, is incoherent.
[13] This is one form of the *diagonal* causal relations cited in the Introduction.
[14] Though Kim (1993a) has taken it to be central to the problem.

largely taken for granted, with critical focus directed elsewhere. But as we will see, none of the strategies is an unambiguous success; all are haunted by the worry of epiphenomenalism, as well as other difficulties. So my motivation in looking at these approaches to the Exclusion Problem is twofold: first, to note that there is room for greater focus upon supervenience, and second, to press home the value of a prospective *dissolution* of the problem. For if we can avoid the problem in the first place, we might avoid the difficulties that come with the solutions.

We now turn to looking at some of those responses from non-reductive physicalists. Non-reductive physicalists clearly have a vested interest in assuaging exclusion-related worries. As physicalists, they will typically endorse Completeness, Supervenience. As *non-reductive* physicalists, they will endorse Distinctness. For reasons outlined at the start of the chapter, they will want to avoid epiphenomenalism and so endorse Causation. But many have also felt compelled by the Exclusion principle, ruling out overdetermination by non-independent sufficient causes. I do not have space here for detailed examination of the field, for the literature on the Exclusion Problem is vast. Instead, I will consider some broad lines of response.

Autonomy Solutions

To start, some solutions have attempted to defuse the problem by effectively denying that M1 is in causal competition with its subvening physical realizer, P1 (see e.g. Thomasson, 1998; Marras, 1998; Gibbons, 2006). The thought here is that, just as properties of events can be causally relevant to explanations, they are also causally relevant to the relations that hold between them. That is to say, the causal relevance of properties is not merely epistemic, but metaphysical.[15] On that basis, we can distinguish between properties of the same event (cause) in terms of their relevance for an effect. Whereas M1 is causally relevant to the mental-behavioural property M2, P1 is causally relevant to the physical effect, P2. On one way of applying this to the above version of the problem, M1 does not cause M2 by causing P2. Rather, we should say that M1 and P1 are both distinct properties instantiated in the same

[15] This is in stark contrast with Davidson's (1970) view, on which it is appropriate to cite causal relevance of properties only when giving causal explanations. Properties are not the relata of causal relations; events are.

causal event, and whilst M1 is causally relevant for the M2 property of the effect event, P1 is causally relevant for the P2 property of the same effect. On this interpretation, there is no diagonal causal relation from M1 to P2; on the contrary, there are only two parallel horizontal relations between causal properties that are distinctly relevant for properties of the effect.

I take the proposal to hold some appeal. The thought that particular properties of an event are causally relevant to an effect is intuitively attractive. (Plausibly: the explosion in the basement caused the house to burn down in virtue of its being an explosion; not in virtue of its being the loudest occurrence in the house this morning.) And the solution it affords the proponent of mental causation is elegant. But it is not without potential cost.

One issue here is the threat of implied parallelism. If we get to assert genuine mental causation only on the grounds that mental properties are causally relevant to mental effects but not physical effects, then this suggests that we cannot assert genuine mental causation of physical effects. If this is the only means of resisting exclusion of mental properties as efficacious, then the implication is that genuine mental causation only obtains between mental-behavioural properties. This may not trouble everyone, but in Chapter 3, we will see reasons for thinking it mistaken to rule out diagonal causation from the mental to the physical.

Another reason for caution is that it's unclear in what sense the mental property is doing causal work. The absence of competition between the horizontal (mental to mental; physical to physical) relations is bought on the basis of the causes' distinct causal relevance for distinct properties of the effect. But it is not obvious that distinct causal *relevance* implies distinct causal *efficacy*.[16] It may be plausible that the event qua explosion in the basement is causally relevant for the fire where the same event qua loudest noise this morning is not. But this does not settle the question of efficacy: it is also plausible that if the event qua loudest noise this morning had not occurred, then the fire would not have. We might therefore worry that even if mental causes are not squeezed out as relevant for effects, this does not secure their causal

[16] Some see a genuine distinction here, and take it as a positive feature. Jackson and Pettit (1988; 1990), whilst not advocating precisely the solution under present discussion, argue that mental properties derive causal relevance from their physical realisers, with only the latter being causally efficacious. In this way, we might preserve mental causal explanation, whilst respecting Completeness and Exclusion. But this may not satisfy those who want mental causation to be more robust than this epistemic notion suggests.

efficacy with regard to those effects. In which case, we might yet worry that epiphenomenalism has not been diverted.

Whilst epiphenomenalism remains a concern here, we should also note that the role of supervenience in the Exclusion Problem is here taken for granted. Autonomy solutions do not question the alleged nature or function of that relation in contributing to the prospect of causal exclusion.

Compatibilist Solutions

Another approach to the Exclusion Problem attempts to solve it by claiming a metaphysical relation, between mental causes and their physical realisers, that respects both their distinctness and the genuine efficacy of the mental. Such relations achieve this because they purportedly secure a form of mental causal determination that is compatible with the Exclusion principle. The causal determination entailed by genuine mental causation does not entail overdetermination.

One early example of this approach was proposed by Yablo (1992).[17] On this view, the key relation between the mental cause and its subvening physical base is the *determinable-determinate* relation. Examples of the relation would be that holding between being red and being crimson, or between being shaped and being round. Causally efficacious mental properties are taken to be determinables, of which the subvening physical properties are determinates. A set of C-fibres firing is a particular, determinate way of being in pain. Mental determinables are then taken to inherit their causal efficacy from their particular physical determinates, securing their status as genuine causes. And the determinable-determinate relation putatively shares formal properties with that of supervenience: the former is plausibly a form of metaphysical dependence-determination, and is plausibly compatible with multiple realisability.

The argument for compatibility between mental causation and the Exclusion principle is then straightforward: we do not typically regard entities related as determinable to determinate

[17] Other, more recent, examples include: Bennett (2003), Shoemaker (2007), Wilson (2009)

to overdetermine effects. The bull's rampage is not overdetermined by the rag's being both red and crimson. If so, then we have grounds for taking the putative overdetermination at work in cases of mental causation to be benign.

As before, one might question to what extent this approach, and others in a similar vein, successfully deliver genuine causal efficacy for mental properties. Indeed, proponents must walk a fine line. The tightness of the metaphysical relation invoked, e.g. the determinable-determinate relation or the set-subset relation[18], is in danger of being so tight that it undermines conviction in the mental property's being efficacious in its own right. But without a sufficiently tight relation, there is no basis for claiming that the mental and physical properties do not compete.

A related concern is that the specific nature of the proposed metaphysical mental-physical relation might not deliver the requisite causal efficacy. This worry, as applied to the present version of the approach, is that the determinable might look to be efficacious only insofar as it is realized by a determinate property with the right causal capacity. Whilst the rag's being both red and crimson plausibly does not overdetermine the bull's rampage, it is not clear that the rag's being red is causally efficacious. One thought here is that, on the assumption that the bull is enraged only by crimson, the rag's being red appears consistent with the bull's stolid refusal to move. For the rag's being red could be determined by its being light scarlet, which by hypothesis, will not cause the bull to charge. But if so, then this suggests that the determinable, being red, is not causally efficacious even in the crimson case, for is not redness but crimsonness that makes a difference to whether the bull charges. On the other hand, consider the pigeon trained to peck at any red patch and only red patches. When presented with a light scarlet patch, the pigeon pecks; when presented with crimson, the pigeon again pecks. When presented with a green patch, the pigeon is unmoved. In this case, the difference seems to be not the determinate shade but rather, the determinable, being red.[19] Given considerations such as these, it is not clear what we should say about the causal efficacy of determinable mental properties, especially given that those properties will plausibly be themselves determinates relative to some wider mental determinable.

[18] Wilson (2011), building upon earlier work from Shoemaker (2001), proposes that the causal powers of the mental property is a proper subset of those of its physical realiser.
[19] Wilson (2009) argues in this direction.

In addition to worries about causal efficacy, this approach hangs on the plausibility of whichever metaphysical relation is proposed to do the job of defeating intuitions of causal competition. As such, these views are open to the criticism that the candidate relation fails to apply to mental causes and their supervenience bases. Some have rejected the determinable-determinate relation on these grounds.[20] For example, Cox (2008) claims that the mental properties cannot be related to their physical realisers as determinable to determinate, because whereas determinates are mutually exclusive, neural properties are not.[21] On that basis, physical (i.e. neural) properties are not suitable candidates for being determinates relative to mental properties. For example, no object can be both red and blue at the same time and in the same area. But, so the objection goes, neural properties are not like this. Whilst an attempt to merge the edges of a blue area with those of a red area would result in effective dis-instantiation of both determinate colours (yielding an area that was neither red nor blue), the same is not true for neural properties. On this line of thought, there is no logical or metaphysical impossibility in there being two neural properties that spatially overlap, e.g. two simultaneous neural instantiations of a mental pain property that share parts of their respective neural networks. But there is a logical / metaphysical impossibility in two determinates of a determinable instantiating at one and the same time in the same place, and hence in their overlapping. For that reason, neural states are not plausible candidates for being determinates for mental determinables.

As with autonomy solutions, compatibilist responses do not typically concern themselves with the nature or function of the supervenience relation. Instead, the focus is upon other synchronic mental-physical relations that might permit evasion of the Exclusion principle. Rather than question the alleged implications of supervenience, they are taken up with explicating other relations that might excuse the mental from overdetermining effects apparently shared with the physical.

[20] Ehring (1996), Funkhouser (2006) and Cox (2008) each take this approach.

[21] It is an open question whether the same line of argument would work when the physical property is characterized at a lower level. However, denying the relation for this level of physical property might be sufficient for denying it in connection with a lower one. In any case, I cite Cox (2008) here by way of illustrating that this approach to the Exclusion Problem is potentially problematic.

Interventionism

The last approach to consider is from within an Interventionist framework (Woodward, 2003; 2015). This framework will be of particular interest in later chapters, where I propose and critique an Interventionist model of the Exclusion Problem and subsequently argue against a recent Interventionist solution to the problem (Campbell, 2020). Here, we confine ourselves to a brief introduction by way of showing some of what is at stake in this approach.[22]

One way of thinking about an Interventionist solution is to see it as rejecting the Exclusion principle[23] by claiming that both the mental and the physical properties, though not independent of each other, are causally efficacious with respect to the same effect.

Roughly, on an Interventionist view, a cause is a variable that can be intervened upon to manipulate an effect.[24] So, for example, pressing the light switch causes the light to come on if my pressing or not is correlated with the light coming on or not. More strictly, a causal candidate variable *c* is a cause of outcome variable *o* if and only if values taken by *c* are correlated with values taken by *o* under interventions upon *c*. An intervention assigns values to a variable (i.e. in this case, 0 for the light switch being off and 1 for it being on) whilst holding fixed all other relevant variables, such as variables that are independent causes of the effect. There might be, for example, two distinct switches at different points in the room for the same light. An intervention will successfully isolate the correlation between one of the switches and the lightbulb from the correlation between the other switch and the bulb. In this way, a variable is a cause of an effect when its values correlate with those of the effect under interventions that isolate that correlation.

On the proposed route to overdetermination above, it is P2 that is putatively overdetermined by M1 and P1. But according to the Exclusion principle, this cannot be so if M1 and P1 are not independently sufficient causes. An Interventionist response might run as follows. M1 is a cause of P2 if M1 is correlated with P2 under intervention. Let's suppose that

[22] There is no single Interventionist response to the Exclusion Problem. Woodward (2015) defends his view, to which Baumgartner (2010) objects, and Campbell (2020) articulates a different strategy. Indeed, there is no single Interventionist account of the problem itself. In Chapter 2, Section 5 I suggest limitations on an Interventionist formulation of the problem, and in Section 6, argue that the problem, when formulated in this way, requires a variety of problematic assumptions.

[23] Exclusion: No effect has more than one sufficient cause unless it is genuinely overdetermined, i.e. unless overdetermined by independently sufficient causes.

[24] Woodward (2003).

M1 is the mental property of feeling a sudden toothache, and P2 is the physical property subvening M2, which is my clutching my jaw. P1 is the physical property subvening my toothache sensation. It is plausible that under intervention, if my toothache does not occur, then the physical effect subvening my jaw-clutching (P2) does not. And it is plausible that under intervention, if my toothache does occur, then so does physical P2. Given Completeness, the same holds for P1 with respect to P2. So in this framework, it is plausible that M1 and P1 both cause P2 because both are correlated with P2 under interventions. Assuming that P1 and M1 do not qualify as independently sufficient causes, this violates the Exclusion principle. Such an approach should therefore provide either an account of how the concerns that motivate the principle can be successfully assuaged under Interventionism[25], or grounds on which to show that the principle is not well-motivated.

One important point from the above is this: a straightforward Interventionist response to the Exclusion Problem is potentially open to epiphenomenalism from within its own principles.

Another salient point for us here is that an Interventionist approach to the Exclusion Problem turns upon issues relating to supervenience, and this has been to some extent appreciated. This is illustrated by Baumgartner's (2010) objection to Woodward (2003). He claims that M1's supervening upon P1 creates difficulties for the Interventionist account above. If M1 supervenes upon P1, then it looks like interventions upon M1 are impossible because when a value is assigned to M1, a value will also be assigned to P1. To use our example, if the status of one of the light switches in the room supervenes upon the status of the other, then it seems that switching on the first might thereby switch on the second.[26] But if so, then it looks like one might not be in a position to isolate the putative causal correlation between M1 and P2. Although supervenience is not the primary focus of an Interventionist response to Exclusion, and indeed is taken for granted as it was by other approaches, some of the challenges for the approach have rendered supervenience visible.

[25] My arguments through Chapters 3 and 4 will suggest that *some* violations of the principle might well respect the motivations for it. The reasons are not confined to an Interventionist approach. I address this specifically in Section 4b of Chapter 3.

[26] I say 'might' because in Chapters 2 and 4, I argue this sort of picture is overly simple.

Conclusion

We have seen the apparent indispensability of mental causation for our folk and scientific psychology. Absent causal efficacy, our mental lives, social interactions, relationships, and agency as we understand them are all under threat. Plausibly, without that understanding, we could not understand ourselves or carry on as we do. But although we might typically assume psychology to be autonomous of physics, and vice versa, mental causation might come under threat from physical causation. If mental properties depend upon physical properties, and the latter populate a causally complete domain, then perhaps there is no real causal work for mental properties to do. Perhaps we have an Exclusion Problem.

I have claimed that supervenience is of central importance in generating this problem. Yet, many of the main strategies for responding to the problem have not focused upon it, and, as we have seen, those strategies are typically beset by persistent worries about epiphenomenalism. In the case of an Interventionist response, this neglect has also manifested in supervenience resurfacing as a critical issue. But we will see that this holds lessons not only for Interventionists but for other parties in the Exclusion debate. For by paying closer attention to the supposed role of supervenience in the Exclusion Problem, we can access a plausible dissolution of the problem. In Chapter 2, I will make the case for that. In Chapter 3, we will see that once the role of supervenience has been explicated, we are free to affirm diagonal mental causation without fear of problematic overdetermination. And, in Chapter 4, we will see how neglecting supervenience plays a key role in originating and sustaining Exclusion worries.

Chapter Two - Dependence, Determination, Intervention

Introduction

In Chapter 1, we saw that the Exclusion Problem has prompted non-reductive physicalists to offer a variety of potential solutions. These have often traded in distinctions, relations or conceptions of causation that open the door to further challenges. Prominent amongst these has been the challenge of avoiding diluted forms of mental causal efficacy that might be closer to epiphenomenalism than is comfortable. We also saw that proposed solutions have tended to neglect the nature and function of supervenience in the Exclusion Problem.

In this chapter, I take my cue from these two features of solutions to Exclusion. I examine the role of supervenience in ostensibly generating systematic overdetermination by way of diagonal causal relations. In so doing, I argue that such diagonal relations will only obtain on the basis of further assumptions that the non-reductive physicalist might plausibly resist. Consequently, the non-reductive physicalist need not – absent those assumptions – accept that systematic overdetermination is entailed by the conjunction of Exclusion principles[27]: Supervenience, Distinctness, Causation and Completeness. For this reason, the arguments in this chapter amount to a proposed dissolution of the Exclusion Problem for non-reductive physicalists: the terms of the problem need pose no threat to mental causation from physics.

My overall argument proceeds by offering three reconstructions of the Exclusion Problem, each intended as a means by which to focus in upon the role of supervenience. We can think of these models as answers to the question, how is it that supervenience, together with Distinctness, Causation and Completeness, might generate diagonal causal relations? As such, they supplement the core Exclusion principles (Supervenience, Distinctness, Causation, Completeness and Exclusion) with further principles by which to explicate the role of supervenience. I should stress here that the supplementary principles are not suggested as entailed by the core principles; they are proposed as auxiliary principles that might help to

[27] Distinctness: Mental properties are metaphysically distinct from physical properties.
Causation: Mental properties – qua mental properties – are causally efficacious.
Completeness: All physical effects have sufficient physical causes.

articulate the thinking of those who take the core Exclusion principles to imply diagonal causation.

In Section 1, I introduce and explain the first of these models: the Dependence model.

In Section 2, I explain why the potential routes, under the model, to overdetermination only obtain if further assumptions are granted. The upshot is that the conjunction of principles in the Dependence model (excluding the Exclusion principle itself) do not themselves entail systematic overdetermination; to do so, they require further substantive assumptions which the non-reductive physicalist has reason to deny.

In Section 3, we look at another potential Exclusion model: the Determination model. Again, we will see (in Section 4) that the model only implies systematic overdetermination given further controversial assumptions.

Our third model – Intervention – is introduced and explained in Section 5. As with the preceding models, the putative routes to systematic overdetermination are only open on the basis of supplementary assumptions (Section 6).

In each case, we will see that supervenience is critical to potentially generating diagonal causal relations from the mental to the physical and vice versa. By the same token, it is supervenience that fails, without further assumptions, to do the job.

Section 1: The Dependence Model

This model attempts to articulate the Exclusion Problem by explicating the role of supervenience in generating ostensible overdetermination. It does so by adding three auxiliary principles to the core Exclusion propositions. The central idea is that synchronic dependence relations between mental and physical properties link with diachronic dependence (i.e. causal) relations m-properties and between p-properties, respectively.

The additional principles taken together with the core Exclusion commitments yield an account of diagonal causation along the lines of Lewis's theory of standard causation, i.e. in terms of chains of counterfactual dependence relations. Crucially, amongst the core Exclusion principles, it is Supervenience that provides both the synchronic dependence relations required for the transitive links in those chains and the generalising function that threatens

systematic overdetermination. Consequently, the Dependence model (hereafter, 'Dependence') is one way that the proponents of the Exclusion Problem might conceive the role of synchronic relations in generating the threat of problematic overdetermination.

Dependence Model

The Dependence Exclusion model comprises the following:

Dependence Assumption (DA): If M supervenes upon P, then M is counterfactually dependent upon P.

Dependence Chains: A dependence chain is any finite sequence of events a,...n such that b is counterfactually dependent upon a, c is counterfactually dependent upon b etc.[28]

Causal Dependence: *c* causes *e* if there exists a dependence chain leading from *c* to *e.*

Core Exclusion Principles:

Supervenience: Mental properties supervene upon physical properties.

Distinctness: Mental properties are metaphysically distinct from physical properties.

Causation: Mental properties – qua mental properties – are causally efficacious.

Completeness: All physical effects have sufficient physical causes.

Exclusion: No single effect can have more than one sufficient cause occurring at any given time—unless it is a genuine case of causal overdetermination.

Routes to Diagonal Causation & Systematic Overdetermination

How do we get diagonal causal relations from these principles? Below are the specific routes to diagonal causation that might be taken as implied by the above principles. In each route,

[28] The Causal Dependence principle requires transitivity of counterfactual dependence such that if, for example, b is dependent upon a, c is dependent upon b, and d dependent upon c, then c is dependent upon a. So the obtaining of a dependence chain between c and e, plus the counterfactual conditions for causation, will imply a causal relation between c and e.

the diagonal relation is delivered by the forming of a dependence chain that comprises links of counterfactual dependence relations. Dependence chains are built, transitively, from such links (see the Dependence Chains principle). According to the Causal Dependence principle, if there obtains a dependence chain between two property instantiations, then those instantiations are related causally. So roughly, dependence chains are built from counterfactual dependence relations between property instantiations, and the properties at different levels are then connected by those chains, resulting in causal relations between properties at different levels; hence, 'diagonal' causation.

Here's an example for one of the routes. I form the intention to lift up my laptop, and then do so. On the Dependence view of Exclusion, my lifting the laptop supervenes upon some physical property(ies) and so counterfactually depends upon the instantiation of that physical property. But that physical property has a sufficient physical cause, because the physical properties are all subject to causal completeness. No physical instantiation lacks a sufficient physical cause. And that physical effect is thus counterfactually dependent upon its physical cause. So we have two physical properties in play – one as the property subvening my lifting the laptop, the other as the sufficient cause of that subvening property – related by counterfactual dependence. We have two counterfactual dependence relations: one holding in virtue of the supervenience relation between my lifting the laptop and its subvening property, and one holding in virtue of the causal relation between the subvening property and its physical cause. Now we join those two relations together to form a dependence chain connecting the lifting of the laptop to the physical cause of the property that subvenes the lifting of the laptop. By transitivity of dependence chains, my lifting of the laptop is counterfactually dependent upon the physical cause of its subvener. And that means that my lifting of the laptop is caused by that physical cause. We have diagonal causation. But now there is a problem; my lifting of the laptop was caused by my intention to do so. So it seems we have causal overdetermination: my lifting of the laptop is caused by both my intention to lift it (something mental) and the physical cause sufficient to bring about the property that subvenes the lifting (something physical).

In more detail, here are the routes to diagonal causation allegedly implied by the Dependence model of Exclusion.

Scenario (1): First Route to Overdetermination

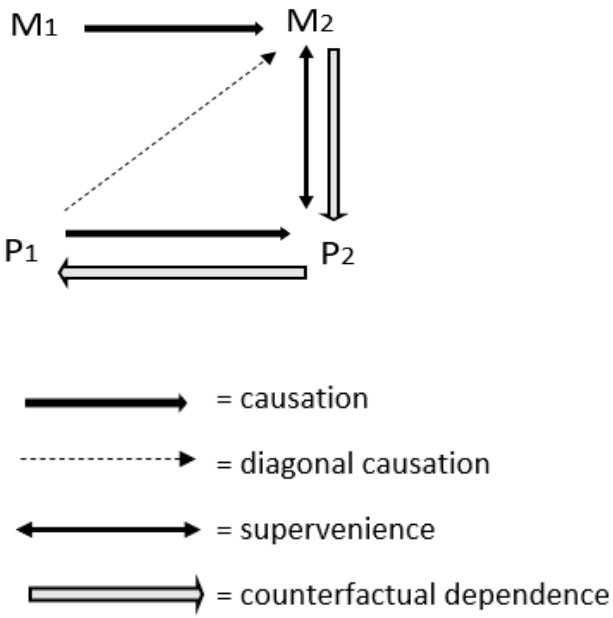
In accordance with the Causation principle, suppose that mental property M1 causes M2. Given the sufficiency of counterfactual dependence[29] for causation (i.e. the Causal Dependence principle), let us suppose that the M1 to M2 causal relation obtains in virtue of counterfactual dependence.[30] On the basis of Supervenience, M2 supervenes upon physical P2. Now, the Dependence Assumption enters: if M2 supervenes upon P2, then M2 counterfactually depends upon P2. So we have a counterfactual dependence relation between M2 and P2. P2, as a physical property that we assume to have a cause, is covered by the Completeness principle, so has a sufficient physical cause. This is P1. Again, we assume that this causal relation from P1 to P2 obtains on account of a counterfactual dependence relation; P2 counterfactually depends upon P1. Although we started by noting a causal relation between the mental properties, M1 and M2, this relation is *not* part of the dependence chain that delivers the diagonal causal relation here. The relevant dependence chain is formed, in virtue of the transitivity of dependence relations, from the two dependence relations between M2 and P2, and between P2 and P1 – as per the Dependence Chains principle. We have a dependence chain linking M2 to P1 via P2. Given the Causal Dependence principle, this dependence chain supports the causal dependence of M2 upon P1. We have a *diagonal causal relation* from P1 to M2 **(Figure 2.)**

On the typical account of Exclusion, this is taken to imply overdetermination of M2 by P1 and M1. That's because in addition to the diagonal causation of M2 by P1, we also have the horizontal causal relation that we started with – the relation between M1 and M2. On this route, we get the result that M2 is caused by both M1 and P2.

[29] The Dependence model assumes Lewisian conditions for counterfactual dependence:
Event e counterfactually depends upon e* iff: (1) It is the case that [if e* had occurred then e would have occurred] (2) It is the case that [if e* had not occurred then e would not have occurred].
These counterfactuals are evaluated as True, at the actual world, by close worlds such that for (1), there is a close world where [e* & e], and for (2), there is a close world where [¬e* & ¬e].

[30] This will be a constant supposition of the Dependence model. For every case of causation discussed, we suppose a counterfactual relation between cause and effect.

(Fig. 2)

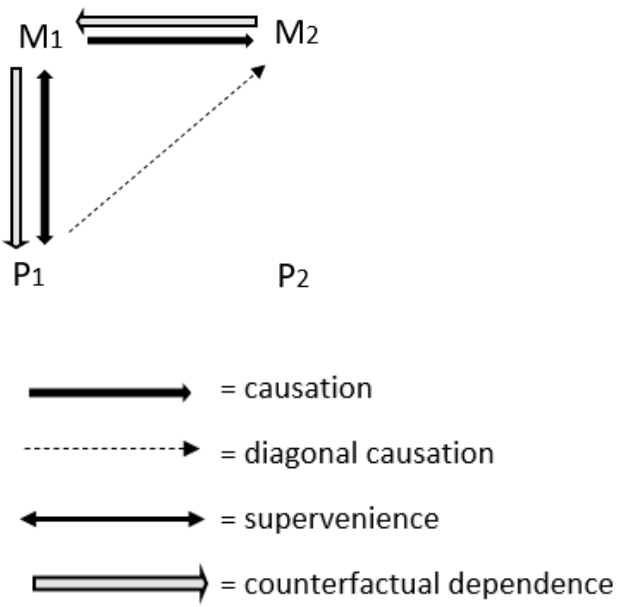


Scenario (2): Second Route to Overdetermination

On the second route, the causal (i.e. counterfactual dependence) relation between M2 and M1 is significant in building the relevant dependence chain. Mental to physical supervenience is taken to imply that the mental cause, M1, supervenes upon some physical property, P1. So by the Dependence Assumption, M1 is counterfactually dependent upon P1. We have two salient counterfactual relations: M2 upon M1 and M1 upon P1. These form a dependence chain, as per the Dependence Chains principle, from M2 to P1 via M1. Then, as per the Causal Dependence principle, we have a diagonal causal relation: P1 causes M2 **(Figure 3)**.

But of course, we started by asserting a causal relation between M1 and M2. So we appear to have causal overdetermination of M2 by M1 and P1.

(Fig. 3)



Clarifications

Whilst the Dependence model is obviously indebted to Lewis's original account of causation, we should also note that Causal Dependence differs from Lewis's condition for causation; Causal Dependence only entails that dependence chains are sufficient for causation, whereas Lewis's condition is both necessary and sufficient. Sufficiency is arguably all that is required to get overdetermination worries going, and is less controversial than claiming necessity as well. I wish to be as generous as possible to the proponent of Exclusion scenarios; imposing a necessary and sufficient condition on causation here would be to commit them to more than is apparently required.

We should also note that the Causal Dependence principle, taken together with the Dependence Chains principle, implies that a single synchronic counterfactual dependence relation qualifies as causal. Since a chain can comprise a single dependence relation, and since, given the Dependence Assumption, a supervenience relation between M and P will constitute just such a relation, it follows that a chain can consist of a synchronic dependence relation from M to P. But such relations are not typically taken as *causal*. So we want to add

this caveat to the model: *dependence chains built from only synchronic dependence relations do not satisfy the Causal Dependence principle.*

Two further clarifications: the two routes to overdetermination articulated above are not intended as mutually exclusive. They are only articulated as distinct routes for the sake clarity. They are, ostensibly, minimal routes to diagonal causation, and thus causal overdetermination. They are the simplest application of the Dependence principles to yield apparent causation between levels. But the obtaining of one route does not rule out the obtaining of the other. Furthermore, the model itself is not intended as exclusive of other models. The suggestion is not that this model might be *the* way of articulating Exclusion for some; rather, it is one way and for all I say, potentially compatible with some other model.

Why Dependence?

Why put this reconstruction of Exclusion in terms of counterfactual dependence at all? Chiefly because, of all the principles forming the conjunction at issue in the Exclusion Problem, it is supervenience that appears to play the crucial role in implying the borrowing of causal relations from corresponding events at different levels. None of the other core Exclusion principles appears to warrant a connection between distinct mental and physical properties. Supervenience presumably does this because of one or both of its core components: dependence and determination. And if, as per the Dependence model, the supervenience relation – as distinct from the causal relation – enters into chains of dependence with causal relations, then we may need a relation which is plausibly common to both. The present model focuses upon the dependence aspect of supervenience as a common element between that relation and causation.

Another key intuition behind overdetermination worries seems to be that relations between mental and physical properties, in virtue of m-properties supervening upon p-properties, *are in competition* with the causal relations holding between properties within the same domain. That is, if we have mental-physical or physical-mental (ostensibly) causal relations obtaining in virtue of supervenience, then those relations compete with the relevant causal relations obtaining between the mental properties or between the physical. So thinking of these inter-level, diagonal relations in terms of counterfactual dependence allows us to accommodate

that intuition: they appear to compete with horizontal, intra-domain causal relations because they are both relations of counterfactual dependence. Because counterfactual dependence is here stipulated as only sufficient, not necessary, for causation (diagonal or horizontal), this permits a means of accommodating the competition intuition without offering a controversial account of causation whereby both cross-domain and inter-domain relations are somehow relations of the very same kind.

Furthermore, framing the Exclusion Problem in terms of counterfactual dependence also has theoretical benefits: It frames the issue in terms that (a) avoid having to specify or justify any particular theory of causation and which (b) would be amenable to those inclined to accept counterfactual dependence as sufficient for causation. Not only does this mean that we can proceed without getting bogged down in metaphysical debates regarding causation, it more importantly means that our claims regarding the assumptions underpinning overdetermination worries are not dependent upon any particular theory of causation – and thus not vulnerable to objection from positions that subscribe to other theories.

Section 2: Latent Assumptions of the Dependence Model

Dependence Assumption: If M supervenes upon P, then M is counterfactually dependent upon P.

As we have already seen, one principle required for Exclusion scenarios to yield apparent overdetermination is the Dependence Assumption principle (DA). In order that the Exclusion scenarios result in overdetermination (given the other conditions), the relevant M properties must counterfactually depend upon the relevant P properties. The DA principle takes the relevant higher-level properties and connects them to their subvening physical properties via counterfactual dependence. Without the DA principle, the model has no means by which to implicate properties at higher levels as included in dependence chains with properties at the physical level. It would have the Supervenience principle, which certainly does imply relations between higher- and lower-level properties. But those relations need to be connected to dependence chains by the Dependence Assumption.

However, as will become clear, this application of the DA principle requires a further assumption. This assumption is required in order for the DA principle to be *true*. The proponent of the Dependence model must assume that mental properties are not *closely multiply realisable* – that is, not multiply realised at close counterfactual worlds. But I will argue that this is not a tenable assumption for the kind of physicalist for whom my arguments are intended.

In my formulation of the Dependence model, I have thus far stated the Supervenience principle in rough terms. But we should now be more precise. For the sake of argument, let's assume that the Dependence model construes the Supervenience principle along the lines of the following, standard formulation of *global* supervenience:

GS: For any possible worlds, w and w*, if w* is a physical duplicate of w then w* is a mental duplicate of w.

There are well-known issues with such a formulation (see e.g. Kim, 1993; Jackson, 1998), and if we were engaged in formulating an adequate thesis for the articulation of physicalism, then this would require revision. However, this captures what we need here. We assume global supervenience for our present purposes, since this is commonly cited as suitable (if not necessarily sufficient) for physicalism (McLaughlin, 1992; Jackson, 1998).[31] Physicalism is supposed to be a thesis, if it is a thesis[32], about what the world is like – not merely individuals or particulars within a world – and moreover what the world must be like, given the nature, distribution and laws of the physical entities. So some form of global supervenience thesis seems apt for the job of expressing the relations holding between properties at different ontological levels. We therefore assume that m-properties globally supervene upon physical p-properties.

[31] Our argument against the Dependence model will not turn upon whether the supervenience thesis is global or otherwise. It would apply if the thesis were local.

[32] Some, such as Ney (2008), argue that it is an attitude whereby one's ontology is formed on the basis of what physics says exists. Nothing here turns upon the issue of whether physicalism is a thesis or an attitude.

Now to explicating the implicit assumptions involved in the Dependence model. We have said that, on this model, the Exclusion scenarios will only result in overdetermination if the DA principle holds. We have also specified that the supervenience thesis at issue is one of global supervenience. We will see that, in order for the DA principle to be true, a number of assumptions are necessary, because there are plausible reasons for thinking it false: reasons stemming from the prospect of close multiple realisability of mental properties.

To show this, we need to show that global supervenience is compatible with multiple realisability of M across close counterfactual worlds. It is worth noting the importance of *close* multiple realisability. For given multiple realisability of M, supervenience might yet (in conjunction with the other assumptions above) entail overdetermination if M could not be differently realised at *close* counterfactual worlds. If M were multiply realisable – but *not* multiply realised at close possible worlds – then the DA principle would still hold. If M were like this, and M supervened upon P, then it would be true that M counterfactually depends upon P. To see this in more detail, we should recall that the Dependence model assumes a Lewisian semantics for counterfactuals. So the counterfactual dependence of M on P requires the truth of two conditionals:

(i) If P had occurred then M would have occurred.

(ii) If P had not occurred, then M would not have occurred.

Crucially, as per Lewis, these conditionals are to be evaluated at close counterfactual worlds. Therefore, if M is *not* alternatively realised at close counterfactual worlds – and so, *a fortiori*, not alternatively realised at close worlds where P is absent – then the second requisite conditional will be true. And the supervenience thesis itself is apparently sufficient for (i) to be true since if any counterfactual world w* is a physical duplicate of the actual world, then it is a mental duplicate – thus if duplicate P* occurs at a close counterfactual world, then duplicate M* will occur. And if (i) and (ii) are true, then the way remains open for the above Dependence principles (i.e. the core Exclusion principles in conjunction with the Dependence Assumption, Dependence Chains and Causal Dependence principles) to entail overdetermination in the Exclusion scenarios. Thus, to block that entailment, we need the second conditional to be false at close counterfactual worlds, that is, we need M to be alternatively realisable at close counterfactual worlds.

So, to the first requirement: showing that global supervenience is compatible with multiple realisability of M. Global supervenience entails only that, for any possible worlds, w and w*, if w* is a physical duplicate of w then w* is a mental duplicate of w. This is compatible with M being multiply realisable. Let the actual world of the Exclusion scenario be world w and the counterfactual world at which P does not occur – but (non-duplicate, distinct) P* does – be world w*. Since w differs from w* in its distribution of p-properties, the above formulation of global supervenience is silent as to whether M is instantiated in w*. As such, the formulation is compatible with multiple realisability of M. Global supervenience entails sameness of global mental property distribution between worlds with the same global physical property distribution, and this is consistent with two worlds (i.e. the actual world and a counterfactual world) having the same mental property distribution but different physical property distribution.

It is true that global supervenience implies that differences between global mental property distributions necessarily depend upon differences between physical property distributions, since sameness of mental distributions holds for any two possible worlds with the same physical property distribution, i.e. the two dimensions of sameness hold across all possible worlds, implying that any difference in mental distributions holds in virtue of differences in physical property distributions. So global supervenience implies that differences of mental distribution across worlds entails differences of physical distribution. But this is distinct from the differences in physical property distributions entailing differences in mental distributions. Therefore, global supervenience is consistent with mental properties being realised by different physical properties across worlds. It is consistent with m-properties being multiply realisable.

To illustrate, let's take a possible world w, comprising a distribution of just two physical properties (P1, P2) and two mental properties (M1, M2), distributed as ordered pairs <P1, P2> and <M1, M2>. Let us suppose that global supervenience is true. Global supervenience implies that a difference in m-property distribution between w and some possible world w* depends upon a difference in distribution of p-properties (where such difference could include non-instantiation of either or both of the mental and either or both of the physical properties. But this is consistent with there being a possible world w** at which M1 and M2 are distributed in the same way they are in w, but at which the physical property distribution

consists of ordered pair <P3, P4>. For the mental property distribution at w to entail the physical distribution, the following conditional must hold true for all possible worlds: if <M1, M2> then <P1, P2>. But world w** is a world at which the conditional is false. It is a world at which we have <M1, M2> but not <P1, P2>; rather, we have physical distribution <P3, P4>. Hence, global supervenience is consistent with multiple realisability of M-entities.

What of other supervenience theses? Might these rule out multiple realisability? I think the answer here is 'no'. Any supervenience thesis will allow for the possibility of multiple realisability of the supervening entities, because any supervenience thesis will involve asymmetry such that the base level fixes the supervening level, but not vice versa.

I have claimed that multiple realisability of m-properties is consistent with supervenience. But I said above that it was important that m-properties are multiple realisable across *close* worlds. I take that our considerations above are sufficient to show that supervenience is compatible with this, too. We can put the matter this way. The consistency between supervenience and multiple realisability holds in virtue of supervenience entailing only upward necessitation of mental properties by physical ones; supervenience does *not* entail downward necessitation, such that if some M instantiates then some subvening P must instantiate. This being so, it follows that supervenience is consistent with close multiple realisability. A supervenience thesis itself places no constraints – at worlds close or otherwise – upon which P co-instantiates with which M.

Close Multiple Realisability

The proponent of Dependence routes to overdetermination must make at least one of two substantive assumptions in order to establish the viability of the Exclusion Problem. They must assume that mental entities are not multiply realisable at all, or that they are not multiply realisable at close counterfactual worlds. Neither of these positions should be dismissed outright.

There are those who reject the plausibility of multiple realisability claims for mental properties (e.g. Bechtel and Mundale, 1999; Polger, 2002) and I will not here attempt to provide persuasive arguments against them. And my claim in favour of multiple realisability

of m-properties at close counterfactual worlds will turn upon – amongst perhaps other things – the proper individuation of underlying physical properties/events. But equally, nor are these claims (i.e. against multiple realisability in general or against its manifestation at close worlds) just obviously true; they count as substantive assumptions on the part of those who accept the Exclusion Problem as explicated here.

The important point to see is that these substantive assumptions may be in tension with other commitments of the non-reductive physicalist who endorses the core Exclusion principles. One reason for the potential tension is the typical status of multiple realisability considerations in motivating non-reductive physicalist views. Historically, one of the key factors in motivating rejection of reductivism – i.e. type identity theory – and optimism regarding functionalism was the ostensible multiple realisability of mental types (Putnam, 1967). Insofar as this is a substantive motivation for those non-reductive physicalists who accept the core Exclusion principles, there will be little temptation to reject multiple realisability. That of course is not to say that they will advocate *close* multiple realisability. But it gets them some way there.

Furthermore, any proponent of the core Exclusion principles must accept the Distinctness principle: mental properties are metaphysically distinct from physical properties. But it is not clear why one would hold Distinctness and deny multiple realisability of mental properties. If, as seems the case, multiple realisability entails that mental properties are not type-identical to physical properties, then multiple realisability entails Distinctness. This does not force the proponent of Distinctness to accept multiple realisability, of course. They can coherently accept the former and reject the latter. But there is a question as to what could motivate rejecting multiple realisability in the absence of rejecting Distinctness. If we accept that mental properties are distinct from physical realisers, then on what grounds would we be inclined to insist that those mental properties are uniquely realisable?

These factors introduce a potential tension between non-reductive physicalism, sympathetic to the core Exclusion principles, and the rejection of multiple realisability. If on these grounds, our physicalist will endorse multiple realisability, then they must – if they want to push for the Dependence model – find reasons to dismiss *close* multiple realisability even whilst accepting it more generally. Again, it is difficult to see what these reasons would be. We do not here offer arguments to show the absence of any such plausible reasons. We only point

out that, absent those reasons, the non-reductive physicalist is advised to at least retain close multiple realisability as an open possibility. In which case, she is advised to deny that diagonal causation results from the Dependence model of Exclusion.

Summary

Let's take stock. The Dependence model is one potential way of interpreting the Exclusion Problem, ostensibly giving rise to two routes to diagonal causation and systematic overdetermination. It attempts to articulate the way in which the core Exclusion principles (Supervenience, Distinctness, Causation, and Completeness) are supposed to entail diagonal causal relations between the mental and physical levels. It does this by supplementing those core principles with the Dependence Assumption, Dependence Chains, and Causal Dependence principles. Of these, it is the Dependence Assumption principle that is key in taking a supervenience relation and yielding dependence relations between levels that then contribute to the dependence chains required for diagonal causation.

I have argued that the Dependence Assumption principle is only true if it is false that mental properties are closely multiply realisable; therefore, the proponent of the Dependence model of Exclusion must deny close multiple realisability.

Now, we should here bear in mind that we are interested in appealing to a particular audience. We want to show that a non-reductive physicalist can endorse the Exclusion principles[33] without inconsistency, i.e. without thereby committing to systematic overdetermination. In the context of the Dependence model, this requires that we provide reasons for thinking that the non-reductive proponent of the core Exclusion principles will not want to deny close multiple realisability. In other words, we should show why she should not rule it out.

I think we have done so.[34] We are *not* here claiming that the non-reductive physicalist, sympathetic to the core Exclusion principles, must endorse close multiple realisability of

[33] Supervenience, Distinctness, Causation, Completeness and Exclusion.

[34] It's worth nothing that we have also shown that one can endorse the conjunction of core Exclusion principles plus the Dependence Chains and Causal Dependency principles consistently with endorsing the Exclusion principle. To do this, one needs to show that the falsity of the Dependence Assumption principle is consistent with those principles. The DA principle is a conditional that is false if the antecedent is true and the

mental properties. We only intend the more modest claim that she has reason to retain it as an open possibility. This is enough to block full-blown endorsement of the Dependence model as a means of entailing diagonal causation and systematic overdetermination.

We can illustrate the upshot of this section by once again considering one of the putative routes to overdetermination under the Dependence model. Let's consider the first Exclusion scenario above. On that route, M2 is allegedly overdetermined by P1 and M1. The crucial move in this route is made by applying the DA principle to yield a diagonal causal relation between P1 and M2: since M2 counterfactually depends upon M1, and since M1 counterfactually depends – according to the DA principle – upon P1, we have a dependence chain between P1 and M2 via M1. But if M2 is multiply realisable at close counterfactual worlds, then at a close world, M1 is realised by not P1, but P*. Thus M1 is not counterfactually dependent upon P1. A similar move is, on the supposition that DA is false because close multiple realisability holds, available for the other scenario. I have argued that the non-reductive physicalist has reason to withhold assertion of the DA principle, and so hold open the above strategy for blocking systematic overdetermination given the Exclusion principles.

Section 3: The Determination Model

The Determination model is the second form of reconstruction for the Exclusion Problem. Like the Dependence model, it attempts to explicate the intuitive appeal of the Exclusion scenarios as typically – and roughly – conceived. And again, like the model above, it does so by focusing upon the role of mental-physical supervenience in those scenarios. And again, whereas overdetermination in the Dependence model turns upon the production of two

consequent false. Because we are assuming the position of the non-reductive physicalist sympathetic to the core Exclusion principles, we assume that supervenience holds. So we assume the antecedent is true. The consequent is false if the relevant m-properties are multiply realised at close counterfactual worlds; so the DA principle is false if m-properties are closely multiply realisable. We have claimed that supervenience is consistent with close multiple realisability; our supervenience thesis does not entail counterfactual dependence of supervening m-properties upon their p-property base. Now, I take it that no other principle, or combination of principles, in the Dependence model (excluding the DA principle) entails the truth of the DA principle. The only prima facie suitable candidate was the Supervenience principle itself. If so, then the consistency of the Supervenience principle and close multiple realisability is sufficient to show that the core Exclusion principles (plus the Dependence Chains and Causal Dependence principles) do not entail the truth of the DA principle. They are compatible with its being false. But if so, then this conjunction of Exclusion principles, under the Dependence model, do not entail diagonal causation and hence, do not entail systematic overdetermination.

competing relations of determination – what I have called horizontal and diagonal causation - the Determination model yields overdetermination through the production of similarly competing relations. It does so by taking the core Exclusion propositions and adding two further principles, Determination and Assimilation, as follows:

Determination Model

Supervenience: Mental properties supervene upon physical properties.

Distinctness: Mental properties are metaphysically distinct from physical properties.

Causation: Mental properties have – in their own right qua mental – genuine causal efficacy.

Completeness: All physical effects have sufficient physical causes.

Exclusion: No single effect can have more than one sufficient cause occurring at any given time—unless it is a genuine case of causal overdetermination.

Determination: If M supervenes upon P, then P necessarily determines M.

Assimilation: If M is necessitated by P, M is a relatum (of the same type, i.e. cause or effect) in any causal relation where P is a relatum.[35]

Below are the routes to overdetermination potentially implied by the Determination model. The rough idea is that, where a mental property M supervenes upon a physical property P, the physical property necessitates M, hence the Determination principles. In virtue of that upward necessitation, M gets to piggyback upon the causal role of the physical property, hence the Assimilation principle. (By 'causal role' I mean here just its being cause or effect; I do not intend the meaning associated with Functionalism.) M is thereby implicated in any causal relation in which P is a relatum. Through M's piggybacking upon P, M gets to enter into diagonal causal relations with physical properties.

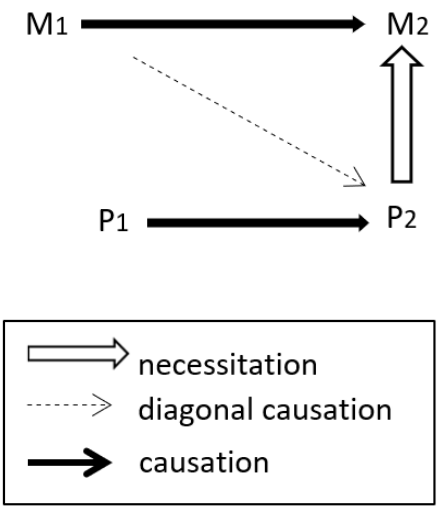
[35] This might appear too strong to be plausible. If so, then there is a weaker version that would still potentially contribute to diagonal causation via the second route below. The weaker version would be: **If M is necessitated by P, M is an effect in any causal relation where P is an effect.**

In more detail:

Scenario (1): First Route to Overdetermination

Suppose that mental property M1 causes mental effect, M2. Given supervenience of mental upon physical properties, M2 and M1 each supervene upon some physical properties. Call these P1 and P2. By Completeness, P2 must have a sufficient physical cause, P1. We suppose that P1, as subvener of M1, is the cause of P2 given that P2 is the subvener of M2. According to the Determination principle, M1 is necessitated by P1. And according to the Assimilation principle, M1 is thereby assimilated to the causal relation between P1 and P2. Hence, M1 is cause of P2: we have a diagonal causal relation. But P2 already has a sufficient physical cause, P1. So P2 is overdetermined by M1 and P1 **(Figure 4)**.

(Fig. 4) Route 1

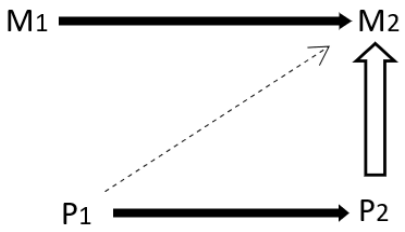


Scenario (2): Second Route to Overdetermination

Suppose that mental property M1 causes mental effect, M2. Given supervenience, M2 and M1 each supervene upon physical properties, P2 and P1. By Completeness, P2 must have a sufficient physical cause. Again, given their respective roles as supervenience bases for M1 and M2, P1 is assumed as the cause of P2. The Determination principle implies that M2 is necessitated by its supervenience base, P2. The Assimilation principle implies that M2 piggybacks upon P2's status as effect of P1, thereby becoming assimilated to the causal

relation between P1 and P2. M2 is implicated as effect of P1. But M2, by hypothesis, has a mental cause, M1. So M2 is overdetermined by P1 and M1 **(Figure 5)**.

(Fig. 5) Route 2



Clarifications

First, we should note that, unlike the Dependence model, I do not here stipulate even a sufficient condition for causation. This is perhaps both a strength and potential weakness. It is arguably a strength in that it might capture some of what is going on in worries about exclusion without relying upon any specific theory of causation. If it does, then it expresses those worries in a way that focuses on the role of supervenience and completeness rather than that of any specific causal notion. It might also be thought a strength since in principle, the Determination model is thereby more flexible and amenable to a range of perspectives depending upon one's preferred theory of causation. It is not ruled out as an articulation of Exclusion on the grounds of invoking a particular theory that parties might happen to reject. (Indeed, it is compatible, for example, with a Lewisian notion. We might plausibly think of the assimilation of mental properties into physical causal relations in terms of dependence: e.g. as per the first route above, M1 gets to cause P2 because assimilated to a dependence relation between P2 and P1.)

However, the absence of any particular notion of causation might also be a potential weakness insofar as there is no guarantee of the model's working for some conceptions. Still, no model is intended as universally appealing. This does not preclude Determination from doing its intended job of providing a potentially viable articulation of the Exclusion Problem.

A second, related point here is that, as previously with Dependence, the Determination model is not meant to be exclusive. That is, one might coherently interpret Exclusion using

this model and some other model. There is no obvious bar to reading it in tandem with Dependence, for instance. That is not to say, however, that it will work in combination with all such models. The last model of this chapter – the Intervention model – would not fit with the first route to overdetermination, for reasons that we will see when we come to it.

Equally, as with Dependence, the two routes to overdetermination here are also not intended as mutually exclusive. One might think of both routes obtaining together. They are distinguished for the sake of clarity and because generosity recommends that we do so; it might be that one of the routes is subject to criticism that the other is not. (This will not be my claim, however).

Why Determination?

The motivation for the Determination model is similar to that in support of the Dependence model. As stated with regard to that model, Exclusion worries have traditionally been generated by the conjunction of: Supervenience, Distinctness, Causation, Completeness, and Exclusion. It is Supervenience that appears to play the crucial role in suggesting to some that somehow, mental properties are *brought along with* their synchronically related physical properties so as to render them relata of competing causal relations. Whereas the Dependence model focuses upon the dependence aspect of supervenience, the Determination model builds upon the determination component.

Additionally, and again similarly to our motivation for the Dependence model, the Determination interpretation of Exclusion is consonant with the intuition of causal competition in putative cases of overdetermination. The intuition seems to be that the so-called 'causal' relations between mental and physical properties in virtue of the former supervening upon the latter are *in competition* with the causal relations holding between properties within the same domain. Framing the issue in terms of Determination respects this intuition, since the relations in competition are both nominally causal relations (albeit holding at different levels).[36]

[36] This is not to say that causal assimilation further elucidates or supports this intuition, since it leaves open the question of whether the higher- and lower-level causal relations are indeed of the very same type, or of types

Finally, the Determination model has a potential advantage over Dependence, since close multiple realisability does not interfere when the critical relation is upward necessitation. The key relation in the model is metaphysical determination (i.e. upward necessitation) of mental properties by their physical subveners. The multiple realisability, close or otherwise, of the mental properties is besides the point when the relation doing the work is upward. In the Determination model, it is the modal fixing of the mental properties by physical ones that helps to allegedly achieve diagonal causation. P's necessitating M is compatible with M's being closely multiply realisable.

Section 4a: Assumptions Required for Applicability of the Determination Principle

When we talk of the applicability or otherwise of the Determination principle, we mean this: the Determination principle is applicable if its antecedent is true. So if M supervenes upon P in some case, then the Determination principle is applicable to that case. To see the assumptions required for applicability, we should first notice that the antecedent of the Determination principle denotes the constants M and P. As constants, these denote specific mental and physical properties. The root of the problems for advocates of the Determination principle is the required individuation of these properties.

As the two mooted routes to overdetermination suggest, the Determination model will only yield apparent overdetermination if one or more of the constituent principles imply horizontal causal relations in putative Exclusion scenarios. Putting it roughly, all exclusion scenarios involve a convergence of causal relations upon some common effect, and that convergence is produced by a horizontal causal relation meeting a diagonal relation. If the Determination model is to deliver diagonal relation, then there must be some principle(s) in that model which implies horizontal causal relations at the physical level. And, of course, there is: the Completeness principle. The Completeness principle underwrites the implication of physical causal relations whenever we have a physical effect.

---

which are coherently thought to be competing. But no matter; the purpose of the model is merely to reconstruct the Exclusion Problem in somewhat more explicit terms.

Because of the role of Completeness in putatively generating diagonal causal relations when in conjunction with the principles, the advocate of Determination needs the physical properties denoted in the Determination principle to be properties individuated by the causal relations of physics. Indeed, plausibly, these properties need to be properties individuated by the laws of physics. Why? Completeness guarantees that for any physical effect, there is a sufficient physical cause. It is this guarantee that entails horizontal physical causal relations whenever a physical effect is implied by the other principles in the Exclusion model. But the physical causes thereby entailed are properties individuated by causal laws of physics. That's because Completeness, if it is not to be arbitrary, is motivated by commitment to maximal generality in physics. That is, Completeness might be seen as expressing a commitment to physical laws that are exceptionless in their scope, underwriting causal relations that hold in all instances of the relevant properties. If so, then the causes of physical effects guaranteed by Completeness are presumably the properties that figure in those laws. Now, the Determination principle states that if a mental property M supervenes upon a physical property P, then P necessitates M. This is to say that, at any possible world, if P instantiates then M instantiates. It is through the implied necessitation from supervenience that instantiations of P are then taken to entail instantiations of M. If the entailed m-properties are to then be assimilated to causal relations at the physical level, the determining physical property P needs to be guaranteed as causally related to another physical property. In other words, the subvening P that – according to the Determination principle – necessitates M had better be one covered by Completeness. Given what we have just said about Completeness and laws, P had better be a physical property individuated by physical laws governing causal relations.

So we have strong grounds for thinking that P needs to be a property individuated by the laws of physics. At the least, I would suggest that the burden lies with those who deny it. But this then causes trouble for the proponent of Determination, because it means that for the Determination principle to be applicable – for the antecedent of the principle to be true – she needs a supervenience thesis that is implausibly local.

We can see this by considering a global formulation of supervenience, as above:

GS: For any possible worlds, w and w*, if w* is a physical duplicate of w then w* is a mental duplicate of w.

Evidently, this formulation of supervenience will not suffice to render the Determination principle applicable. It will not entail that mental properties supervene upon specific physical properties as individuated by the laws of physics, so will not entail that specific physical P necessitates M, which is what the antecedent of the Determination principle requires. For the global formulation is too rough-grained for that. The global formulation is compatible, for instance, with a specific mental property – say, my sensation of toothache – supervening upon a vast configuration of physical properties. If w is the actual world and w* is a physical duplicate world, then (GS) entails that w* is a mental duplicate of w, sharing both its physical nature and structure and its mental nature and structure with w. Given that such duplication would, according to (GS), occur between any two possible physical duplicates, this warrants the inference to the complex physical base of w determining – with necessity – the mental nature and structure of w. But because (GS) only entails duplication of the entire higher-level structure, it is silent as to *which particulars of the physical base are responsible for synchronically bringing about which particulars of the mental superstructure*. It is too rough-grained to support the isolation of specific physical base properties as necessitating specific mental properties.

So a global supervenience thesis will not help to establish the antecedent of the Determination principle. This means that in order to secure the applicability of the Determination principle, the proponent of this model needs a much more local formulation. Indeed, as I argue below, she needs a formulation so local as to be implausible. But plausibility is not the main issue here – I want to show that in committing to a suitably local formulation, the proponent of Determination must also commit to other assumptions. The burden of defending the Determination model grows heavy.

Section 4b: Radical Local Supervenience – Assumptions

If a global supervenience thesis will not suffice to support the Determination principle, what form of supervenience would? A standard local formulation, where we quantify over

individuals across worlds, rather than over worlds, will not be local enough. This would only entail that sets of physical properties relative to individuals necessitate sets of mental properties of those individuals. Again, this would be too rough-grained. For the applicability of the Determination principle, we require a thesis that entails necessitation of mental properties by specific physical properties.

Such a radically local supervenience thesis, entailing determination of mental properties by specific physical properties, is costly. It requires assumptions that incur a significant burden of proof.[37]

One assumption that the proponent of radical local supervenience must make is that content externalism – even what I will call, 'mild' content externalism – is false.

Following seminal arguments from Putnam (1975) and Burge (1979), many are sympathetic to the claim that the content of at least some mental states is determined by factors beyond the intrinsic physical states of the individual. A famous example comes from Burge (1979). Actual Larry truly believes a variety of propositions about arthritis. He believes that he has it; that he has had it for many years; that it can be severely painful, etc. He also mistakenly believes that he has arthritis in his thigh. This is false, because it is a condition that affects only the joints. But it is – according to Burge – also a genuine belief about arthritis. Counterfactual Larry is a physical-functional, experiential and behavioural duplicate of actual Larry. He finds himself in a linguistic community that, unlike that of actual Larry, uses the word 'arthritis' to mean a condition that not only affects the joints but can also affect the thigh. When counterfactual Larry forms the belief that 'I have arthritis in my thigh', he forms a true belief. And, again contrary to the situation with actual Larry, counterfactual Larry's belief is not about arthritis: it is about 'tharthritis', the condition that includes affliction of thighs as well as joints. So on Burge's view, actual Larry has a false belief about arthritis; counterfactual Larry has a true belief about tharthritis. Given that counterfactual Larry is a physical-functional, experiential, and behavioural duplicate of actual Larry, it seems to follow that the different content of their beliefs is down to their different linguistic communities' practices

[37] In the present section, I draw attention to the arguments and views that the proponent of radical local supervenience must deny. In Chapter Four (Section 3c), we will also see that this form of supervenience lacks positive support from psychology or physics, and cannot be coherently established via correlations between psychological and neurophysiological properties.

regarding their respective word, 'arthritis'. If so, then it seems that their propositional attitudes about arthritis – that take propositions about arthritis as their content – also differ in their content due to their differing linguistic communities. If we individuate (some) mental states by their content, then we have here an argument to the effect that the individuation of (some) mental states is not wholly determined by intrinsic physical facts about the individual who instantiates those states. Call such content, 'broad'.

So plausibly, some content of some mental states is broad; hence, some mental states of individuals are not determined only by intrinsic physical properties of their bodies. This is particularly plausible for those states that are typically thought of as propositional attitudes. But this is inconsistent with radical local supervenience because this thesis entails that for any mental state M, M is wholly determined by intrinsic physical states of the individual.

Some respond to arguments for content externalism by positing two kinds of content: narrow and broad (Loar, 1988, 2003). Narrow content is that which is wholly determined by intrinsic states of the individual. On this view, if the arguments for content externalism are successful, they do not show that the relevant mental states are wholly individuated by external factors because those states also include content that is narrowly individuated. This is the dual-content view.

However, we should note that a radical supervenience thesis is challenged not only by full-blown content externalism such that all content of all mental states is externally individuated, or even that all content of some mental states is so individuated; but also by *mild* content externalism whereby some content of some mental states is broad. For if only some content of some mental states is determined by factors outside of the intrinsic physical properties of the individual, we still have a negation of radical local supervenience. That thesis has it that all mental states of the individual are wholly determined by physical properties of that individual. So the thesis is incompatible with the view that some mental states are not wholly determined by those physical properties.

Mild content externalism is compatible with the dual-content view. But radical local supervenience is incompatible with mild content externalism. So those who wish to retain a notion of narrow content for some mental states are not thereby ruled out from opposing radical local supervenience.

I am not here claiming that the non-reductive physicalist who wishes to defend radical local supervenience is straightforwardly defeated by arguments for mild content externalism. They do have options available to them.[38] But they must respond, and their response must contend with the plausibility of the arguments and their conclusions.

Even relatively simple, common-sense considerations can serve to render plausible that notion that some content of some psychological states is broad. These include considering the kinds of folk-psychological explanation that invoke mental causation and which largely motivate our concern with mental causation in the first place. The explanandum is given at the level of description that goes beyond mere physiological events; it includes descriptions of bodily movement that implicitly or explicitly refers to social or environment conditions. For example, when we ask, 'why did he reach for the knife?', or 'why did he frown when I mentioned the thesis?'. Now, if these events, which are partially individuated via environmental or social factors, are to be causally explained by some psychological state, then it is plausible that the causal state will also be individuated, at least partly, by these factors. For how could one hope to *causally* explain, by citing a mental state, an action that involves relations to my environment without the content of that state being similarly related? And if the content of my mental state is so related, then it is plausible that the constitutive conditions for that state will include more than just specific, synchronically related physical properties. I suggest the burden of proof lies firmly with those who would deny it, as the defender of radical local supervenience must.

We should also note that whilst claiming content internalism would be sufficient to block the externalist objection to radical local supervenience, it would not be sufficient for *motivating* that supervenience thesis. That's because content internalism does not entail radical local supervenience.[39] So, as regards the issue of content, blocking mild content externalism is necessary but not sufficient for motivating the supervenience claim.

---

[38] There are objections to both Putnam and Burge. Boghossian (1997) and Segal (2000) object to Putnam's 'twin earth' thought experiment; Crane (1991) and Georgalis (1999) reply to Burge's. More recently, Gertler (2012) argues that the distinction between internal and external properties, upon which the internalist-externalist debate appears to depend, is not sufficiently well-defined. Furthermore, no available conception will perform the task of permitting the kind of division that parties to the debate require.

[39] This will be true for either of the two following formulations of content internalism:
Physical Internalism: The content of S's mental states is wholly determined by physical properties that do not depend for their instantiation upon any properties outside of S's biological boundaries.

But the proponent of radical local supervenience would need to go further than rejecting mild content externalism. She would need to deny *any view* on which the content of any of a subject's mental states depends, even partially, on the causal history of the subject. So she would need to rule out any view on which mental content depends in any way on prior psychological *or* physical states. A radical local supervenience thesis would entail that specific physical properties *synchronically determine* mental states. So specific physical states at time *t* are sufficient to constitutively bring about mental states. If sufficient, then no states prior to time *t* are required to determine the instantiation of the relevant mental states.

Given the plausible grounds for thinking that the content of at least some mental states is at least partly determined by prior states of the subject, this means that advocating radical local supervenience is costly indeed. One can, for instance, read Burge's argument as making plausible the notion that the causal history does make a difference to the content of mental states, since it is not only factors *beyond the biological boundaries* of the subject that allegedly

---

Phenomenal Internalism: The content of S's mental states is wholly determined by phenomenal states of S. (e.g. Farkas, 2008)

Taking physical internalism first, this thesis, if true, would not entail that specific physical properties necessitate specific psychological properties. Put roughly, this form of internalism is still too wide to capture the very narrow constitutive correlations specified by radical local supervenience. Consider, for instance, S's psychological property, instantiated at time t1, of believing that it is now raining. Assume that physical internalism is true, so that the propositional content of this belief is wholly determined by physical properties that do not depend for their instantiation upon any properties outside of S's biological boundaries. On this formulation, internal content is content that is determined wholly by physical properties that instantiate independently of properties outside of S's biological boundaries. Let's assume that such properties are properties within S's body and brain. This will not entail that S's belief that it is raining is necessitated by a specific physical property in the sense intended by radical local supervenience. That's because this form of internalism is compatible with a holistic thesis such that the psychological properties of S are necessitated by conjunctive physical states comprising combinations of specific physical properties. Such holistic states are not the specific physical properties individuated by the laws of physics, and so are not the specific properties designated by radical local supervenience as determinants of specific psychological states. Furthermore, whilst it is possible that such holistic states decompose into the specific properties designated by the laws of physics, internalism says nothing to entail that these subsets of holistic states are sufficient for the propositional contents of the psychological states in question.

Rather different considerations apply for the second formulation of internalism. This formulation is consistent with both property and substance dualism about the actual world. The key notion of phenomenal states is neutral on the question of what those states consist in: they might be physical or non-physical; they might be states of a physical substance or otherwise. This ontological neutrality makes the formulation ill-suited to implying radical local supervenience, because the latter view is inconsistent with both property and substance dualism. (RLS) says that, in the actual world, psychological properties metaphysically depend upon physical properties such that instantiations of the latter necessitate the former. But substance and property dualism claim that (at least some) psychological properties are not necessitated by instantiations of physical properties; there are metaphysically possible worlds at which – as per the zombie argument or Descartes' conceivability argument – physical properties are realised but (some or all) psychological properties are not. If (RLS) is inconsistent with both substance and property dualism, and phenomenal internalism is not, then phenomenal internalism cannot entail (RLS).

determine the semantic content of the subject's words; those factors are also parts of the subject's causal history (in Burge's example, the socio-linguistic history). The proponent of radical local supervenience would also have to claim that Davidson's Swampman (1987) was, despite entirely lacking Davidson's causal history, a mental duplicate of Davidson.[40]

In addition to the above, there may be further assumptions required in defending a radical local formulation of supervenience. It is plausible that some mental states, such as intentions, are standing states that require a range of constitutive psychological background conditions (such as the holding of certain beliefs). Call such states 'wide'.

Now we suppose that radical local supervenience is true, and that specific physical P determines wide M. The psychological background conditions constitutively required for M are, given supervenience, also supervenient upon physical properties. If M is wide, and radical local supervenience holds, then instantiation of P needs to entail instantiation of the subvening physical properties that determine the psychological background conditions.

It seems that if P-instantiation entails instantiation of the background conditions, then P must 'pack in' the background conditions; P must be a conjunctive physical property. But to hold P as a conjunctive property, it seems that one must assume that every conjunct has (confers) distinct causal dispositions because otherwise it's not clear in what respect they are distinct conjuncts. But in order to *be* conjuncts of the *single* conjunctive P, every conjunct would need to confer these distinct causal dispositions under the *same laws*. (P must be covered by some law, since P is covered by Completeness.) This is a significant assumption.

Furthermore, every conjunct of P must necessarily co-instantiate, since every instantiation of P across worlds, by radical local supervenience, determines an instantiation of M. If the conjuncts are distinct physical properties, then the clearest reason for why they would necessarily co-instantiate would be that they are covered by some law that entails their co-

[40] In Davidson's story, he is wandering through a swamp when lighting strikes and reduces his body to its basic physical elements. Simultaneously, the lightning transforms the basic elements in a nearby dead tree into a perfect physical replica of Davidson's body (including the brain). Davidson takes it that the replica (i.e. Swampman) is not him, for Swampman – though behaving exactly as Davidson would – lacks the causal history of Davidson. On his view, Swampman's words and thoughts lack meaning since Swampman lacks the historical connections with the world and others required for attributions of such meaning.

instantiation. But this involves not only the assumption of the requisite law but the assumption that physical laws are metaphysically necessary.

Summary

In the above, we have seen that the Determination model requires a radically local formulation of supervenience. Adoption of that thesis requires:

(a) Denial of mild content externalism
(b) Denial of any form of causal theory of content

It might also involve assumptions about conjunctive physical properties if some mental states are wide. Such assumptions are substantive, controversial and open to rejection by non-reductive physicalists.

Section 5: The Intervention Model

The Intervention model is my third reconstruction of the Exclusion Problem. It attempts to establish diagonal causal relations between mental and physical properties, and hence systematic overdetermination, on the basis of the core Exclusion principles plus auxiliary principles that connect supervenience to a particular conception of causation: Interventionism. The component principles of the model are as follows:

Intervention Model

Distinctness: Mental properties are metaphysically distinct from physical properties.

Supervenience: Mental properties supervene upon physical properties.

Causation: Mental properties have – in their own right qua mental properties – genuine causal efficacy.

Completeness: All physical effects have sufficient physical causes.

Exclusion: No single effect can have more than one sufficient cause occurring at any given time—unless it is a genuine case of causal overdetermination.

Correlation: If m-properties supervene upon p-properties, then m-effects are correlated with p-causes and some p-effects are correlated with m-causes.

Intervention Causation: Variable *c* is a cause of variable *e* iff values of *e* are correlated with values of *c* under interventions upon *c.*

The key principles here are the Correlation and Intervention Causation principles. Correlation plays a similar role in this model to the Dependence Assumption and Determination principles in our previous two models. It seeks to articulate the implication of mental-physical supervenience in terms of correlations between causal relata across levels. This then sets up diagonal causal relations in virtue of the Intervention Causation principle which defines causation in terms of particular kinds of correlation. I further explain these principles below.

The rough idea of the Intervention model is this. Take a horizontal causal relation between two mental properties, e.g. my sensation of intense heat and my deciding to move away from the heat. On an interventionist notion of causation, there must be correlations between values of these two properties. For the sake of simplicity, let us assume that my heat sensation takes a value of 1 if instantiated, and a value of 0 if not. We assume the same for my decision to move away from the source of the heat. If my sensation causes my movement, then there is a correlation between the two properties such that if my sensation has a value of 1, then so too does my decision to move. If my sensation has a value of 0, then so too does my decision. Now, on the Intervention model, my decision to move away from the heat source supervenes upon some physical base, which we'll call D. If so, then the thought is that values (1 or 0) taken by the decision property also correlate with values (1 or 0) taken by the physical base, D. But we already have a correlation between the two mental properties: the sensation property and the decision property. So we have a correlation between the sensation property and the physical D property. Since such correlations are supposedly sufficient for causation, we have a causal relation between my sensation and a physical property D subvening my decision.

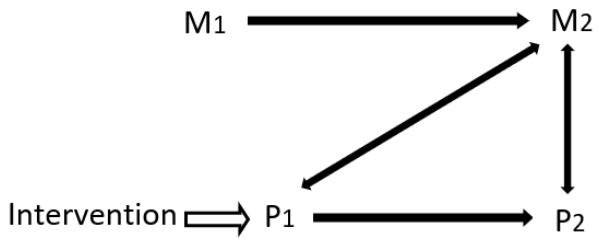
The two routes to overdetermination are as below.

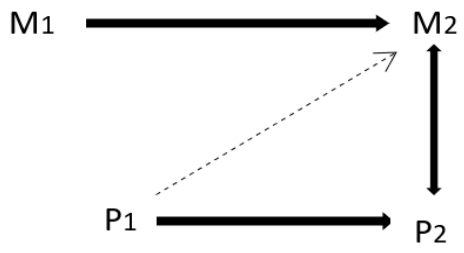
Scenario (1): First Route to Overdetermination

We suppose, as per Causation, that mental property M1 causes M2. We also suppose that, as per Supervenience, M2 supervenes upon some physical P2. It follows that P2 is correlated with M2. By Completeness, P2 has a sufficient physical cause, P1, and by the Intervention Correlation principle, it follows that P1 is correlated with P2. But if P1 is correlated with P2, and P2 correlated with M2, then P1 is correlated with M2, as per the Correlation principle. If so, then by the Intervention Causation principle, P1 is a cause of M2. But M2 already has a cause, M1. So M2 is overdetermined by M1 and P1 **(Figure 6)**.

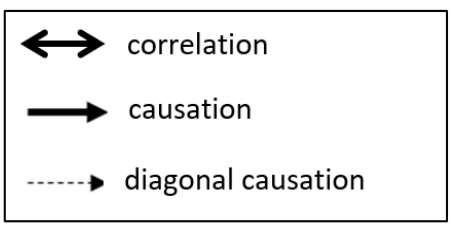
We should also note that this route illustrates the first conjunct of the consequent in the Correlation principle ('…m-effects are correlated with p-causes').

(Fig. 6) Route 1





Intervention

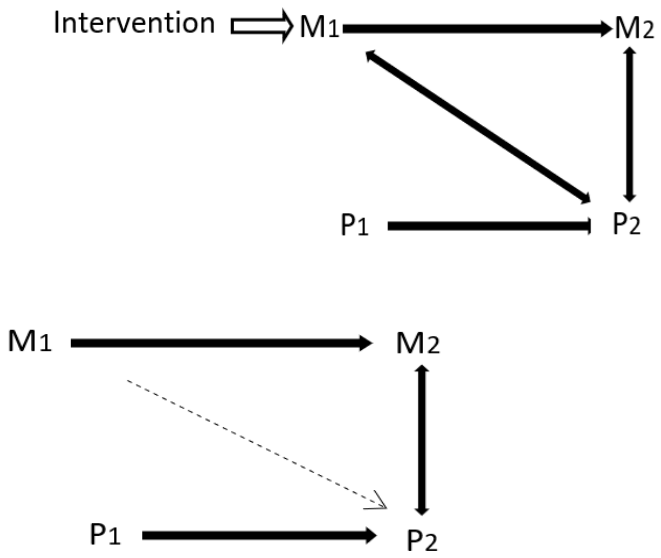


Scenario (2): Second Route to Overdetermination

On the basis of the Causation principle, we suppose that M1 causes M2. The necessary condition for causation stipulated by the Intervention Causation principle implies that values of M1 are correlated with those of M2. We also suppose, from Supervenience, that M2 supervenes upon some physical P2. M2's supervening upon P2 implies a correlation between M2 and P2. But such a vertical correlation itself implies a diagonal correlation between P2 and M1 because P2 is correlated with M2 and M2 is correlated with M1. Given that such correlations are, by the Intervention Causation principle, sufficient for causation, it follows that M1 causes P2. But P2 is physical, so by Completeness has a sufficient physical cause. We appear to have a case of causal overdetermination: P2 is caused by P1 and M1 **(Figure 7)**.

This route illustrates the second conjunct of the consequent in the Correlation principle ('…some p-effects are correlated with m-causes').

(Fig. 7) Route 2



Clarifications

The Intervention model differs from both the Dependence and Determination models in its specifying a necessary and sufficient condition on causation. This reflects the central importance in this model of a specific theory of causation: that of Interventionism. The formulation of the model makes clear the key role of correlations of certain kinds in

supporting diagonal causal relations, and it is this notion of a causal correlation that has a particular meaning in the context of Interventionism.

The basic idea behind an Interventionist conception of causation is that causes render their effects amenable, in principle, to manipulation (Woodward, 2003). If *c* causes *e*, it is possible to influence the occurrence of *e* through manipulation of *c*. One could use the cause as a kind of switch for bringing about the effect. This is possible because causes are statistically correlated with their effects, with the values taken by causes correlated with those taken by their effects. To influence the occurrence of an effect by manipulating a cause is to assign values to the cause such that values are also taken by the effect.

But more needs to be said because causal correlations are not the only kind. The investigator needs to distinguish causal correlations from non-causal. The most obvious form of non-causal correlation would be a correlation based upon common cause of two or more effects. We will speak here of candidate causes and effects as being *variables,* i.e. the bearers of values, because causal relations are understood in terms of correlations between values. Property instantiations or events can be accommodated by characterising the occurrence of these as values of a variable, e.g. the variable 'throwing a stone' takes the value of 1 if the property is instantiated, the value of 0 if not. Suppose that we find a correlation between values taken by variable X and those taken by variable Y. We might be tempted to conclude that X is causally related to Y. But the correlation between X and Y might be due their both being causally related to some other variable, Z. Whenever variable X takes a value of 1, we find that Y does, too. But this might be because both are effects of a common cause, Z. To illustrate, consider the case of lung cancer and yellow fingers.[41] The naïve investigator observes a statistical correlation between values taken by the lung cancer variable and the yellow fingers variable. He then mistakenly concludes that lung cancer causes yellow fingers. But of course, he has missed something: the two variables are correlated because both are effects of a variable that has been missed – smoking. This kind of correlation, between effects of a common cause, is not the only kind of misleading non-causal correlation, but it illustrates the present point: causal correlations need to be carefully identified and defined.[42]

[41] This example is taken from Woodward (2015).
[42] I discuss other kinds of non-causal correlation in Chapter 4.

To that end, Interventionism defines causal correlations in terms of interventions. An interventionist correlation is one which successfully isolates the difference-making capacity of a variable, with respect to an effect, from the influence of other variables. This is achieved by the use of an intervention variable that effectively switches on (assigns values to) a candidate causal variable and switches off the influence of any other variables upon the target outcome (candidate effect). In our smoking case, an intervention would assign a value of 1 to the cancer variable and simultaneously switch off the influence of the smoking variable, thereby isolating the difference-making capacity of lung cancer for yellow fingers. In this case, the capacity is negative – the investigator will find that values taken by the lung cancer variable under interventions are not correlated with those of the yellow fingers variable. The influence of the common cause, smoking, has been switched off to enable elucidation of the direct causal relationship between lung cancer and yellow fingers. If the smoking variable were intervened upon, then the investigator would observe a correlation between that variable and the yellow fingers variable.

The notion of an interventionist correlation is central to the Intervention model, as can be seen by the Intervention Causation principle: Variable *c* is a cause of variable *e* iff values of *e* are correlated with values of *c* under interventions upon *c.* The necessary and sufficient condition for causation is here expressed in terms of an interventionist correlation between *c* and *e*.

We should also note that the two routes to overdetermination under this model both operate via vertical correlations between effects. That is, neither of the routes turn upon a correlation between stipulated causes, P1 and M1. One might think, for all we have yet said, that we should expect vertical correlations between P1 and M1, in virtue of M1's (allegedly) supervening upon P1. But no such routes are permissible within an Interventionist framework.

These routes are ruled out by the constraints on what can qualify as an intervention variable for any causal candidate. We have seen the importance of interventions – to articulate the difference-making potential of a causal variable with respect to an effect, or outcome variable. But interventions are themselves effected by a causal relation between an intervention variable and the causal candidate. This causal relation is taken to, so to speak, stand outside of the variable space containing the candidate cause and the outcome variable

of interest.[43] The key point here is that one of the rules concerning what counts as a legitimate intervention variable is the following: Any directed path from I (i.e. the candidate intervention variable) 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 − X − Y connection itself (Woodward, 2003, p.98). Here, I is candidate intervention variable for causal candidate X and outcome variable Y. For present purposes, this constraint means that for I to be a legitimate intervention variable for articulating the causal relation between, in our model, P1 and P2, I cannot be correlated – independently of its intervening upon P1 – with variable P2. I cannot have a causal path to P2 that does not pass through P1. To illustrate what this means, suppose that I is the candidate intervention variable for the causal relation between P1 and P2. Given Distinctness, P1 is a distinct variable from supervening M1. But M1, in virtue of its (ostensible) supervening upon P1, will be correlated with P1. Values assigned to P1 by I will be correlated with values taken by M1. But this then means that I will either have a causal path to P2 independently of P1 because when I assigns a value to P1 it will also assign a value to M1 which will also be correlated with the value of P2.[44] So I effectively causes M1 to take a value which is then correlated with that taken by P2 – independently (because M1 is distinct from P1) of the causal path from I to P2 via P1. This would mean that I cannot qualify as a legitimate intervention variable for the causal relation between P1 and P2.

The same problem arises in respect of candidate intervention variables for articulation of the causal relation between M1 and M2.

Another note about the model concerns the formulation of the Correlation principle. The principle says that if m-properties supervene upon p-properties then m-effects will be

[43] The intervention variable is, for this reason, an 'exogenous' variable, with the candidate cause and outcome variables being 'endogenous'. We need not elaborate for present purposes, but the exogenous/endogenous distinction and its logic will be of significant importance in Chapter 5.

[44] This issue can be read in either of two ways. One way is that presented here, where the intervention variable I does have a causal path to P2 via M1. The other way is that I *appears* to have a causal path to P2 via M1 but does not actually have such a path. On the first way, intervention variable I fails to satisfy the constraint; on the second way, the interventionist is unable to ascertain whether I satisfies the constraint. We hold the second way open as a reading of the situation because some may wish to claim that M1 is not distinct from P1 in a way that would imply an independent causal path from I to P2. Either, it seems to me, would be problematic and provide reason to rule out routes to overdetermination that depend upon correlations between causal variables. I say more about this in Chapter 4.

correlated with p-causes, and some p-effects will be correlated with m-causes. By 'm-effects', we mean instantiations of mental properties that are taken as effects of mental instantiations. By 'm-causes', we mean instantiations of mental properties that are taken as causes of mental effects that are correlated with the relevant p-effects. It is also worth noting that the consequent in the principle is taken to follow from the antecedent on the assumption of: (a) the Causation principle, i.e. assuming that some mental properties qua mental properties are causally efficacious, and (b) the Completeness principle, i.e. assuming that all physical effects have sufficient physical causes.

The final point of clarification echoes points about previous models. Again, the routes to overdetermination mooted by the Intervention model are not intended as mutually exclusive. There is nothing internal to the model to prevent both from obtaining.[45]

Why Intervention?

Some of the appeal of the Intervention model derives from similar considerations to that of our two preceding models. It clearly assigns a central role to the mental-physical supervenience relation, via the Correlation principle. It also does justice to the common intuition that part of the problem with Exclusion cases is that they involve competing causal relations. The notion of a diagonal causal relation is here similar to that notion as it figures in our other models. There is no reason provided by the model to think that a diagonal causal relation here is somehow benign, given that it obtains in virtue of the same feature – i.e. correlations under intervention – as horizontal causation.

That said, why propose a model that commits us to a particular theory of causation, as the Intervention model does? One reason is that Interventionism enjoys relatively widespread support amongst philosophers concerned with mental causation.[46] But furthermore, Interventionism also provides a way of conceiving of causation that might be thought of

[45] It need not detain us here, but this is not a straightforward claim. Campbell (2020) argues that variable sets, relative to which causal relations between variables are articulated, cannot include specific variables related by supervenience. If so, then this might rule out the diagonal causal relation between M1 and P2 from obtaining along with (i.e. relative to the same variable set as) the diagonal relation between P1 and M2, because M1 supervenes upon P1. I discuss, and reject, this argument in Chapter 4.

[46] See e.g. Shapiro and Sober (2007); Shapiro (2010); Woodward (2015); Baumgartner (2009); Raatikainen (2010).

particular relevance for and application to psychology (Campbell, 2006). And as we will see in later chapters, there are reasons to think that, for all its potential viability with respect to special science causation, it is nonetheless limited in its scope. Specifically, there are reasons to think that Interventionism might not be well-suited to explication of causal laws in physics. So we want to introduce an Interventionist perspective here, to start clearing the ground for further consideration of how it might impact upon broad pictures that inform and motivate Exclusion worries.

Section 6: Latent Assumptions of the Intervention Model

Correlation: If m-properties supervene upon p-properties, then m-effects are correlated with p-causes and some p-effects are correlated with m-causes

As we saw, the Correlation principle is crucial in the Intervention model, for it links supervenience to diagonal correlations, the latter then taken as sufficient for causal relations as per the Intervention Causation principle. In this respect, it functions similarly to the Dependence and Determination principles. Because of this, the Intervention model involves similar assumptions to both of our preceding Exclusion models. We should therefore be able to see the implicit assumptions relatively swiftly. The assumptions divide into two groups: one group pertinent to the first route to overdetermination, the other pertinent to the second route.

Assumptions for First Route: M-effects are Correlated with P-causes

The first route to overdetermination turned upon a diagonal correlation between P1 and M2. This diagonal relation itself was supposed to result from a vertical correlation between P2 and M2, which we can call a 'bottom-up' correlation because it is supposedly implied by the causal relation between the properties at the bottom level: P1 and P2. Bottom-up correlations are crucial in supporting the first conjunct of the consequent in the Correlation principle: '…m-effects are correlated with p-causes'. The idea was that if M2 supervenes upon P2, then M2 is vertically correlated with P2; and if P2 is caused by P1, then P2 is horizontally

correlated with P1; so M2 is diagonally correlated with P1 via its correlation with P2. An m-effect is correlated with a p-cause, as per the Correlation principle.

The assumptions required for these bottom-up correlations are the same as those required for the Determination model. There, the Determination principle was only applicable on the assumption that supervenience is a radically local thesis. But assuming that thesis itself requires a host of assumptions against: (a) mild content externalism and (b) any form of causal theory of content. Indeed, we further argued that assuming a radically local supervenience formulation may plausibly involve assumptions concerning (c) conjunctive physical properties that may be required on the basis of some mental states being 'wide'.[47]

In my discussion of the Determination model, I said that the key principle, Determination, was only applicable on the basis of these further assumptions. What was meant was that the principle is only (non-vacuously) true if it has a true antecedent, and it only gets to have a true antecedent if supervenience is radically local. Due to the formulation of Correlation, I put things slightly differently here. We should say that *the first conjunct of the consequent in the Correlation principle is only true if radical local supervenience is true.* On that basis, the range of further assumptions outlined in our discussion of Determination then become relevant for the truth of the Correlation principle, too.

The first conjunct of the consequent in Correlation is this: '…m-effects are correlated with p-causes'. For reasons familiar from my section on the Determination model, the p-causes and p-effects involved in putatively generating diagonal correlations need to be specific properties of physics. For if they are not, then Completeness will not guarantee that the p-effect has a p-cause, and will therefore not guarantee the p-cause (P1) of the diagonal correlation to M2. This first conjunct expresses the obtaining of diagonal correlations, and as we said above, these correlations are supposed to be implied by vertical correlations between p-effects and m-effects. In our present context, this would be a bottom-up vertical correlation between P2 and M2.

This bottom-up vertical correlation is supposed to follow from supervenience of M2 upon P2. Bottom-up correlations are taken to follow from the necessary determination aspect of supervenience. The thought is that if P2 subvenes M2, then P2 necessitates M2; and if P2

[47] See Section 4 of this chapter.

necessitates M2, then any case in which P2 takes a value of 1 (i.e. any case in which P2 instantiates), is also a case in which M2 takes that value. Hence, we would have a correlation between P2 and M2. Now, I do not wish to question the claim that P2's necessitation of M2 would imply a correlation between their values. But as we have seen, for P2 to play the role required in generating systematic overdetermination, it must be a specific physical property. If so, then the supervenience relation referred to by the Correlation principle must be radically local. As we saw in Section 4, that is a significant assumption, and arguably requires a range of further significant assumptions.

Assumption for Second Route: Some P-effects are Correlated with M-causes

In the case of the Dependence model, my main argument was that the core principle – the Dependence Assumption[48] - is only true on the assumption that mental properties are not multiply realisable across close counterfactual worlds. This is also an assumption that the proponent of the Intervention model must make. Again, this assumption is required *for the truth* of the key principle connecting supervenience to causation: the Correlation principle.

Correlation: If m-properties supervene upon p-properties, then m-effects are correlated with p-causes and some p-effects are correlated with m-causes.

More specifically, the assumption is required for it to be true that if m-properties supervene upon p-properties, then some p-effects are correlated with m-causes.[49]

Suppose, with the Intervention model, that M2 supervenes upon some physical base, P2. A variable taking a value of 1 here entails that the corresponding property is instantiated, in which case, M1's causing M2 is a matter of M1's bringing about M2's instantiation. Similarly, a property that does not instantiate entails that the corresponding variable takes a value of 0. To say that M2's supervening upon P2 implies a correlation in values of M2 and P2 is to say that M2 takes a value of 1 when P2 does, and M2 takes a value of 0 when P2 does. And M2

[48] Dependence Assumption: If M supervenes upon P, then M is counterfactually dependent upon P.

[49] The consequent of the Correlation principle is a conjunction. Different routes to overdetermination deploy different conjuncts. The first assumption required for the principle concerns one route, and so one conjunct. The later assumptions will concern the other route, and so the other conjunct (i.e. 'm-effects are correlated with p-causes').

and P2 take a value of 1, respectively, when the corresponding properties instantiate; otherwise they take a value of 0.

Now, the Correlation principle (specifically, the second conjunct of the consequent) requires that P2 instantiates when M1 does. We should recall the reason for thinking that some p-effects (i.e. P2) are correlated with m-causes (i.e. M1) if m-properties supervene upon p-properties. The thought was that *if M2 supervenes upon some p-property (P2) then it will be correlated with that property*. This is what we can call a 'top-down' correlation, since it supposedly follows from the causal relation between properties at the top, mental, level. This is taken to imply that M1, as cause of M2, will also be correlated with P2 via the latter's correlations with M2; hence, the p-effect is correlated with the m-cause. So correlations between m-effects (M2) and p-effects (P2) are crucial in supporting diagonal correlations from m-causes (M1) to p-effects (P2).

But if M2 is closely multiply realisable, then there are close cases in which M2 instantiates and P2 does not. This is a case in which M2 takes a value of 1 and P2 takes a value of 0. Given that M2's being correlated with P2 is a matter of their both taking the same value in the same cases, close multiple realisability of M2 is sufficient to block top-down correlations, and hence sufficient to falsify the Correlation principle. The truth of that principle requires an assumption against close multiple realisability of m-properties.

Conclusion

In this chapter, I have shown that three reconstructions of the Exclusion Problem fail to imply diagonal causal relations, and hence fail to imply systematic overdetermination. Each model requires further assumptions in support of the principles by which supervenience is taken as pivotal in generating diagonal causation.

The Dependence model requires assumptions against close multiple realisability of mental properties. The Determination model requires the assumption of radical local supervenience, which itself requires assumptions against mild content externalism and any form of causal theory of content. The Intervention model requires some of the same assumptions, depending upon which overdetermination route one considers.

The upshot is that none of these models can be deployed by the non-reductive physicalist without adopting substantive, controversial positions on multiple realisability or supervenience and mental content. Without assuming those positions, the models fail to imply systematic overdetermination.

We should also note that this result, whilst achieved by way of assessing the above models, is plausibly not limited to those models. For the difficulties do not arise from features peculiar to them, such that one might reasonable expect some other model to avoid the issues. For each model, the requisite assumptions are needed to facilitate the move from supervenience (in conjunction with Causation) to diagonal causal relations between specific mental and physical properties. As such, we can reasonably expect that those assumptions will be needed any model that leans on supervenience in this way.

On that basis, the non-reductive physicalist need not feel pressed to formulate solutions to the Exclusion Problem. Without the assumptions outlined above, the non-reductive physicalist need not fret about physics – via Completeness – excluding mental causes. Hence, they have available a dissolution of the Exclusion Problem.

Chapter Three - Direct Diagonal Causation

Introduction

The broad message of my argument in Chapter 2 was that non-reductive physicalists sympathetic to the Exclusion Problem are free to both endorse the principles included in any of the models *and* to reject systematic overdetermination. I wanted to show that non-reductive physicalists need not be bound by the Exclusion Problem in their views on mental causation. What I now want to show is that those same physicalists need not - and should not -oversimplify and reject *all* diagonal causation.

In this chapter, we take inspiration from a common-sense case of diagonal mental causation, i.e. a case in which a mental cause has a physical effect. Our starting point is the following case:

Hadron Collider (HC)

> I form the intention to flick the switch on the Hadron Collider (mental event). This causes me to flick the switch, and my flicking the switch causes the Collider to start smashing protons (physical event).

My central claim in this chapter is that Hadron Collider can be read as a case of diagonal causation consistently with rejecting the claim that Dependence, Determination or Intervention entail systematic overdetermination. Furthermore, the Hadron case can be read in this way consistently with endorsing the core Exclusion principles[50]. The latter is important, because the message from the previous chapter was intended as a call for optimism on the part of a certain kind of physicalist. For non-reductive physicalists sympathetic to the principles that constitute the Exclusion Problem, my claim that those principles do not entail systematic overdetermination should be good news. I do not wish to spoil that news in this chapter by proposing mental causation that requires the rejection of those principles.

[50] These are: Distinctness, Supervenience, Causation, Completeness, and Exclusion.

So in what follows, I want to urge that common-sense attributions of diagonal mental causation are heeded, and that nothing in those cases requires the rejection of my earlier claims, arguments or assumed principles.

In Section 1, I clear the ground for diagonal mental causation by identifying and articulating the common element running through the Dependence, Determination and Intervention models of Exclusion: alleged *supervenience-based* causation. In each of these readings of the Exclusion Problem, diagonal causation - and hence systematic overdetermination – was supposedly entailed by the relevant principles because of supervenience. It was the supervenience that was, on each model, key to systematically generating diagonal causal relations. So given my stated ambitions, any viable model of diagonal mental causation must be one on which such diagonal relations do not obtain in this way.

In Section 2a, I articulate the Hadron case with a model that I call, 'Direct Causation'. The idea is that the new model of diagonal causation at work in Hadron Collider is sufficient to *accommodate* intuitions about the case whilst remaining consistent with the core Exclusion commitments. The principles that compose the model are those entailed by my reading of Hadron as a case of diagonal causation that is not *supervenience-based causation* plus the core Exclusion principles.

In Section 2b, I marshal arguments in support of Direct Causation as a viable model of the Hadron case. These arguments focus on showing the plausibility of the mental cause in Hadron Collider being efficacious not in virtue of its instantiating supervenience-based causation.

Section 3 addresses the issue of consistency, both between the model and my earlier claims against entailment of systematic overdetermination, and between the model and the wider dialectical point of the previous chapter. That point was that the non-reductive physicalist is free to endorse the Exclusion propositions *and* deny systematic overdetermination. Here, I show that Direct Causation is a model that is consistent with my earlier arguments against the Dependence, Determination and Intervention models of Exclusion. I also show that Direct Causation is not a model that itself entails systematic overdetermination.

The final section considers the coherence of the Direct Causation model. I argue the model is coherent, with particular focus on the logical relationship between the Exclusion principle itself[51] and the other constituent principles.

Section 1: Supervenience-based Causation

Each of the Exclusion models previously examined are cases of ostensible *supervenience-based causation*. That is, the Dependence, Determination and Intervention models all supposedly entail diagonal causal relations between the mental and physical domains via relations entailed, in turn, by supervenience. Let us recap the main principles of each model to see how this plays out. We start with the Dependence model:

Dependence Model

Dependence Assumption (DA): If M supervenes upon P, then M is counterfactually dependent upon P.

Dependence Chains: A dependence chain is any finite sequence of events a,...n such that b is counterfactually dependent upon a, c is counterfactually dependent upon b etc.

Causal Dependence: *c* causes *e* if there exists a dependence chain leading from *c* to *e.*

Supervenience: Mental properties supervene upon physical properties.

Distinctness: Mental properties are metaphysically distinct from physical properties.

Causation: Mental properties – qua mental properties – are causally efficacious.

Completeness: All physical effects have sufficient physical causes.

We can illustrate the key notion of supervenience-based causation by considering one of the routes to diagonal causation, implied by the Dependence model.

[51] Exclusion: No single effect can have more than one sufficient cause occurring at any given time—unless it is a genuine case of causal overdetermination

On supposition of the Causation principle, mental property instantiation M1 causes M2. In accordance with Supervenience, M2 supervenes upon some physical property, P2. As per the Completeness principle, P2 has a sufficient physical cause, P1. Now, with the Dependence model, we assume a Lewisian notion of causation. So we assume that if P2 is caused by P1, then P2 counterfactually depends upon P1. But the Dependence Assumption principle tells us that M2 also counterfactually depends upon P2, because m-properties supervene upon p-properties. So given the definition of a *dependence chain,* there is one such chain running from M2 through P2 to P1. M2 counterfactually depends upon P2; P2 counterfactually depends upon P1. Hence, given our causal dependence principle, there is a causal relation between P1 and M2: P1 causes M2. But M2, by hypothesis, already has a cause in M1. So M2 is overdetermined by M1 and P1.

The key move in securing the diagonal causal relation between P1 and M2 is made by applying the Dependence Assumption: if M supervenes upon P, then M counterfactually depends upon P. This implies, as per the above, that M2 counterfactually depends upon P2. As such, we can see how the supervenience relation between mental and physical properties is supposed to support, via the generation of dependence chains, the diagonal causation of M2 by P1. Here we have an example of (alleged) *supervenience-based causation.*

Determination Model

Let's now see how supervenience-based causation works in the context of the Determination model. These were the key principles:

Supervenience: Mental properties supervene upon physical properties.

Distinctness: Mental properties are metaphysically distinct from physical properties.

Causation: Mental properties have – in their own right qua mental – genuine causal efficacy.

Completeness: All physical effects have sufficient physical causes.

Determination: If M supervenes upon P, then P necessarily determines M.

Assimilation: If M is necessitated by P, M is a relatum (of the same type, i.e. cause or effect) in any causal relation where P is a relatum.

To illustrate the role of supervenience here, consider the following alleged route to diagonal causation. Suppose that some physical property, P2, instantiates. By Completeness, P2 has a sufficient physical cause, P1. Suppose also that P1 is the supervenience base for a mental property instantiation, M1. The Determination principle tells us that, because M1 supervenes upon P1, P1 necessitates M1. This then sets up M1 for being drawn into the causal relation between P1 and P2 by applying the Assimilation principle. If M1 is necessitated by P1 (which, given M1's supervening upon P1 and the Determination principle, it is), then M1 is assimilated by the causal relation between P1 and P2. Hence, M1 is also a cause of P2. But P2 – by Completeness – already has a sufficient physical cause in P1, so P2 is overdetermined by P1 and M1.

As before, we can see that supervenience plays an essential role in generating the ostensible diagonal causal relation between M1 and P2. Without supervenience, we would have no reason for asserting the necessitation of M1 by P1, and so no reason for thinking that M1 is assimilated by the causal relation between P1 and P2. Supervenience is the basis for (alleged) diagonal causation in the Determination model. Once again, if it is a case of diagonal causation, then it is supervenience-based causation.

Intervention Model

Finally, we consider the Intervention model:

Distinctness: Mental properties are metaphysically distinct from physical properties.

Supervenience: Mental properties supervene upon physical properties.

Causation: Mental properties have – in their own right qua mental properties – genuine causal efficacy.

Completeness: All physical effects have sufficient physical causes.

Correlation: If m-properties supervene upon p-properties, then m-effects are correlated with p-causes and some p-effects are correlated with m-causes.

Intervention Causation: Variable *c* is a cause of variable *e* iff values of *e* are correlated with values of *c* under interventions upon *c.*

Again, let's see how supervenience plays a crucial role in generating diagonal causation on this model. One potential route to such causation is as follows.

Assume that, as per Causation, psychological variable M_1 causes psychological variable M_2. Assume, as per Supervenience, that M_2 supervenes upon physical P_2. Assume, as per Completeness, that P_2 has a sufficient physical cause, P_1. Given Correlation, P_2 is correlated with M_1: values assigned to M_1 are correlated with those taken by M_2, and M_2 supervenes upon P_2, so those interventions will also be correlated with values taken by P_2. By Intervention Causation, P_2 is hence caused by M_1. But P_2 has a sufficient physical cause in P_1. Hence, P_2 is overdetermined by M_1 and P_1.

Here, supervenience is vital to establishing the correlation between M1 and P2. It is the basis for affirming that correlation, which is in turn the basis for taking M1 to cause P2. So again, we have supervenience as the basis for the alleged diagonal causal relation.

Having now illustrated the supervenience-based causation allegedly entailed by the previous models of the Exclusion Problem, we are in a position to see how our model of mental causation in the Hadron case differs from each of them. I now introduce that model.

Section 2a: The Direct Causation Model of Diagonal Causation

The Direct Causation interpretation of Hadron Collider consists of the following principles.

Direct Causation

Causation: Mental properties have – in their own right qua mental properties – genuine causal efficacy.

Distinctness: Mental properties are metaphysically distinct from physical properties.

Direct Causation: For some physical *p* and some mental *m*, *p* is directly caused by *m.*

Directness[52]: A causal relation *R* between *c* and *e* is direct iff *R* does not obtain in virtue of a causal relation that obtains between properties *p*, upon which *c* supervenes, and *e.*

Supervenience: Mental properties supervene upon physical properties.

Completeness: All physical effects have sufficient physical causes.

Exclusion: No single effect can have more than one sufficient cause occurring at any given time—unless it is a genuine case of causal overdetermination.

Direct Causation & Hadron Collider

Intuitively, Hadron Collider differs from the scenarios ostensibly implied by Dependence, Determination or Intervention, as those turned upon some specific higher-level properties supervening upon some physical properties. Those supervenience relations were the mechanism by which causal relations with a higher-level relatum were supposedly brought into competition with causal relations obtaining – by Completeness – between physical events. By contrast, it is tempting to say that, in the Collider case, a mental property instantiation bears a *direct* causal relation to a physical instantiation. It looks to involve *Direct Causation* – the macro-level instantiation (flicking the switch) causes the micro-level instantiation (collider initiation) directly, i.e. not in virtue of a causal relation that obtains between some micro-level instantiation, upon which my switch-flicking supervenes, and the collider's initiating. When presented with a common-sense description of the case, it might seem natural to simply say that if I had not flicked the switch, then the subatomic activity constitutive of the initial collision process would not have occurred. Certainly, this is how it seems to me.

If so, then we can differentiate between a case like Hadron and the kinds of case associated with the Dependence, Determination or Intervention models. In those cases, causal relations between levels required a horizontal causal relation – i.e. a relation between two physical properties or between two mental properties – in conjunction with some synchronic

[52] Here we should also acknowledge that the Direct Causation and Directness principles are both stipulative, initially motivated by the intuitive appeal of HC as read in this way. I take this to be reasonable, given that Direct Causation is supposed to be a means of explicating an intuitive response to the Hadron description. The suggestion here is that reflection upon Hadron Collider would motivate Direct Causation.

supervenience relation by which physical causes were implicated in the bringing about of mental effects and mental causes implicated in the bringing about of physical effects. In other words, the putative causal relations were all forms of supervenience-based causation. But on the Direct Causation interpretation of Hadron Collider, we have the following picture. My intention to flick the switch causes me to flick the switch, and my flicking the switch causes the subatomic activity of collider initiation. My intention and my flicking the switch supervene, respectively, upon some physical base. Furthermore, the causal relation between my flicking the switch and the collider's activity does not depend upon causal relations that obtain, respectively, between those supervenience bases and the collider's initial activity. On the Direct Causation model of Hadron Collider, we have a direct causal relation between the subatomic effect and the flicking of the switch: it follows by transitivity of causation that the subatomic effect is caused by my intention to flick the switch. On the Direct Causation reading of HC, the causal relation between my intention and the collider's smashing of protons is not parasitic upon specific supervenience relations between the relevant higher- and lower-level properties.

Direct Causation, Completeness, and Supervenience

The classic Exclusion principles most obviously required for the Direct Causation view of Hadron Collider are the following: Causation and Distinctness. Causation is implicated in my intention causing my flicking of the switch; Distinctness implies that the mental cause, and my flicking the switch, are metaphysically distinct from any underlying physical properties – which is required for any bona fide diagonal causal relation.

We include Supervenience and Completeness here for two reasons. First, we do so in order to forestall misconceiving our view as a much more radical claim than intended. The more radical claim would be that diagonal causation could obtain in worlds where Completeness and Supervenience were, as general principles, false. I am not making that claim. I only want to say that the diagonal causal relation in Hadron does not depend upon any *specific* causal relation between the relevant supervenience base (of my switch-flick) and the physical effect (the collider's initiation).

My second reason for including Supervenience and Completeness, and my reason for including the Exclusion principle, in my model is to make explicit the availability of the model to the non-reductive physicalist who wishes to endorse the principles that constitute the Exclusion conjunction: Distinctness, Supervenience, Causation, Completeness, and Exclusion itself.

Causation in Direct Causation

We should note that the Direct Causation model does not include any particular specification of causation. It includes no necessary or sufficient conditions for a causal relation to obtain (the Directness principle pertains only to the directness of causal relations; not to their being causal). This is deliberate, because I want my model to be amenable to a non-reductive physicalist who accepts the core Exclusion propositions and who wishes to endorse the auxiliary principles of any of the Dependence, Determination of Intervention models. So I have not stipulated any particular notion of causation in our model of direct diagonal causation in Hadron.

For the same reason, I have specified Directness in a way that is compatible with a Lewisian or Interventionist conception of causation, or indeed others. As we saw previously, a Lewisian conception is implicit in the Dependence model and obviously an Interventionist one is deployed in the Intervention model. The Determination model does not specify. This is part of the advantage of a negative formulation. A causal relation's obtaining *not* in virtue of any causal relation between the relevant supervenience base *p* and the effect is open as to what the causation *does* consist in.

Direct Causation and the Dependence and Intervention Models

Although the Dependence and Intervention models are both, in their own way, forms of (alleged) supervenience-based causation, the particular routes to diagonal causation suggested by these models are not pertinent to the Hadron case as described. For that reason, they are also not pertinent to the Direct Causation model, since that model only includes

principles required to accommodate the direct mental cause of the collider's starting as outlined by the description.

Let's consider the Dependence model first. These are the principles:

Dependence Assumption (DA): If M supervenes upon P, then M is counterfactually dependent upon P.

Dependence Chains: A dependence chain is any finite sequence of events a,...n such that b is counterfactually dependent upon a, c is counterfactually dependent upon b etc.

Causal Dependence: *c* causes *e* if there exists a dependence chain leading from *c* to *e.*

Supervenience: Mental properties supervene upon physical properties.

Distinctness: Mental properties are metaphysically distinct from physical properties.

Causation: Mental properties – qua mental properties – are causally efficacious.

Completeness: All physical effects have sufficient physical causes.

According to this model, we get diagonal causal relations in the following ways:

Route 1: M1 causes M2, so M2 counterfactually depends upon M1. M1 supervenes upon P1, so M1 counterfactually depends upon P1. We therefore have a dependence chain from M2 to P1 via M1. Causal dependence, sufficient for causation, is transitive. So the chain from M2 to P1 via M1 implies causal dependence of M2 upon P1.

Route 2: M2 supervenes upon P2, so counterfactually depends upon P2. P2 is caused by P1, so counterfactually depends upon P1. There is a dependence chain running from M2 to P1 via P2. Again, causal dependence is transitive so the chain implies causal dependence of M2 upon P1.

Both routes deliver the same causal relation: P1's causing of M2. But in the Hadron case, we are interested in a putative causal relation between M1 and P2. So Dependence is not pertinent to the causal relation at issue in Hadron.

Now we turn again to the Intervention model. Let's briefly recap the Intervention route to causation of P2 by M1. The Intervention model comprises:

Distinctness: Mental properties are metaphysically distinct from physical properties.

Supervenience: Mental properties supervene upon physical properties.

Causation: Mental properties have – in their own right qua mental properties – genuine causal efficacy.

Completeness: All physical effects have sufficient physical causes.

Correlation: If m-properties supervene upon p-properties, then m-effects are correlated with p-causes and some p-effects are correlated with m-causes.

Intervention Causation: Variable *c* is a cause of variable *e* iff values of *e* are correlated with values of *c* under interventions upon *c.*

The route by which, according to this model, P2 would be caused by M1 is this. M1 causes M2 and M2 supervenes upon P2. Given that M1 causes M2, values assigned to M1 are correlated with M2. Given M2's supervening upon P2, and given the Correlation principle above, there is a correlation between values assigned to M2 and those assigned to P2. On that basis, and given the Intervention Causation principle, M1 is a cause of P2.

On the Intervention model of supervenience-based causation, higher-level properties such as M1 get to cause physical property instantiations such as P2 in virtue of the physical effect's subvening a higher-level effect. We might say that M1 causes P2 in virtue of M1's causing M2 and the (ostensibly) implied correlations with P2.

How does the Intervention model relate to Hadron Collider? Again, like the Dependence model, the routes to ostensible diagonal causation are not pertinent to Hadron. That's because neither the initial common-sense description of Hadron Collider nor our formulation of the Direct Causation model include any mention of M1's causing some higher-level M2. Nor do they include any mention of P2 subvening some higher-level M2. If we are inclined to judge, in response to the initial description of the case, that my intentional flicking of the switch causes the collider to start, then presumably we are not tempted to do so in the basis of the route set out by Intervention. The description says nothing to elicit the latter thought. The same holds for the Direct Causation model. So in supporting the directness of the mental

cause in Hadron Collider, we need not show that it obtains not in virtue of the route articulated by Intervention. There is no suggestion of it doing so given the description of the case or formulation of the model for the case.

Section 2b: Arguments for Direct Causation in Hadron Collider

We can offer further support for the Direct Causation principle and its application to the Hadron case. To do so, I will first offer arguments in support of my HC description such that my intentional flicking of the switch is a bona fide cause of the collider's activity. In other words, I argue for the claim that a *mental* property instantiation is efficacious with respect to the collider's activity. From there, I will argue for the directness of the causal relation, as expressed by Direct Causation together with Directness.

A natural question to ask at this point is this: given that we earlier argued that the Dependence, Determination and Intervention models failed to entail diagonal causation, and would only imply such causation when appended with costly assumptions, why should we now argue that Hadron is not a case of supervenience-based causation? Why should we want to argue further for the directness of the mental cause?

My arguments in Chapter 2 showed that the non-reductive physicalist need not accept diagonal causation on the basis of the Exclusion principles. Those principles do not entail diagonal causation. The overall purpose of this chapter is to caution against taking that to mean there cannot be diagonal causal relations. But the motivation for being careful about that is the plausibility of cases, such as Hadron, where mental diagonal causation is direct. So my arguments below are intended to bolster the intuition that the mental cause in Hadron *is* a direct diagonal cause of the physical effect.

The Mental Cause in Hadron Collider

I take there to be some preliminary intuitive appeal to the notion that the relevant macro-level cause is an event or property instantiation that is (partially) individuated by a relation to agential intention. Roughly, upon consideration of the Hadron Collider scenario, we want to take seriously the potential causal relevance of an action that is motivated by and directed

towards bringing about a represented end. The agent flicks the switch purposefully to bring about the physical effect constitutive of the Collider's initial state, as represented abstractly by the agent. The agent forms an intention that includes this representation of the desired states of affairs (the collider's starting) and acts accordingly (flicks the switch). This chimes with our intuitive take on the case as described. In Hadron Collider, we have preliminary reason to take the macro-cause as being 'my intentionally flicking the switch in order to initiate the collider's physical activity'.

Here is an argument by way of support for our initial intuitions. I should stress that the purpose of the argument is to not to prove the psychological causation of the collider's initial activity; it will not persuade those who flatly deny psychological causation in the Hadron case. Rather, the purpose is to bolster the intuitions of those sympathetic to such causation by showing that a psychological description of the macro-level cause better accords with the dependence of the micro-effect upon that cause. We will call the cause, under the intentional description just given, 'macro-*c*'. And we will call an alternative instantiation, 'my arm moving flick-wise', 'macro-*c**'. My arm moving flick-wise is movement of the arm in *precisely the same purely physical way* that it would move when I intentionally flick the switch – but where this description is open as to agential intentions or their absence. We assume a Lewisian framework whereby counterfactual dependence of effect *e* upon *c* is sufficient for causation of *e* by *c*, and assume that such counterfactual dependence is appropriately cashed out by the familiar counterfactual conditionals (I specify below).

To start, we grant the assumption that in the Hadron case, there is *some* macro-cause of the collider's activity. We do this because to do otherwise violates basic facts that we already accept about the collider, e.g. that it's proton-smashing will not be initiated unless it is switched on. And furthermore we accept, on the basis of the description in HC, that the macro-cause is denoted by the description 'flicking the switch'. That is, we take the macro-cause to be a property instantiation denoted by that description. We also assume, for the sake of argument, that the macro-cause is efficacious in virtue of the micro-effect counterfactually depending upon it, in line with our assumption that such dependence is sufficient for causation.

But now we have options, because the description might be treated as denoting the partially mental (intentional) macro property mentioned above – i.e. 'intentionally flicking the switch

to initiate the collider's physical activity' – or as denoting a macro property that is *only* (broadly) physical – i.e. 'my arm moving flick-wise'. There need arguably be no intentional component in this latter property, since it is plausible that one might trip as one walked past the switch and spontaneously, or from momentum, move one's arm in the way required to flick the switch.

It therefore looks as if we have a choice as to how to interpret the description given in Hadron. But I think we have reason to prefer the intentional, mental interpretation. The central point here is that a counterfactual required for dependence of the collider's activity upon my intentional flicking of the switch (macro-*c*) is plausibly satisfied by close worlds, but the equivalent counterfactual for the collider's activity depending upon my moving my arm flick-wise (macro-*c**) is not. (Recall that we are assuming that the causal relation between macro-cause and micro-effect obtains in virtue of counterfactual dependence; the corresponding counterfactuals are required only in view of this assumption - not in the sense that we take such dependence to be necessary for causation.)

The requisite counterfactuals in which we are here interested are: [if macro-*c* had not occurred, then micro-*e* would not have occurred] and [if macro-*c** had not occurred, then micro-e would not have occurred]. On the standard Lewisian semantics, the first conditional requires that the closest counterfactual world at which my intentional flicking of the switch does not occur is a world at which the collider fails to start. The second conditional requires that the closest world at which my arm's moving flick-wise fails to occur is also one at which the collider fails to start.

Intuitively, at a close world where the antecedent of the second conditional [if micro-*c** had not occurred] holds, the consequent does not. It is plausible that my arm moves in some other, very similar but strictly distinct, way and that micro-e occurs: the collider starts. (Call this other arm movement macro-*c***.) So the counterfactual [if micro-*c** had not occurred, then micro-e would not have occurred] is false.

If, on the other hand, we treat 'flicking the switch' as denoting an instantiation of a partially mental property, then the requisite counterfactual evaluates as True. For the counterfactual [if macro-c had not occurred, then micro-e would not have occurred] to be non-trivially true, there must be a close world at which both the antecedent and consequent are true. In

contrast to the situation above, where consideration of macro-*c** suggests that there are close worlds where some other instantiation, macro-*c***, causes micro-e and hence falsifies the counterfactual [if macro-*c** had not occurred, then micro-e would not have occurred], there is no plausible close replacement for macro-*c*. I take it that, when we interpret 'flicking the switch' as a mental property instantiation, it is intuitively plausible that close worlds at which that property is not instantiated are worlds at which the collider does not start smashing protons.[53]

Given that we started the argument with the assumption that there is some macro-cause of the collider's starting (i.e. M1) and that, given the description provided by the Hadron case, that cause is 'flicking the switch', we need an interpretation of that designation that supports the requisite counterfactuals. On the basis of the above, we have reason for thinking that partially mental, intentional interpretation better supports those counterfactuals than one on which my arm moving flick-wise is the putative cause. For this reason, I hold that we should take seriously the attribution of mental causation in the Hadron Collider case. We should specify the cause of the collider's initial, functioning physical states as an intentional act individuated by the agent's represented end in performing the action. The case, as initially described, is plausibly a case of mental causation of physical effects. If so, then it is also a case of diagonal causation simply in virtue of the stipulated diagonality of the mental and physical domains. It exhibits *diagonal* mental causation.

At this point, it might be suggested that the problem with the physical description is that the description of the bodily movement is too narrow. Each physical description so far (macro-*c** and macro-*c***) has picked out the putative cause in terms of the precise movement. The first description (designated micro-*c**), as relative to the actual world, describes the cause in terms of the precise movement of the arm that actually occurs. The second description (designated micro-*c***) describes the cause in terms of the precise movement of the arm that occurs at the counterfactual world at which micro-*c** fails to occur.

On this way of thinking, the way to avoid the above problem is to describe the movement in broader terms – say, as 'my arm moving in some way sufficient to flick the switch'. Then

[53] Again, I emphasise that this is not intended as a *proof* of the mental property's efficacy; the argument here is only meant to further clarify and draw out the initial intuition in its favour.

perhaps the description will capture the multitude of similar movements, and hence the counterfactual [if macro-*c** had not occurred, then micro-*e* would not have occurred] will evaluate as True at close worlds. It will not suffer the problem outlined above, where close worlds are worlds at which a very similar, but distinct, movement occurs and hence worlds at which the counterfactual evaluates as False because the consequent is false (i.e. because the collider does initiate its activity despite the specific bodily movement not occurring).

However, to attribute my arm's moving in some way sufficient to flick the switch as the cause of the collider's activity provides a less informative explanation of the collider's starting than citing my intentional flicking of the switch. For the description picks out the arm movement only insofar as it brings about the collider's activity (i.e. by flicking the switch). To the question, 'why did the collider start?', we would on the present suggestion answer, 'because my arm moved in some way sufficient to flick the switch'. This does give us *some* information – it picks out arm movement rather than, say, leg movement – as the cause. But arguably it does not provide as much information as citing my intentional flicking of the switch, which includes information about my mental state and its motivating the action. If the fuller explanation cites the better causal candidate, we still have grounds for taking the mental instantiation as causal.

Direct Causation in Hadron Collider

The purpose of the immediately preceding section was to support our view that HC is a case of *mental*, diagonal causation. We now turn to arguments that are intended to support our claim that the mental cause in HC is indeed efficacious *not* in virtue of a causal relation between *p*, upon which the mental cause supervenes, and the relevant physical effect: that the mental cause is *direct.*

Let's recall the Direct Causation and Directness principles:

Direct Causation: For some physical *p* and some mental *m*, *p* is directly caused by *m.*

Directness: A causal relation *R* between *c* and *e* is direct iff *R* does not obtain in virtue of a causal relation that obtains between properties *p*, upon which *c* supervenes, and *e.*

We want to show direct causation is plausibly at work in Hadron Collider. According to Directness, the directness of the putative causal relation between M1 and P2 requires that M1 causes P2 not in virtue of M1's supervening upon a physical property or in virtue of P2's subvening a higher-level property. My argument focuses only upon the relation obtaining not in virtue of M1's supervening upon a physical property because the issue of P2' subvening some M2 is not relevant to the Hadron case, where there is no suggestion of the Intervention or Dependence routes to causation that depend upon a supervenience relation between M2 and P2.

Direct Causation: Counterfactual Worlds Argument

Here we assume, for the sake of argument, that M1 does supervene upon P1, and that P1 is the sufficient physical cause of P2. That is, we assume that my intentional flicking of the switch supervenes upon some physical property P1, and that physical property is sufficient to bring about the collider's initial subatomic activity. (As we will later see, the Direct Causation model provides no reason for thinking this specific supervenience relation obtains. But we are here offering motivation for thinking that model apt for articulating what's going on in the Hadron case. So we are entitled, in attempting to show this, to our assumption.)

The central idea of the argument is this: even if we assume that M1 does supervene upon P1, qua sufficient cause of P2, we have reason to suppose that M1's causing of P2 is not in virtue of this relation. That is to say, we have grounds for denying that M1's causing of P2, even given the present supervenience assumption, obtains by piggybacking upon the horizontal causal relation between P1 and P2.

The key move here will be warranted by a supposition I first introduced in my previous discussion of the Dependence model of Exclusion: that mental properties are multiply realizable at close counterfactual worlds. My argument has it that P2 causally depends upon M1 independently of any counterfactual dependence of P2 upon P1 because if M1 is multiply realizable at close counterfactual worlds, P2 *does not counterfactually depend* upon P1.

Here are our assumptions:

(a) M1 causes P2.

(b) M1 supervenes upon P1 qua sufficient cause of P2.

(c) Lewisian counterfactual dependence is sufficient for causation, and M1 causes P2 in virtue of such dependence.

(d) M1 is multiply realized at close counterfactual worlds.

We begin by considering the following counterfactual, required on assumption (c) for causation of P2 by M1: [If M1 had not occurred, then P2 would not have occurred]. The non-trivial truth of this counterfactual requires that the closest [not-M1]-world is also a [not-P2]-world. So, on the assumption of M1's causing P2 in virtue of P2's counterfactually depending upon M1, the closest [not-M1]-world is a world where [not-M1 & not-P2].

Given Completeness, it follows that the collider's activity has a sufficient physical cause, P1. So we assume that P1 causes P2. Now, assuming Lewisian semantics for counterfactual conditionals, the counterfactual dependence of P2 upon P1 would require that the closest [not-P1]-world also be a [not-P2]-world. But the closest world at which P1 fails to occur is plausibly a world at which M1 does occur, alternatively realized – as per its close multiple realisability – by P*, and brings about – as per P2's counterfactual dependence upon M1 – P2. If this is the world pertinent to evaluation of the counterfactual [if P1 had not occurred, then P2 would not have occurred], then the counterfactual is false, for we have here the absence of P1 with the instantiation of P2. In which case, it would follow that P2 does not counterfactually depend upon P1.

Why should we think that the closest [not-P1]-world is a world at which M1 instantiates? Why not just evaluate [if P1 had not occurred, then P2 would not have occurred] by envisaging a world at which we, so to speak, *snip out* P1 without committing to M1's occurrence? Whilst I take there to be no reason that compels us to take M1 as instantiated at the closest [not-P1]-world, there are some plausible grounds for so doing. For if M1 is closely multiply realizable, then it is alternatively instantiated in some close counterfactual worlds. And whilst, as per M1's supervening upon P1, every world at which P1 instantiates is one at which M1 instantiates, this is compatible with close worlds at which M1 instantiates whilst P1 does not. Given this, it seems that we need more than a Lewisian 'small miracle' to envisage a

counterfactual world at which both P1 and M1 fail to instantiate; P1's failure does not entail that of M1. On the basis that the closest world is one at which we have, so to speak, the smallest miracle required for the antecedent to hold, we should assume that M1 instantiates at the closest world.

Why think that, if M1 instantiates at the closest [not-P1]-world, then P2 instantiates at that world? Recall that M1 causes P2 in virtue of P2's counterfactually depending upon M1. It follows that the following counterfactual conditional obtains: [if M1 had occurred, then P2 would have occurred]. In order for that conditional to obtain, the closest [M1]-world must also be a [P2]-world. So if the closest world at which M1 is realized by alternative realiser P* (rather than P1) is the closest [not-P1]-world, that world is one at which P2 instantiates.

We thus have reason to think that P2 does not counterfactually depend upon P1, despite P1's status as sufficient cause of P2.[54] However, as per our assumptions, P2 does counterfactually depend upon M1. On this basis, we might reasonably deny that M1's causing of P2 obtains in virtue of the causal relation between P1 and P2. It would follow that the mental cause of the collider's starting is directly causally efficacious. We are reasonably entitled to treat my intentional flicking of the switch as the direct diagonal cause of the physical effect. The mental causal relation is not supervenience-based causation.

We now have an argument that supports our intuitions of Hadron Collider as a case of direct mental causation. As such, the argument justifies aspects of our practice in assessing and deploying the counterfactuals pertinent to that direct causal relation. These practices reflect our pre-theoretic attitude towards cases like Hadron, where, I claim, we would typically regard the mental cause as direct. The first practice is that of assessing the relevant counterfactual in connection to the mental cause. The second practice is that of deploying the relevant counterfactual when thinking about future cases.

[54] This result is, admittedly, somewhat counterintuitive in itself: it is perhaps natural to expect that, if P1 causes P2, then P2 is counterfactually dependent upon P1 (at least if we make the equivalent assumption viz. P2 and M1). So this raises a puzzle. However, I return to this in Chapter 5, where I suggest that Interventionists might profitably drop their conception of causation for the physical domain and embrace causal pluralism. This position would involve endorsing a nomological regularity conception of causation for the physical, hence permitting the absence of counterfactual dependence alleged here.

Practice: Counterfactual Assessment

The relevant counterfactual here is [if M1 had occurred, then P2 would have occurred]. The key aspect of our practice in assessing this counterfactual is that we hold fixed the mental instantiation (M1) but we *do not* hold fixed any particular physical property that is, by Completeness, causally sufficient for our physical effect, P2. It is true that, given Supervenience, there must be *some* physical property (or properties) subvening M1, and given Completeness, there must be *some* physical property causally sufficient for the instantiation of the physical effect, P2. But when evaluating the counterfactual [if M1 had occurred, then P2 would have occurred], we do not hold fixed any particular physical property underlying M1. We consider close worlds at which the switch is intentionally flicked and whether the micro-effect occurs in those worlds, with no thought to the instantiation of whichever particular physical properties might subvene the macro-cause at those worlds or might be causally sufficient for P2 at those worlds. Evaluation of the counterfactual involves no presupposition of a particular physical subvenient base of M1 or a particular physical cause for P2 at close worlds.

Practice: Projection on Basis of Counterfactual

Here, we focus on the apparent irrelevance of any counterfactual holding between a physical property subvening M1 (or indeed *any* physical property sufficient for P2) and P2 when we reason about, and project forwards, the counterfactual relation obtaining between P2 and M1.

Not only do we not hold fixed any subvening P1 when assessing the counterfactual [if M1 had occurred, then P2 would have occurred] – as claimed above – we do not regard the counterfactual [if P1 had occurred, then P2 would have occurred] as important in inferences about the causal relation between my intentional flicking of the switch and the collider's activity. Suppose that I am scheduling Hadron research periods for 2023. In my plans, I implicitly make use of the counterfactual supposition that were I to intentionally flick the Hadron switch, the collider would start its process. Accordingly, I intend to flick the switch at the beginning of the next scheduled period. But I take it as plausible that I do not thereby implicitly appeal to the thought that were some particular physical cause P1 to occur, then

the collider's activity would occur. This thought is simply not pertinent to my plans; it is not required for my prospective exploitation of the causal relation between my intentional flicking of the switch and the collider's initiation. My practice of deploying counterfactuals in making plans centering upon the mental cause of the collider's starting simply ignores the physical cause of P2. The relevant counterfactuals are those that include the mental cause - my intentional flicking of the switch – and only those.

Direct Causation: Summary

I have presented two arguments in support of the physical effect, P2, being directly caused by M1. I intended the first argument to give grounds for taking the mental description of the switch-flick as the pertinent one. The second argument was marshalled in support of the claim that the mental cause of the collider's activity is not efficacious on the basis of some causal relation between a posited supervenience base P1 and that activity. The mental cause is not supervenience-based.

Section 3a: Direct Causation & My Preceding Arguments

At this point, we should recollect the broad aim of the chapter. My central message is one of caution: we should take care not to overgeneralize on the basis of the preceding arguments concerning systematic overdetermination, and rule out all diagonal causal relations. We *should* not, for two reasons. First, nothing in those earlier arguments entails that there be no diagonal causal relations; second, cases like Hadron Collider are plausible candidates for such relations. I have so far addressed the second reason. Now I address the first.

To start, I make explicit an obvious but nonetheless important point. My earlier arguments concerning Dependence, Determination, and Intervention concluded that none of the models entails diagonal causation, and hence none entails systematic overdetermination. I denied that those models respectively entail diagonal causation. But that is logically distinct from denying diagonal causation. The distinction between these two denials is the crux of what I'm trying to show in this chapter. The conclusion of my earlier arguments and my claim about Hadron Collider are logically compatible.

This leaves the question of compatibility between the claims of my earlier arguments and the affirmation of diagonal causation in Hadron (or in other similar cases).

The relevant argument here is the one regarding the Determination model, since the potential routes to overdetermination under that model are the most plausible candidates for the Hadron case. My main critical claim for that model was that it required an implausibly local formulation of supervenience. But given that Hadron is here taken as a case of direct causation, and hence not supervenience-based, it will not contradict my earlier argument by way of standing as a counterexample, i.e. as an example of supervenience-based causation that does not require a radically local supervenience thesis for its entailment by the core Exclusion principles.

I therefore take Direct Causation, as a model of direct causation in Hadron Collider, to be compatible with my earlier arguments of Chapter 2.

Section 3b: Direct Causation - A New Exclusion Problem?

The compatibility of Direct Causation with my arguments of Chapter 2 is important. But the model can only be of use if it avoids implying systematic overdetermination in some other way.

I commented earlier that I have deliberately included the core Exclusion principles in my articulation of the Direct Causation model. That is because I want it to be available to those non-reductive physicalists who want to endorse those core principles. This was, after all, why the Exclusion Problem originally worried such physicalists; they wanted to endorse all the principles, but were inclined, after Kim, to accept that they could not, on pain of inconsistency. The message of Chapter 2 was intended as a message of optimism for those physicalists, and so I wish to address them with my cautionary message here. As such, I want to retain the core Exclusion principles in my model of direct mental causation. But I must avoid spoiling my earlier message with a model that entails systematic overdetermination.

I take it that there is no need for worry here. Direct Causation is not a model of diagonal causation that threatens to entail problematic overdetermination. Or at least it doesn't if it is the *systematicity* of overdetermination that would render it problematic. On this view, only

if diagonal causal relations are guaranteed as systematic by a suitably general – and so iterative – principle will those relations constitute the kind of problematic overdetermination cited by the Exclusion Problem.

This is indeed how we have characterised *systematic* overdetermination. In the Introduction, I stipulated the following notion of problematic overdetermination: overdetermination is problematic only if systematic, and systematic iff entailed by a modally strong, general principle (together with Completeness). As I pointed out earlier, this constraint makes a stipulation against systematic overdetermination compatible with widespread cases of the firing squad kind. It also makes the absence of systematic overdetermination compatible with there being cases of overdetermination which are (in some sense) the contingent result of adopting arbitrary social norms, for example.[55] But what my formulation, together with a prohibition on systematic overdetermination, rules out are cases of overdetermination that are entailed by Supervenience and Completeness.

There are grounds, evident in the Exclusion Principle itself, for taking overdetermination to be problematic insofar as it is systematic:

Exclusion Principle: No single event can have more than one sufficient cause occurring at any given time—unless it is a genuine case of causal overdetermination. (Kim, 2005)

As it's commonly construed, a 'genuine case of causal overdetermination' is one where each sufficient cause occurs independently of the other. Standard illustrations of such overdetermination include the classic firing squad case. Given this reading, the Exclusion Principle is concerned with *ruling out* cases of overdetermination where the competing causes are not independent occurrences. It is therefore reasonable to think that the Exclusion Principle is motivated by the implausibility of systematic non-independent overdetermination.

In its basic form, the Exclusion Problem deploys the Exclusion Principle against the putatively non-independent, and systematic, diagonal causal relations taken as implied by the core Exclusion principles, with Supervenience being particularly salient. Since such causal

[55] See the clarifications section of the Introduction.

overdetermination is taken to be implausible, the Exclusion Principle then rules it out: hence the alleged inconsistency of the conjunction of Exclusion propositions.

But Direct Causation is not a model of diagonal causation that entails systematic overdetermination. Systematic overdetermination is here defined as overdetermination that is entailed by a general, modally strong principle (together with Completeness). The obvious candidate is the Supervenience principle in the Direct Causation model. But Direct Causation is a model of diagonal causation that is direct, and so does not affirm supervenience-based causation. The mental causation of Direct Causation does not arise from supposition of supervenience relations between mental and physical properties; since it was those relations that were supposedly responsible for systematically generating overdetermination in the earlier Exclusion models, it follows that mental causation under Direct Causation does not imply systematic overdetermination on grounds of supervenience.

Nor does this model include some other generalizing principle by which to generate systematic overdetermination. To recap, Direct Causation comprises:

Causation: Mental properties have – in their own right qua mental properties – genuine causal efficacy.

Distinctness: Mental properties are metaphysically distinct from physical properties.

Direct Causation: For some physical *p* and some mental *m*, *p* is directly caused by *m.*

Directness: A causal relation *R* between *c* and *e* is direct iff *R* does not obtain in virtue of a causal relation that obtains between properties *p*, upon which *c* supervenes, and *e.*

Supervenience: Mental properties supervene upon physical properties.

Completeness: All physical effects have sufficient physical causes.

Exclusion: No single effect can have more than one sufficient cause occurring at any given time—unless it is a genuine case of causal overdetermination.

I take it as self-evident that none of the above generalizes the diagonal causal relation in our Hadron case. None, that is, imply systematic direct diagonal mental causation. Causation entails only that some mental properties are causally efficacious. Distinctness is general in

claiming that all mental properties are metaphysically distinct from physical ones but, when conjoined with the other principles, only entails that an efficacious mental property will be metaphysically distinct from any physical property. It does not entail systematic diagonal causation. Direct Causation is not general. And Completeness, whilst general, is alone consistent with there being only one direct diagonal mental cause. Direct Causation does not provide the resources with which to press a new Exclusion Problem: it lacks the requisite generalizing principle with which to imply systematic diagonal causation.

Section 4a: Coherence – Completeness & Supervenience

I have now argued that Hadron Collider should be read according to the principles of Direct Causation; that those principles articulate our intuitive judgements about diagonal mental causation in the case.

I have also shown that the Direct Causation model is not in tension with my earlier arguments against systematic overdetermination. Those arguments attacked the claim that the set of Exclusion principles , when articulated by a Dependence, Determination or Intervention model, entail systematic overdetermination. But the attack was consistent with the claims here made for direct diagonal causation as per the Direct Causation model.

Finally, we have just seen that Direct Causation does not threaten to imply systematic overdetermination.

But there is still the question of whether a Direct Causation reading of Hadron Collider is fully coherent. We should remember that the significance of Direct Causation / Hadron Collider for our overall discussion rests ultimately upon its potential for articulating diagonal causation in a manner consistent with the commitments of the non-reductive physicalist sympathetic to the core Exclusion propositions. Accordingly, we must also examine whether a direct, diagonal causal relation in this case is really consistent with the collider's activity having a sufficient physical cause, and really consistent with the supervenience of m-properties upon p-properties. Is the Direct Causation model fully coherent?

In this section, I defend the coherence of the Direct Causation model. The Direct Causation and Directness principles are consistent with the other core Exclusion principles, Distinctness, Supervenience, Causation, Completeness, and Exclusion.

I start by considering the consistency of Completeness and Supervenience with asserting direct causation in Hadron Collider, i.e. with invoking the Direct Causation and Directness principles. I will consider Completeness primarily; the question of consistency with Supervenience will be answered along the way. Our question is this: in the HC case, does the directness of the diagonal relation entail the absence of a sufficient physical cause of P2 – the collider's initial activity?

Directness in HC does not entail the absence of a sufficient physical cause of P2. According to Directness, a causal relation R is direct iff it does not obtain in virtue of a causal relation that obtains between properties *p*, upon which *c* supervenes, and *e*. But clearly, to assert a direct causal relation between my intentional flicking of the switch and the collider's initial activity does not thereby imply that there are no specific supervenience relations borne by either of them. It only implies that whatever supervenience relations they might bear, the cause and effect do not qualify as such *because* they bear those relations. Nor does a direct causal relation in Hadron imply the negation of Supervenience formulated globally. We have answered the question about consistency with Supervenience.

Now, to address Completeness, there is no reason to suppose that Completeness requires supervenience relations of any sort. Completeness is not a principle motivated by theses concerning synchronic relations between physical properties and higher-level properties. However Completeness is motivated, it is only causal relations *between physical* properties that are salient. It seems that a world in which every physical effect had a sufficient physical cause could be a world in which mental properties were dualistic, metaphysically independent of physical properties. There is no apparent logical entailment from Completeness to Supervenience; in which case, no reason to worry that, even if we were here implying absence of supervenience relations borne by mental cause and physical effect, respectively, we would also imply the negation of Completeness.

There is another reason why direct causation in Hadron potentially might be thought to conflict with Completeness. This is the basic worry that drives some expressions of the

Exclusion Problem: that Completeness entails that the physical effect (the collider's activity) has a sufficient physical cause, the latter excluding any higher-level candidate cause, such as my intentional flicking of the switch. But to respond to this is effectively to show that the Exclusion principle itself is consistent with the other principles in Direct Causation, particularly the Direct Causation and Directness principles. For the question of consistency between direct causation in Hadron and the Exclusion principle is equivalent to the question of whether asserting direct causation here entails systematic overdetermination. This is the question of my next section.

Section 4b: Coherence – The Exclusion Principle

Exclusion: No single event can have more than one sufficient cause occurring at any given time—unless it is a genuine case of causal overdetermination.

The Exclusion principle effectively asserts that the only kind of overdetermination is independent overdetermination, i.e. cases such as the firing squad. Above, I speculated that the reason that non-independent overdetermination is to be excluded is that this form of overdetermination is systematic. Now, it is true that I have just claimed that Direct Causation does not imply systematic diagonal causation, and hence does not imply systematic overdetermination. To that extent, we have considered the relationship between Direct Causation and the Exclusion principle. But though this model of Hadron Collider does not imply systematic overdetermination, one might worry that it could yet contravene the Exclusion principle if the causal relation between M1 and P2 is not independent of a causal relation between P2 (as effect) and some other physical property. For the principle rules out *any* instance of non-independent overdetermination.

There is thus more to be said in respect of the *independence* or otherwise of the direct diagonal relation between the mental cause and physical effect in Hadron Collider, and so more to be said about how Direct Causation relates to the Exclusion principle.

In this section, I argue for the following claims:

1) In the absence of supervenience-based causation, there are no grounds for taking M1 to supervene upon the sufficient physical cause – implied by completeness – of P2. So in the

Hadron case, there are no grounds for worrying that M1's causing of P2 is non-independent in the relevant sense.

2) Even if M1's causing of P2 *were* non-independent in the relevant sense, the diagonal mental causation in Hadron would not contravene the substantive prohibition expressed by the Exclusion principle.

Independent Causation in Hadron Collider

For the sake of illustrating my central point, I can borrow Bennett's (2003, p.476) counterfactuals as necessary conditions for independent causation[56].

IC1: If M1 had not occurred, and P1 had occurred, then P2 would have occurred.

IC2: If P1 had not occurred, and M1 had occurred, then P2 would have occurred.

The question of causal independence in Hadron Collider is then partly the following question[57]: does the directness of causal relation between M1 and P2 imply the satisfaction of either or both of these conditions?

My first, and main, claim here is that we need not ask the question of the Hadron case: the debate here is unnecessary. A debate around independent causation in contexts of direct causation would be predicated upon a mistake.

To see this, we should step back and ask another question: Why are we concerned with whether M1 and P1 are independent causes of P2? Because if M1 and P1 are independent causes, then their overdetermining of P2 is not systematic and so – on our terms – not problematic. It is a case that, like the over-efficacious firing squad, has a free pass.

But why do we suppose that the question of causal independence arises here? Why think that M1 and P1 are candidates for non-independent causation? Why worry? The answer, once again, is that M1 is assumed to supervene upon P1, where 'P1' denotes the sufficient cause

[56] We should note here that Bennett's concern with these conditionals is different to mine. On her view, these are necessary conditions for *problematic* overdetermination. On my view, these are necessary conditions for independent overdetermination. When conjoined with our view that independent overdetermination is not systematic, this means that these are necessary conditions for overdetermination that is *not* problematic (or at least, not problematic in the sense ruled against by the Exclusion principle).

[57] It is partly this question, but not wholly – these are stated as only necessary, not sufficient conditions for independence.

that P2 must, by Completeness, have. It is the synchronic metaphysical relation that supposedly holds between M1 and P1 that is the source of the concern. To put it roughly: if M1 is bound to P1 by supervenience, then any causal move made by P1 seems to bring M1 along with it. And if that is right, then – so the worry seems to go – we cannot think of M1 as independently causing P2.

What is the problem with *non*-independent causation here? Well, given the foregoing, it is the supervenience relation between M1 and P1 that renders M1's causing of P2 non-independent. So assuming our conception of systematicity, if M1 non-independently causes P2 then M1 systematically causes P2. That's because we defined systematicity of causation as causation entailed by a modally strong, general principle (together with Completeness).[58] The general principle at work here is Supervenience: the causal efficacy of M1 is entailed by its supervening upon P1. And if M1 systematically causes P2 whenever P1 causes P2, we have systematic overdetermination of P2.

So the question of independent causation is supposed to be germane because only independent causation enters into benign overdetermination; non-independent causation implies problematic overdetermination.

But now we ask: why worry about the question of independent causation in the Hadron context – i.e. when thinking about my intentional flicking of the switch and its causing of the collider's subatomic activity?

I want to claim that no such worry is here necessary. The reason is straightforward: given that M1 does not cause P2 in virtue of any supervenience relation – i.e. that M1's causing of P2 is not supervenience-based causation - there is no reason to suppose that M1 does supervene upon P1 qua sufficient cause of P2.

In the Hadron case, we posit my intentional flicking of the switch as the cause of the collider's starting. So we have M1 causing P2. We also have it that M1 – as a member of the global class of m-properties – supervenes upon physical properties. And we have it that P2 – as a physical property covered by the Completeness principle – has a sufficient physical cause that we designate as P1. But there is no reason to suppose that P1 is the supervenience base of M1.

[58] See the Metaphysical and Terminological Clarifications section of the Introduction.

What *would* motivate the claim that M1's supervenience base is P1? If M1 got to cause P2 in virtue of its supervening upon a physical cause of P2, then to that extent it would be plausible that M1's supervenience base was that physical property that causes P2. If my intentional flicking of the switch caused the collider's subatomic activity in virtue of piggybacking, via supervenience, upon the causal relation between that activity and a prior physical cause, then there would be reason to take my intentional action to supervene upon that physical cause.

But we have argued that Hadron is not that kind of case. On the contrary, it is a case of direct diagonal causation, where M1's causing directly is to cause not in virtue of a causal relation between M1's *supervenience* base and the relevant effect. So mental causation in Hadron does not raise the question of causal independence – does not demand evaluation of counterfactuals (IC1) and (IC2) – because the grounds for worrying that it might be a case of non-independent causation are absent.[59]

I have argued that there are no grounds for taking the direct mental cause in Hadron to be non-independent with respect to P2's physical cause, P1. We therefore have no grounds for taking M1's causing of P2 to violate the Exclusion principle. The upshot is this: if the direct mental causation in Hadron is consistent with the Exclusion principle, then the Exclusion principle is consistent with all the principles of Direct Causation. Direct Causation is a coherent model.

[59] What would happen if we did? We should bear in mind that establishing the non-vacuous truth of (IC1) and (IC2) would not thereby establish causal independence; these are only proposed as *necessary* conditions. Still, there are grounds for thinking that they are non-vacuously true in the Hadron context. For all Causal Convergence / Hadron Collider tell us, (IC1) is not vacuous: there are possible worlds in which P1 instantiates but M1 does not. This is due to our main point, i.e. that we have no grounds in the present context to take M1 as supervening upon P1 qua sufficient cause of P2. It is also reasonable, for the same reason, to think that (IC2) is not vacuous. (Furthermore, even if M1 did supervene upon P1, multiple realizability of M1 would support the possibility of M1's instantiating in the absence of P1.) I think (IC1) is true when considered in the context of Hadron Collider partly, again, because there are no grounds for taking M1 to supervene upon P1. So if P1 were to occur in the absence of M1, I see no reason to think that P2 would not occur; P1 is by hypothesis sufficient for P2. (Bennett (2003) thinks otherwise, but her view assumes that M1 supervenes upon P1.) Finally, there is no reason to suppose that (IC2) is false. This is for two closely related reasons. First, in HC, M1 causes P2, and not in virtue of its supervening upon P1 (this formulation is *not* intended to imply that M1 does supervene upon P1; only that M1's efficacy with regard to P2 does not obtain in virtue of supervening upon P1 even if it does so supervene). So there is no reason to think that it would not cause P2 in a world lacking P1. Second, due to the direct causation of P2 by M1, there is anyway no reason for taking M1 to supervene upon P1. In which case, there seems no reason to think that a world lacking in P1 would make any difference to M1.

Consistency with the Substantive Spirit of Exclusion

However, even if the direct mental causal relation in HC is *not* causally independent, it is not the end of the road for the Direct Causation model. That's because non-violation of the Exclusion Principle is arguably not necessary for the proponent of Direct Causation who wishes to respect the motivation and spirit of the Exclusion propositions.

We speculated above that the motive for the Exclusion Principle is to rule out systematic overdetermination. If, as per my reading of the principle, any case of non-independent causation is deemed to threaten systematicity of diagonal causal relations, then a general principle that outlaws such causation is prudent. However, if this is the motive then the diagonal relation in HC can respect the *spirit* of the law laid down by the Exclusion Principle, even *if* it does not respect the letter.

The reason for this is straightforward. I argued above that Direct Causation does not imply systematic overdetermination because it lacks any generalizing principle by which to systematically generate diagonal causal relations. So even if, contrary to what I have claimed, the causal relation in HC is not causally independent, and hence violates the Exclusion Principle, it is a harmless violation. It does not imply the kind of systematic generation of overdetermining relations that the Exclusion Principle is concerned to exclude and so, in this sense, is consonant with the spirit of the principle.

If the purpose of the Exclusion principle is to rule out systematic diagonal causation then, even if it strictly violates it, direct mental causation in Hadron is consistent with its motivation. We might think of it this way: if the purpose of the Exclusion principle is to rule out systematic diagonal causation, then we could, in a manner consistent with that purpose, amend the principle so as to permit non-systematic, local cases of diagonal causation. If we're right about the purpose, then this amendment would be arguably permissible. We will not do so; the point is just that the more important principle here is surely the motivating one. With that in mind, non-independent causation under Direct Causation would plausibly be benign in its technical violation of Exclusion. If so, then Hadron Collider as articulated via Direct Causation is – in the sense that matters – consistent with the Exclusion principle, and hence Direct Causation is coherent.

Conclusion

The primary message of this chapter has been this: whilst we can deny the entailment of diagonal causation by the core Exclusion principles, we need not thereby think that no diagonal causation occurs or could occur. I have argued that common-sense cases such as Hadron Collider should be considered examples of such diagonal relations, from the mental to the physical. Such cases are not undermined by the arguments of Chapter 2, because they are not cases of supervenience-based causation; rather, Hadron Collider is a case of *direct* diagonal causation.

Furthermore, I have claimed that the corresponding model, Direct Causation, does not threaten systematic overdetermination, and so does not threaten a new form of Exclusion Problem. I have also argued that this model plausibly does not imply non-independent mental causation, and so is consistent with the Exclusion principle itself. If so, then it is consistent with the core set of Exclusion principles and is available to our non-reductive physicalist. But even if it is a case of non-independent causation, I take Hadron Collider and the Direct Causation model to be broadly consistent with the underlying concern of the principle: to rule out systematic overdetermination. For on my formulation, overdetermination is systematic only if entailed by modally strong, general principles such as strong supervenience. It follows that Hadron Collider, as a case of non-supervenience-based causation, is not a case indicative of such overdetermination.

In summary: the non-reductive physicalist can endorse the core Exclusion principles, reject systematic overdetermination and embrace non-systematic, direct diagonal mental causes. The psychological and the physical need not run parallel.

Chapter Four - Pluralism & Exclusion

Introduction

In this chapter, I consider a recently proposed solution to the Exclusion Problem from Campbell (2020). Roughly, the solution is as follows. First, causation is conceived in Interventionist terms, so causal relations are relativized to variable sets. A mental cause qualifies as such only by reference to an appropriate variable set; the same holds for a physical cause.

Second, mental and physical causes related by supervenience, that might be thought of as competing, must be assigned to mutually exclusive variable sets. So take a mental cause and the subvening physical cause that threatens to exclude it, and assign each to distinct variable sets. The set relative to which the mental cause is explicated as such, and the set relative to which the physical cause is explicated, each exclude the other cause.

Third, causal overdetermination is defined as itself something that only happens relative to a single variable set; the mental and physical causes could only overdetermine an effect if they were both causes relative to the same variable set. Given that they are not, they cannot combine to overdetermine an effect. Indeed, given that they *could not* be, they could not overdetermine an effect.

I will argue that the solution fails, because the key move in step two – the exclusive assignation of mental and physical causes to distinct variable sets – is neither methodologically viable nor warranted.

In Section 1, I present Campbell's solution in more detail. In Section 2, I consider the potential warrant for the assignation of mental and physical causes to exclusive variable sets. In Section 3, I marshal arguments against the viability of and warrant for this exclusive assignation. I then offer, in Section 4, a diagnosis of how a proponent of the solution might get into the position of taking it to be warranted. Finally, this diagnosis will suggest a more general diagnosis of how proponents of the Exclusion Problem might find themselves worried by the threat of exclusion for mental causes.

Section 1: The Pluralist Solution to Exclusion

Campbell's solution to the Exclusion Problem consists of the following claims, the conjunction of which I shall henceforth call, 'pluralism':

Relativity: Any causal relation is relative to some variable set.

Leanness: For any variable *V* in a set of endogenous variables for explication of causal relations, there must be no variables related to *V* by constitutive dependence.

Overdetermination: If causal variables *c* and *c** overdetermine effect *e*, then *c* and *c** are causally related to *e* relative to the same variable set.

The Relativity principle is a standard Interventionist principle. As outlined in the section of Chapter 2 on the Intervention model of Exclusion, causal relations under Interventionism are explicated only relative to a variable set. The relevant variable set for explication of a causal relation will include the candidate cause, the outcome variable, and a number of other variables that are to be controlled for. The latter are essential because the key idea of Interventionism is that causes of an outcome variable are variables the manipulation of which would enable the manipulation of the outcome. Or to put it slightly differently: the causal variable makes a difference for the outcome variable; when a causal relation holds, to nudge the cause is to nudge the outcome. But whether or not a causal variable makes a difference to an outcome is a matter that can only be ascertained or explicated when certain other variables are controlled for or held fixed. This is predominantly because causal variables and their outcomes do not obtain in a vacuum; other variables might also make a difference to the outcome. So to isolate the difference-making capacity of a causal variable, we must effectively switch off the influence of other causal variables.

Consider the following toy example. We notice correlations between eating extremely spicy curry and stomach cramps. Eating extremely spicy curry is a candidate causal variable for the outcome variable of stomach cramps. (Let us also suppose that this is a *bona fide* cause of stomach cramps, but that we do not yet know this). We want to ascertain whether we can manipulate the occurrence of stomach cramps by manipulating the ingestion of such curry. But we might also have noticed on these occasions correlations between the occurrence of

stomach cramps and the imbibing of lager. Lager is, after all, often consumed at the same time as extremely spicy curry. What should we do?

The Interventionist tells us that we must construct a variable set that includes both the causal candidate (eating extremely spicy curry) and the other correlated variable (drinking lager), in addition to the outcome variable (the onset of stomach cramps). These are all *endogenous* variables. Then we must contrive an intervention variable that enables manipulation of the causal candidate whilst holding fixed the other correlated variable: we need to be able to nudge the curry variable whilst holding fixed the lager variable. The intervention variable is an *exogenous* variable because it is not to be included in the variable set itself. It, so to speak, comes from outside the endogenous correlations and serves to isolate the difference-making capacity of eating extremely spicy curry with respect to experiencing stomach cramps. In practice, such interventions are often achieved by randomized controlled trials. So in this case, we would want to take the group of people in which we have the correlations between eating curry and having stomach cramps, and between drinking lager and having stomach cramps. Then we randomly assign each individual to one of two groups. Call them A and B. In group A, we give each individual extremely spicy curry and lager to drink; in group B, we give each individual just the curry. Thus we effectively switch off lager-drinking variable in group B. By doing this, we should be able to observe the correlation – in isolation – between the curry-eating and stomach cramps.

If we can do this, then we will be able to see that there is an isolated correlation between values taken by the curry variable and values taken by the stomach cramps variable. We will see that eating extremely spicy curry is a cause of stomach cramps. It is only by relativizing the articulation of that causal correlation against the background of other (causal and non-causal) correlations with stomach cramps that we can isolate the difference-making potential of our causal candidate variable.

Next we have the Leanness principle. This principle plays the central role in Campbell's solution, because it is ensures that constitutively related variables do not belong to the same endogenous variable set. It therefore underwrites the construction of mutually exclusive variable sets for explication of causal relations where two or more constitutively related variables are causal candidates.

How does Leanness apply to the Exclusion Problem? Campbell takes it that in approaching the Exclusion Problem, we must assign M_1 and its realiser P_1 to different variable sets. This is because, given Supervenience, M1 and P1 are not (according to Campbell) *constitutively independent*. P1, as subvening realiser of M1, *constitutes* M1.[60] Given the Leanness constraint on variable sets, such that no single set can contain variables that are constitutively dependent upon other variables within the set, it follows that M1, as constitutively dependent upon P1, cannot be assigned to the same set as P1.

Campbell works with the assumption that both M1 and P1 cause mental effect, M2.[61] So when constructing variable sets relative to explication of M1 and P1 as causes, we have one which includes M1 as causal variable with respect to M2, and one which includes P1 as causal variable with respect to M2. Given the Interventionist commitment to causation as relative to a variable set, this exclusive assignment of variable sets implies two distinct causal relations, one relative to the set containing M1, the other relative to the set containing P1. If M1 causes M2, that relation is relative only to the variable set that contains M1 as causally relevant variable for M2; if P1 causes P2, that is relative only to the other variable set. Since, when considering causes of M2, no single set can contain both M1 and P1, it follows that correct specification of M2's cause depends upon which variable set we are considering. Campbell asserts that M1 does indeed cause M2, relative to the variable set containing M1 as relevant variable, and that P1 causes M2 relative to its set.

This brings us to the third principle of Pluralism: Overdetermination. With this, we have a single-set constraint on overdetermination. Campbell claims: "Causal overdetermination is when we have two sufficient causes for an effect within a single variable set, such as when we have both a bullet from the right striking the heart and a bullet from the left striking the heart (2020, p. 157)".

---

[60] In the immediate context (2020), it is not entirely clear what Campbell intends by his term, 'constitution'. However, in his "Independence of Variables in Mental Causation"(2010, p. 65), he indicates that for X to constitute Y is for X and Y to be the "same phenomenon…identified in two different ways". He offers the example of John Campbell constituting the oldest person in his family. On this basis, we will assume that in the present context, M1 is constituted by P1 if M1 is realised by P1, where to be realised by P1 is for M1 to be necessitated by P1. Thus we connect the notion of constitution with Supervenience, which Campbell evidently endorses given his assumption that P1 constitutes M1 in the context of cases pertinent to Exclusion. Given that we are also assuming that mental properties are distinct from their realising properties, the relation of constitution holds between tokens not types.

[61] I do not question this assumption here. But in Section 3, I will; there are grounds for rejecting it.

By insisting upon the Pluralist nature of causal relations, and adopting a same-set constraint on causal overdetermination, Campbell ostensibly provides a solution to the Exclusion Problem. On his account, both M1 and P1 are causally efficacious with respect to M2, but M2 is not thereby overdetermined.

This result appears to respect each of the core Exclusion principles: Supervenience, Distinctness, Mental Causation, Completeness, and Exclusion. Presumably, it is Supervenience that motivates the claim that M1 and P1 are not constitutively independent. If we take Supervenience to entail P1's necessitating M1, and P1's necessitating M1 to imply P1's realising M1, and take P1's realising M1 to entail P1's constituting M1, then constitutive dependence is consistent with Supervenience. Distinctness appears consistent with the claim that M1 and P1 can be assigned to distinct variable sets. Mental Causation yields the claim that M1 is causally sufficient for M2, whilst nothing in Campbell's treatment contravenes Completeness. Finally, if the claim regarding causal overdetermination is right, then the account is consistent with the Exclusion principle. That principle rules out systematic overdetermination; if Pluralism and the same-set constraint on overdetermination hold, then M1 and P1 do not overdetermine M2, and so this principle is not violated.

Ruling Out Exclusion

Indeed, the Pluralist response to Exclusion, if successful, goes beyond showing how diagonal causation (in this case, from P1 to M2) is consistent with the Exclusion principle. It guarantees that, wherever some diagonal causal relation obtains in virtue of Supervenience, that causal relation is not a participant in overdetermining any effect. We might put this by saying that the Pluralist solution *inverts* the role of Supervenience. On standard readings of the Exclusion Problem, Supervenience generalises the implied diagonal causation, rendering it systematic. In this way, Supervenience plays the essential role in potentially violating the Exclusion principle. But on the Pluralist conception, Supervenience instead plays the role of generalising the allocation of constitutively related causes to different variable sets, in a systematic defence *against* overdetermination.

To see this, we should recall the Exclusion principle:

Exclusion: No single effect can have more than one sufficient cause occurring at any given time—unless it is a genuine case of causal overdetermination.

The principle does not, as formulated here, explicitly rule out systematic overdetermination. But it is typically interpreted as doing so on the grounds that 'genuine' cases of overdetermination are also typically taken to be only those anomalous scenarios of the kind familiar from examples such as the over-efficient firing squad or Billy and Suzy's bottle-breaking. In such cases, a single effect is thought of as brought about by distinct, and independently sufficient, causes. But the alleged cases of overdetermination resulting from the conjunction of mental causation, mental-physical supervenience, mental-physical distinctness, and Completeness are taken to be significantly different. In these, a common effect is the result of two distinct, but constitutively related – and so non-independent – causes. And, we might reasonably suppose, it is because of this constitutive relation that such cases are to be ruled out. If we assume a global supervenience thesis, then *all* mental properties are related by supervenience to the physical base. If we assume that some of those mental properties are causally efficacious, then every one of them will supervene upon some physical base. But, so the proponent of Exclusion will argue, if diagonal causation results from supervenience and the causal completeness of the physical, then diagonal causation will be systematic. On this basis, we can appreciate how the non-independence of distinct, overdetermining causes might appear problematic in a way that perhaps cases of independent overdetermination do not. The former case implies systematic overdetermination; the latter does not.[62]

On this reading, the Exclusion principle is violated by any case of non-independent overdetermination resulting from two distinct causes related by supervenience. It is supervenience that underwrites the unacceptability of such cases, since it threatens to guarantee (in conjunction with Completeness) the systematic production of causal overdetermination.

Whereas supervenience is typically taken to be the engine behind systematic overdetermination, it is, on Campbell's account, the engine behind systematic barring of such

[62] In the Introduction, I explained that I treat overdetermination as problematic only if systematic. I stipulated that overdetermination is systematic iff entailed by a general, strong modal principle such as global supervenience (in conjunction with Completeness).

overdetermination. On the Pluralist model of mental causation, the supervenience relation between mental and physical properties serves to rule out any m-causes from being variables in the same set as their subvening p-properties. Taken together with the principle by which all causes are efficacious relative to a variable set, this in turn implies that for any variable set relative to which M is causally efficacious with respect to effect *e*, subvening P is not – and vice versa. If, as Campbell contends, overdetermination requires sufficient causes relative to the same variable set, it follows that no m-property can overdetermine an effect with its subvening p-property.

Because the Pluralist strategy *rules out* variable sets including both mental and subvening physical variables, there can be no variable set relative to which both M and its subvening P are causally efficacious with respect to the same effect. So the same relation that seems to open up the potential for systematic overdetermination is also the relation which, on the Pluralist view, closes that potential down. In this way, Pluralism inverts the role of supervenience. Supervenience becomes the benign guarantor of causal distinctness for constitutively related variables.

Section 2: Motivating Leanness

Introduction

In this section, I want to introduce the motivation for applying the Leanness principle to variable set selection. This will be crucial to showing that, in defence of Campbell, Leanness is not merely an ad hoc constraint, drafted in for the purposes of dealing with Exclusion concerns. But furthermore, understanding the motivation for Leanness will put us in a position to appreciate why Campbell's appeal to the principle fails. My underlying assumptions in arguing against Pluralism will be these: (a) the methodological application of Leanness is warranted insofar as it is practically viable, and (b) the application of Leanness is warranted insofar as it is practically required. In Section 3, I will argue that the Leanness principle fails on the first count generally, and fails on the second when considered in Exclusion contexts, that is, in those contexts implied by the conjunction of Exclusion

principles. Before we can see how these arguments work, we need to understand the putative motivation for the Leanness principle.

Before proceeding, a number of clarifications are in order. First, we will be assuming that variables correspond to property types, construed as universals. Second, we will be assuming that the relation of *constitutive dependence*, central to Campbell's view of the Exclusion Problem, is entailed by mental-physical supervenience.

We assume that variables correspond to types for several reasons. For one, Campbell's approach to Exclusion assumes, as is standard, that the relevant higher- and lower-level entities subject to causal relations are metaphysically distinct. Since the entities typically implicated by a Distinctness thesis are property types, we should assume that the entities considered here (variables) correspond to property types. Furthermore, variables are defined as entities capable of taking a range of values. In the simplest case, this range will be binary, with the values being 0 and 1. If we assume that the value 0 corresponds to non-instantiation, it would make little sense to think of a token taking the value of 0; tokens are instances, so are here most naturally conceived as instances of a type taking the value of 1.

We assume that the relation of *constitutive dependence* is entailed by mental-physical supervenience because Campbell clearly takes the relation to hold in virtue of the standard Exclusion principles, and Supervenience is the principle most naturally taken to imply a form of synchronic dependence of higher-level variables upon physical variables. We have taken mental-physical supervenience to imply the necessary, synchronic dependence of mental properties upon physical properties, such that for any M dependent upon any P, P determines (with necessity) M. I take this notion of dependence to be plausibly conceived as 'constitutive' in the sense that it implies the determination (or equivalently, necessitation) of the relevant m-properties by the relevant p-properties. So if M is dependent upon P, then P determines M across all possible worlds and this necessitation is commensurate with P's *constituting* M. We need not take a view on whether P's constituting M involves more than this dependence-determination; nothing in what follows will turn on whether constitutive dependence is more than supervenience understood in the sense outlined here.

Furthermore, at this stage we accept, for the sake of illustration and argument, that the variables (strictly: the instances of those variables) allegedly related by constitutive

dependence are indeed so related. Later, when I come to discuss my objections to Pluralism, we will see that this assumption should be rejected in Exclusion contexts.

The section splits into two. In Section 2a, I outline two kinds of misleading correlation, the obtaining of which the Leanness principle is designed to preclude: *pseudo-causal* and *confounding* correlations.

In Section 2b, I examine in detail how these two kinds of correlation might be thought to arise given that m-variables supervene upon p-variables. We will see that there are four specific forms of potentially misleading correlation, two from each kind.

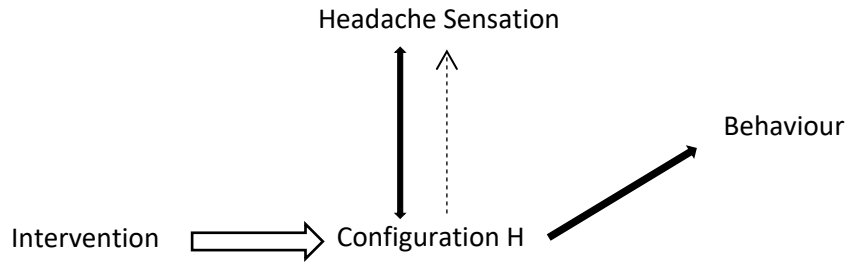
Section 2a: Pseudo-Causal & Confounding Correlations

(i) Pseudo-Causal Correlations

The motivation for Leanness is pragmatic, rooted in the perceived need to safeguard the explication of bona fide causal relations from two kinds of misleading correlations. I will call the first kind, 'pseudo-causal correlations', and the second, 'confounding correlations'. Campbell himself explains that Leanness is a necessary condition on variable set selection because if variables related by constitutive dependence were permitted in the same set, then we might find variables to be correlated under interventions "even though the correlation was explained by their lack of independence rather than by a causal relation between them" (2020, p. 152). Consider a variable set that includes my acute, specific headache pain and neural configuration H as distinct variables (**Figure 8**). Suppose for the sake of illustration that the occurrence of headache pain is constitutively dependent upon the occurrence of neural configuration H. Without prior knowledge of the constitutive dependence relation between these events, the interventionist investigator might, so the worry goes, be led to treat them as causally correlated. She might erroneously take my headache to be caused by the activation of neural configuration H. Why? Because interventions on the configuration would also, unbeknownst to our investigator, be interventions upon the headache. If the configuration were to be activated, my headache would manifest. Conversely, if the configuration were to be left alone, my headache would leave me in peace. A stable correlation between the values of the headache variable and the values of H would emerge.

But they are not distinct existences; so by Campbell's (and Hume's) lights, they cannot be causally related. Our investigator has been duped. Leanness is motivated in part by the concern to preclude, through judicious variable selection, such scenarios.[63] It is there to ensure that misleading, pseudo-causal correlations are excluded.

(Fig. 8)



The black double-arrow between H and the Headache Sensation represents the constitutive dependence relation. The black arrow between H and Behaviour is causal. The white arrow between Intervention and Configuration H is the causal relation by which the intervention acts upon the configuration. The intervention sets the value of the Configuration H variable which causally correlates with the value of the Behaviour variable. But the Headache constitutively depends upon H so the intervention also sets the value of H. Interventions upon the neuronal configuration produce pseudo-causal correlations between the configuration and the headache.

(ii) Confounding Correlations

Pseudo-causal correlations are one kind of misleading correlation, the prospect of which motivates Leanness. And, although Campbell does not himself cite it, we might reasonably suggest another. For without applying the Leanness constraint to variable set selection, one potentially leaves oneself open to the problem of *confounding* variables. The rough idea here is this: Suppose that variable *c* – say, the neuronal configuration H – is candidate causal variable for outcome variable *o* – say, my wincing and clutching my forehead. We carry out interventions on *c* and observe correlations with values taken by *o*. Intervening upon neuronal configuration H is found to be correlated with whether I wince and clutch my forehead. But suppose further that variable *c* constitutes variable *c** - my acute headache sensation. Then

[63] I should stress that I am not here taking a stance on whether this or the potential problem immediately below really are genuine problems in relation to Exclusion scenarios. I consider the extent to which these potential concerns are genuinely worrying below, in Sections 3b and 3c.

it might be thought to follow that any intervention upon neuronal configuration H is also an intervention upon my pain sensation. If so, then it seems we will have a correlation between *c** and *o* that confounds the candidate causal correlation between *c* and *o.* The correlation between my headache sensation and the behavioural outcome gets in the way of isolating the correlation of interest: that which holds between H and my wincing.

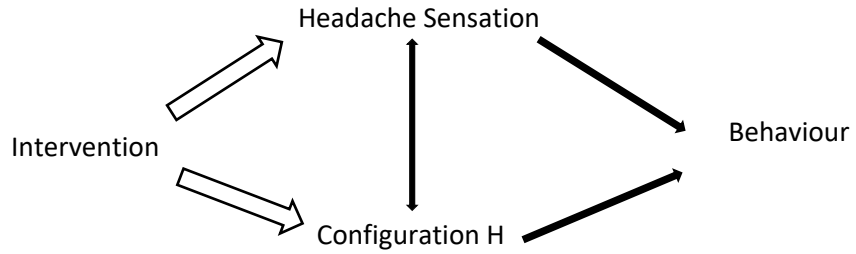
We might say that the problem here is that variables related by constitution might be impossible to separate so as to enable a legitimate intervention upon one as opposed to the other. If a legitimate intervention upon variable *c* effectively suspends or 'switches off' the influence of other causal variables with respect to the outcome variable of interest, then the issue here is that in the case of variables related by constitution, it might not be feasible to effect an intervention upon these variables. It might not be feasible to effectively isolate the difference-making capacity of either variable.

This notion of confounding variables, and the threat they pose to legitimate interventions, has been formulated in the work of Woodward (2003). To clarify, we should note the necessary and sufficient conditions he proposes for intervention variables. I is an intervention variable for candidate cause X with respect to outcome Y iff:

> **1.** I causes X;
>
> **2.** I acts as a switch for all the 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;
>
> **3.** 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 − X − 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;
>
> **4.** I is (statistically) independent of any variable Z that causes Y and that is on a directed path that does not go through X. (Woodward, 2003, p.98)

The most salient condition here is (3), especially the following: Any causal path from the intervention variable I to the outcome variable Y must go through X; I is not a cause of any causes of Y that are distinct from X. Or, in terms of our example above, interventions upon neuronal configuration H do not cause any causes of my wincing that are themselves distinct from the neuronal configuration. If neuronal configuration H is intervened upon by intervention variable I then, so the worry goes, my headache sensation will (or might) also be intervened upon. If so, then given that my headache sensation is causally correlated with my wincing, intervention variable I will fail to satisfy necessary condition (3). Intervention variable I will not be a legitimate intervention variable. And one reason, as I suggested above, for this necessary condition on legitimacy of interventions – that is, the reason this scenario is problematic – is that such correlations confound the correlations of interest (in this case, the correlation between neuronal configuration H and my wincing) **(Figure 9).**

(Fig. 9)



The intervention variable acting upon Configuration H also acts upon Headache Sensation. If the headache is a distinct variable from H, then we have a direct causal path from the intervention to Behaviour that is independent of the causal path via Configuration H.

Now, reading the condition this way does assume that a constitutively related variable, such as my headache sensation, qualifies as 'distinct' in the relevant sense. And this is a potentially contentious matter, given that some of what's at stake in disputes over the Exclusion Problem is whether, and in what sense, mental properties should be considered 'distinct' from their physical realisers. We are, in accordance with the core Exclusion principles, supposing that mental properties are metaphysically distinct from physical properties, so there may be some grounds there for thinking the relevant variables distinct in the present case. But even if

mental properties and their physical realisers are not distinct in the sense at issue in this condition, there is still a potential problem here.

The rationale for these necessary conditions on legitimate intervention variables is that interventions must isolate the causal candidate variable's correlation with the outcome variable. If, as with our example, the intervention variable for candidate cause *c* and outcome *o* also sets values for a distinct cause *c** that independently causes *o* (i.e., causes *o* via a directed path that does not pass through *c*), then we have not successfully disentangled the influence of *c* on *o* from that of *c**. So whilst there is a constitutive problem if the intervention fails to satisfy the conditions above – because these are necessary conditions for the variable to be a legitimate intervention variable – this reflects an epistemic concern: to isolate and ascertain the difference-making influence of our causal candidate *c* with respect to *o*. We want to find out whether my neuronal activity H makes a difference to my wincing; if whenever we attempt to switch on the neuronal activity we also inadvertently switch on my pain sensation, then our investigation is thwarted. This means that, even if mental properties do not qualify as distinct in the sense intended by condition (3) above, there will still be a problem if the interventionist investigator is unable to identify which variables are constitutively related – and so non-distinct – and which are not. If the causal candidate variable co-instantiates with a distinct variable under intervention, then the intervention fails condition (3); if the causal candidate variable co-instantiates with a non-distinct variable, and the investigator is unable to tell that it is non-distinct, then the investigator is unable to judge that the intervention satisfies the condition and hence unable to judge that she has successfully isolated the difference-making influence of the causal candidate.

For these reasons, we can appreciate why an Interventionist might want to rule out from the same variable set variables constitutively related to their candidate causes. If the above concerns are genuine dangers, then the Leanness principle might seem appealing in its apparent promise to circumvent them.[64] We can, at least, see that Campbell's appeal to

[64] We should note that if the second problem holds, then it seems that this would prevent the first from holding and vice versa. If pseudo-causal correlations obtain with the relevant variables, then confounding correlations do not, and vice versa. Consider M1 causing M2 and constitutively depending on P1. Intervention variable (IV) is the candidate intervention. Suppose (IV) intervenes upon M1 and inadvertently intervenes upon P1. Assume that M1 causes M2. If so, then interventions upon M1 would produce correlations between P1 and M2. So (IV) would have a causal path through to M2 that was independent of M1. So (IV) is disqualified. But if (IV) is disqualified, no causal explication of M1's relation to M2 is viable. In which case, no downward pseudo-causal correlation between M1 and P1 will show up – because M1 is not a viable causal candidate for any outcome variable, because M1 will lack any viable intervention variable for any outcome variable.

Leanness is not ad hoc; it is not motivated only by its utility in facilitating a solution to the Exclusion Problem.

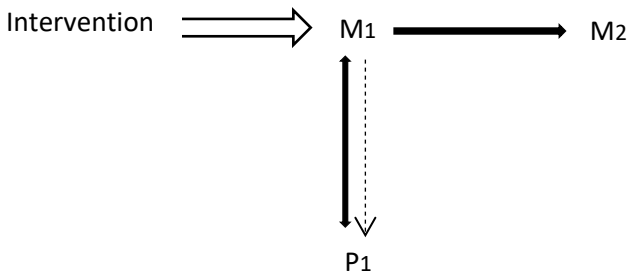
Section 2b: Top-down & Bottom-up Correlations

In this section, I examine the putative motives for Leanness in more detail. In considering how these might show up when assuming constitutive dependence relations between variables, we will see that there are four forms of potentially misleading correlation that might arise.

I will take the primary motive for Leanness first. I outlined above this motive – explicitly stated by Campbell – for the Leanness constraint: to avoid mistaking correlations between variables related by constitutive dependence for correlations between causally related variables. I now examine this putative concern in more detail. We should at this point note that misleading correlations might be generated from two directions. We might have top-down correlations produced by interventions upon causal candidate M1, or we have might have bottom-up correlations produced by interventions upon subvening P1. In the first case, the thought might be that intervening upon M1 will also intervene upon P1, hence we will find correlations between M1 and P1 such that M1 might be the cause of P1 (**Figure 10**). In the second, intervening upon P1 will also intervene upon M1, hence we will find correlations between the two such that P1 might be the cause of M1. So, broadly, there are two ways in which misleading pseudo-causal correlations between constitutively related variables might be feared.

The first problem holding also prevents the second from holding – if you had a pseudo-causal correlation between M1 and P1, and so treated M1 as cause of P1, then the intervention variable (IV) for M1 would have a causal path to M2 via P1, but this would not be a causal path independent of M1; it would be a causal path from (IV) to M2 via M1 and P1. This does not render the distinction between two kinds of problematic correlation pointless, because it might be that one obtains and the other doesn't for reasons independent of the misleading correlation that does obtain.

(Fig. 10)



Top-Down Pseudo-Causal Correlation. Interventions upon M1 produce correlations between M1 and P1.

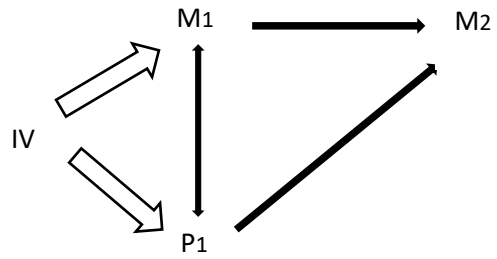
The second motive for Leanness was the worry that where M1 is constitutively dependent upon P1, and where they are included in the same variable set, candidate interventions will either fail to satisfy a constraint on legitimate interventions, or the investigator will be unable to determine whether they do. Again, this worry rests upon the potential for misleading correlations from two directions, top-down and bottom-up, but these are generated differently to the above.

The top-down correlation might be produced by attempted interventions upon P1. Consider the case in which P1 is the candidate cause and P2 the outcome variable. If P1 constitutes M1, then the concern is that interventions upon P1 reveal correlations between M1 and P2 in addition to those between P1 and P2. If M1 is distinct from P1, in the sense intended by Woodward's constraint, then the candidate intervention variable for P1 is disqualified on the grounds that there is a causal path from the intervention variable to the outcome (P2) that does not pass through P1; on the contrary, it passes through M1. If M1 is not distinct from P1 in the relevant sense, then the investigator is not in a position – assuming she does not already know that M1 is not distinct from P1 – to judge that there is no causal path between M1 and P2, independent of that between P1 and P2. The top-down worry is that M1 might be implicated, or appear possibly implicated, as independent causal variable for P2.

The bottom-up generation of the problematic correlation would result from interventions upon M1 (**Figure 11**). Here, we consider M1 as candidate causal variable for outcome variable M2. The concern is that if M1 is distinct from P1, then interventions upon M1 will yield correlations between P1 and M2 such that the interventions are related by a causal path that gets to M2 independently of M1, i.e. the path that goes through P1. In which case, the candidate intervention variable will either violate the constraint on legitimate interventions,

or the investigator will be unable to discern whether it has. The bottom-up worry is that P1 might be implicated, or appear so, as independent causal variable for M2.

(Fig. 11)



Bottom-Up Confounding Correlation. (IV) is the candidate intervention variable in respect of M1 for outcome M2. But interventions upon M1 are also interventions upon P1, and P1 is also causally correlated with M2. So (IV) has a direct causal path to M2 via P1 – a causal path that obtains independently of that running via M1.

We can now see that the Leanness principle might find motivation in worries about four potential correlations:

1) Top-down pseudo-causal correlations from M1 to P1.
2) Bottom-up pseudo-causal correlations from P1 to M1.
3) Top-down confounding correlations from M1 to P2.
4) Bottom-up confounding correlations from P1 to M2.

These forms of correlation, if genuine threats, would lead to potential misattribution of causal relations and might prevent the legitimate explication of bona fide causation through the needless ruling out of intervention variables for those causal candidates. The Leanness Principle would have important work to do.

Section 3: Objections to Pluralism

Introduction

In this section, I argue that Pluralism fails as a solution to the Exclusion Problem. This is because Pluralism depends upon the Leanness principle, and that principle has no *viable* or *warranted* application in Exclusion contexts.

Section 3a presents the argument against the viability of Leanness. I claim that Leanness is methodologically redundant because it can only be applied when causal and constitutive relations have already been distinguished. But since the distinguishing of these relations requires that the bona fide causal relations have already been explicated, there is no job for Leanness to do.

Section 3b presents the first argument against the warrant of applying Leanness in Exclusion contexts. By Exclusion contexts, I mean scenarios in which the core Exclusion principles – Supervenience, Distinctness, Causation, Completeness, Exclusion – hold and in which we assume close multiple realisability of mental properties.[65] I argue that close multiple realisability of mental properties implies that two of the four specific forms of misleading correlation that putatively motivate Leanness will not arise in Exclusion contexts. Part of the warrant for applying Leanness is thereby removed.

Section 3c goes further. Here I argue that there are no grounds for worrying about any of the misleading correlations that might motivate Leanness. At the heart of the worries that potentially motivate Leanness is the assumed constitutive dependence of specific mental properties upon specific physical properties. My arguments turn upon two points. I first show that, given the limits of Completeness and on a plausible formulation of supervenience, there is no reason to suppose that such specific constitutive relations obtain. I then argue that such relations are inherently implausible.

[65] I assume close multiple realizability of m-properties because this was, in earlier chapters, part of my argument against the entailment of systematic overdetermination by the core Exclusion principles. As previously explained, it is a reasonable assumption on behalf of non-reductive physicalists, given the typical prominence afforded multiple realizability in motivating commitment to the distinctness of mental and physical properties.

On the basis of these arguments, I conclude that Leanness has no viable or warranted application in Exclusion contexts, and that Pluralism fails to provide a cogent solution to the Exclusion Problem.

Section 3a: Methodological Redundancy of Leanness

The Leanness Principle is methodologically redundant. As I explain in the sections immediately below, this is because the principle cannot be applied prior to ascertaining causal relations. Given its supposed methodological priority, this renders the principle redundant. The only context in which it could be applied would be one in which there were no call for it, for we would already have identified the bona fide causal correlations in contrast to correlations that obtain on the basis of constitution. We would, in other words, have successfully sorted the causal from the constitutive. Given the pragmatic justification for Leanness, it follows by the interventionist's own lights that Leanness cannot be a methodological constraint on articulating causal correlations. If so, then this central plank of Pluralism is removed.

To appreciate the redundancy problem, we should again consider the implications of setting up interventions on candidate causal variables. When considering it as a methodological principle, we ask: assuming only that M1 supervenes upon *some* p-property (ies), would an investigator be able to distinguish between p-properties that cause M1, and p-properties that subvene M1?

Whether for interventions upon M1 or upon P1, the investigator will be unable to confidently distinguish causal from constitutive relations. And as we will see, the reasons are similar to the reasons discussed in Section 2 for why misleading correlations might arise when intervening upon M1 or P1.

I start by considering the case where we intervene upon M1. We again assume a simple case where M1 is a variable with a binary range of possible values, 0 and 1. Let us imagine that an interventionist investigator suspects that M1 is a causal variable for outcome variable P1. Accordingly, she proceeds to intervene upon M1 by 'switching off' confounding variables (here we assume that this is itself feasible) and assigning it a value of 0. In so doing, she

observes that variable P1 also takes a value of 0. She then intervenes upon M1 again, assigning it a value of 1, and now observes that P1 takes a value of 1. The trouble is that this result is consistent with both M1's being a cause of P1 and M1's being constituted by P1. After all, if M1 is constituted by P1, then any case in which M1's value is set to 0 will be a case in which P1's is also 0. And whilst it is true that M1 may be multiply realisable and that therefore there may be cases in which M1's value is set to 1 and P1's 0, there is will only aid the investigator in ruling out M1 as cause of P1 if she observes a sufficient number of such cases. M1's being a causal variable for P1 is consistent with cases in which P1 is caused by alternative m-properties or some other kind of property entirely.

We obtain a similar result when considering the scenario in which P1 is suspected as causal variable for outcome variable M1. Interventions upon P1 assign it a value of 1, and we find that M1 takes a value of 1 also. This is consistent both with P1 causing, and with P1 constituting, M1. Again, multiple realisability of M1 would mean that there are some cases in which we assign P1 a value of 0 and M1 has a value of 1 – but this is also consistent with both P1's causing and P1's constituting M1, since P1's being a causal variable for M1 is consistent with cases in which some other p-property or another kind of property causes M1 instead.

If the above holds, then the Leanness principle is redundant in the sense that it is practically inapplicable: we are not in the requisite position of being able to distinguish between causal and constitutive relations.

Given this, if we are required to select our variables for the relevant variable set – that is, apply the Leanness Principle – *prior* to causal investigation, then of course we will be unable to do so. For it is only via causal investigation that we could, in principle if not in practice, begin to sort the constitutive from the causal.

The only alternative would be to use a supervenience thesis in conjunction with Leanness to motivate wholesale rejection of variable classes from each other's variable sets. So on the assumption of a global formulation, whereby m-properties supervene upon p-properties globally, we rule out mental properties from inclusion in variable sets for physical candidate causes and vice versa. But this has the theoretical disadvantage of also ruling out, a priori, any inter-level causal relations. Not only do cases such as Hadron Collider suggest diagonal causal relations, to prohibit inter-level causation on the basis of a prior commitment to Leanness

would seem arbitrary. Leanness is a methodological principle, intended to facilitate clarification of causal correlations; it should not prohibit such relations between whole classes of variable from the start.

I have thus far argued that Leanness is practically inapplicable because it is reasonable to think that an investigator would be unable to confidently distinguish between causal and constitutive relations between m- and p-properties. We should also note that if this is true, then Leanness cannot be a constraint upon causal relations, even by the lights of Interventionism itself. Woodward (2003) contends that an interventionist account of causation can persuasively explain the motivation for our causal investigations by locating them within a practical concern for which aspects of the world are to be manipulated to achieve a given outcome. If this is so, then any principle that is inapplicable from a practical point of view is unsuitable as a constraint on causation. Whilst Interventionism does not need to claim that all causal relations involve causal variables that are practically amenable to intervention, the notion of causation is nonetheless rooted in such practical manipulability.[66] If Leanness were a constraint on causation, then such practical manipulation would be impossible, since we would be unable to sort the permitted variables from the prohibited in our sets relative to which candidate causes are specified.

On the basis of the foregoing, the Leanness Principle is not available for deployment in an Interventionist solution to the Exclusion Problem. Campbell's solution was to invoke Leanness in excluding causal candidate P1 for effect M2 from the variable set relative to which M1 causes M2, and excluding causal candidate M1 for effect M2 from the variable set relative to which P1 causes M2. Having achieved this separation of variable sets for ostensibly competing causal variables, the solution then defines specious overdetermination in terms also relative to variable sets. So cause *c* and cause *c** can only overdetermine effect *e* if *c* and *c** if both are causes of *e* relative to the same variable set. Given that, assuming Leanness, our candidate M1 cause for M2 cannot belong to the same variable set as candidate cause P1, it follows that M1 and P1 cannot causally overdetermine M2. Leanness is crucial to this line of argument,

[66] This is intended in two senses. First, our concept of causation reflects our practical concerns with manipulating features of the world. Second, such manipulability provides the ideal case against which particular judgements of causation are to be judged. On this view, a sufficient condition for causation is that, under *ideal circumstances* the causal variable, related by interventionist counterfactual to the outcome variable, would be such that interventions upon it *would* result in the changes in the outcome variable described by the counterfactual.

and so putting it to one side leaves Campbell's solution without motivation. Without Leanness, his account offers no principled reason for excluding M1 and P1 from each other's variable set, relative to which they are causal candidates for M2.

Section 3b: Objections to Pluralism from Close Multiple Realisability

I have argued that, taken as a methodological principle, Leanness is redundant. But my rejection of Leanness does not rest solely on that claim. For even if it were not generally redundant in the manner described above, Leanness would not be justifiably applied to causation in Exclusion contexts.

In Section 2b, I outlined four ways in which the worries that might ostensibly motivate Leanness might manifest. Top-down and bottom-up *pseudo-causal correlations* might be generated by interventions upon M1 and P1, respectively. In addition, top-down and bottom-up *confounding correlations* might be generated by interventions upon P1 and M1, respectively. The former kind of misleading correlation are taken to threaten false judgements of causal relations between constitutively related variables; the latter kind is taken to either violate a constraint on legitimate intervention variables (if M1 is relevantly distinct from P1) or present the interventionist with a case that cannot be ruled out as not violating that constraint (if M1 is not so distinct).

However, whatever their status in other contexts, these worries will not arise in contexts of causation implied by the core Exclusion principles together with close multiple realisability of m-properties. If Leanness, as a methodological principle, is motivated – and so its application warranted – by the potential for these misleading correlations, then it is not warranted in contexts where no such correlations will obtain.

The arguments in support of my claim fall into two broad camps. We have arguments against the alleged potential correlations generated by interventions upon M1, and arguments against those generated by interventions upon P1. Due to our taxonomy of potential worries, this means that our arguments in respect of interventions upon M1 will address top-down pseudo-causal correlations and bottom-up confounding correlations. Our arguments in respect of interventions upon P1 will address bottom-up pseudo-causal, and top-down confounding, correlations.

In the section below, we will find that close multiple realisability of m-properties prevents misleading correlations arising on the basis of interventions upon M1. But we will also see that such multiple realisability will not prevent correlations of those kinds on the basis of interventions upon P1. The immediate upshot will be that insofar as applying Leanness to Exclusion contexts would be warranted by the need to preclude misleading correlations produced by interventions upon M1, it is not in fact warranted. However, similar considerations do not seem sufficient to reach the same conclusion in respect of potential correlations produced by intervening upon P1. I address those in Section 3c.

Correlations produced by interventions upon M1

Exclusion contexts are cases implied by the core Exclusion principles:

Distinctness: Mental properties are metaphysically distinct from physical properties.

Supervenience: Mental properties supervene upon physical properties.

Causation: Mental properties have – in their own right qua mental properties – genuine causal efficacy with respect to other folk-domain properties.

Completeness: All physical effects have sufficient physical causes.

Exclusion: No single effect can have more than one sufficient cause occurring at any given time—unless it is a genuine case of causal overdetermination.

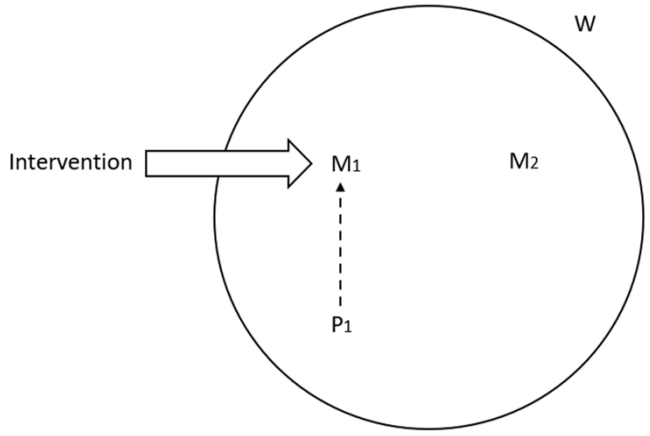
In order to see, given our commitments from previous chapters, what these principles imply, we need to supplement and refine them. In particular, for our present concerns, we need to add our principle of close multiple realisability for this principle will be integral to our arguments concerning putative worries arising from interventions upon M1:

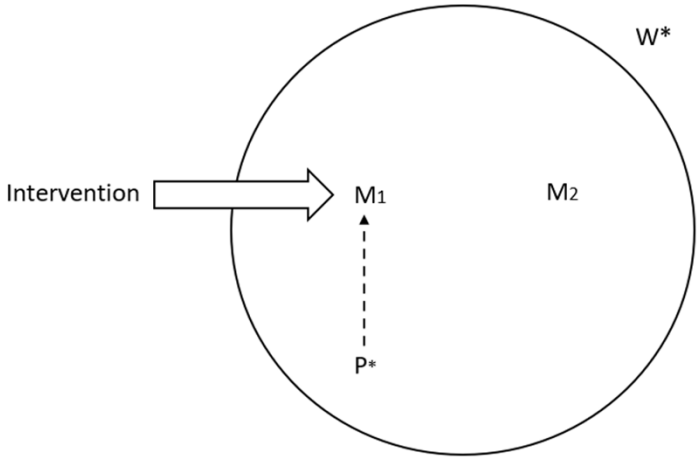
Close Multiple Realisability: Mental properties are multiply realisable at close possible worlds.

We start by considering the question of top-down pseudo-causal correlations between M1 and P1. What we want to see is whether, in Exclusion contexts, interventions upon M1 as causal candidate for outcome M2 will plausibly generate correlations between M1 and its subvening P1. To put it roughly, we need to consider whether interventions upon M1 will also function as interventions upon P1, hence setting up a correlation between the values assigned to M1 an the values assigned to P1 (see Figure 12).

For the sake of simplicity, we take the relevant variable set V to include only M1, M2 and P1. So V = {M1, M2, P1}. For the same reason, we take our range of values to be 0 and 1. Intuitively, if M1 instantiates, it takes a value of 1; if it does not, then it takes a value of 0. Given the close multiple realisability of M1, it is reasonable to deny that interventions upon M1 will set up pseudo-causal correlations between M1 and P1. Suppose we intervene upon M1, assigning it a value of 1: M1 instantiates. But now suppose that M1 is multiply realised at close worlds, so that if P1 co-instantiates with M1 at the actual world, it does not do so at a close world w* because at w*, M1 is realised (in our parlance, 'constituted') by P* instead. It is reasonable to think that interventions assigning values of 1 to M1 across cases will not thereby assign values of 1 to P1; if so, then it reasonable to deny that interventions upon M1 will yield pseudo-causal correlations between M1 and P1.

(Fig. 12)





- - - - - - -> = constitution/realisation

This is not yet a conclusive consideration, since close multiple realisability of M1 is strictly compatible with P1's co-instantiation across a wide range of intervention cases. Just because M1 could have been realised by P* does not entail that intervention cases will not be ones in which P1 in fact instantiates when M1 does. P1 might be instantiated under interventions upon M1 in enough cases to perhaps warrant a (mistaken) judgement of causal correlation between M1 and P1, despite there being possible cases in which P1 is not so instantiated. But it does provide reason to block the entailment of such pseudo-causal correlations from the fact that M1 constitutively depends upon P1. (We will see below, in Section 3c, that there are other reasons to shore up our claim that no such correlations will arise in Exclusion contexts.)

The same line of thought will apply to the prospects for bottom-up confounding correlations when intervening upon M1. Here the worry was that if we intervene upon M1 as causal candidate for M2, and if M1 is constituted by P1 (and distinct from P1), the candidate intervention variable will have a causal path to M2 that is independent of M1: it will have a path to M2 via P1. Such confounding correlations are sufficient, under Woodward's necessary conditions on intervention variables, to rule out the candidate intervention variable as viable for explication of M1's correlation with M2.

Assume again an endogenous variable set V, with M1, M2, and P1 as members. Here we also need a candidate intervention variable (IV). If we suppose that M1 is multiply realised across close worlds, then there is no reason to assume that interventions assigning a value of 1 to

M1 will also assign that value to P1. At close worlds, M1 co-instantiates with P*, and the value of P1 is 0. If so, then there will be cases in which variable (IV) assigns to M1 a value of 1, M2 takes a value of 1 as a result of its causal correlation with M1, and P1 takes a value of 0. The potentially worrisome confounding correlation between P1 and M2 does not obtain.

Correlations produced by interventions upon P1

We cannot however use close multiple realisability of m-properties to preclude all worries in support of Leanness. I now consider bottom-up pseudo-causal correlations from P1 to M1. Here we take P1 to be the candidate causal variable for outcome variable P2. Again, M1 supervenes upon P1, and is closely multiply realisable. However, the situation when thinking about the potential for these kinds of correlations is significantly different. Whereas close multiple realisability plays an important role in showing that top-down worries are not entailed by constitutive dependence of M1 upon P1, it will not help alleviate the concern regarding bottom-up pseudo-causal correlations from P1 to M1.

The difference results from the asymmetric necessitation of M1 by P1, as implied by M1's supervening upon P1.[67] This entails asymmetric co-instantiation of P1 and M1, so if we are thinking in terms of binary ranges of associated values for P1 and M1 – i.e. values of just 0 and 1, corresponding to instantiation or otherwise – then every case in which P1's value is set to 1 will be a case where M1's value is also 1. If so, then it seems plausible that interventions on P1 will also assign values to M1 across cases resulting in pseudo-causal correlations.

Close multiple realisability of M1 will not foreclose the prospect of bottom-up pseudo-causal correlations here, because if there are cases in which M1 is instantiated (i.e. takes a value of 1) but where P1 is not (takes a value of 0), this is consistent with P1's causing of M1 and consistent with P1's constituting M1. It is commonplace to acknowledge that an effect

[67] As cited in previous chapters, supervenience is typically taken to imply necessitation of supervening properties by their supervenience bases. P1 asymmetrically necessitates M1 so that the following holds, necessarily: [if P1 then M1] but the following does not: necessarily: [if M1 then P1]. For present purposes, it does not matter whether we treat the necessity as physical or metaphysical. If the modal strength of supervenience is physical, then it is still the case that P1's occurrence entails M1's occurrence across intervention cases.

can have a range of possible causes; if M1 takes a value of 1 when P1 takes a value of 0 – but where every case of P1 taking value 1 is also a case of M1 taking value 1 – then this is consistent with M1's sometimes being caused by P1 and sometimes being caused by some other variable. So just as asymmetric necessitation of M1 by P1 would ostensibly produce observed correlations that are ambiguous between P1's causing and P1's constituting M1, close multiple realisability does nothing to ameliorate this ambiguity.

The situation is ultimately similar for top-down confounding correlations. Here, close multiple realisability *might appear* to fend off the threat. But we have reason to judge that it will not. To see this, let's first consider how one might come to worry about these correlations when intervening upon P1 as causal candidate for P2.

Given asymmetric determination of M1 by P1, it appears that every intervention upon P1 will also be an intervention upon M1. As before, we assume an endogenous variable set V comprising P1, P2, and M1, and an exogenous candidate intervention variable (IV). Each endogenous variable has a range of two values, 0 and 1. If we use (IV) to intervene upon P1, assigning it a value of 1 then it seems that we will also assign that value to M1. M1, after all, is necessitated by P1 so any instantiation of P1 implies instantiation of M1. Given that P1 causes P2, assigning a value of 1 to P1 will result in P2 taking a value of 1. If so, then we have P1, M1 and P2 all taking a value of 1. And whenever we use (IV) to assign that value to P1, we will have the same situation. On this basis, it seems plausible that the candidate intervention variable (IV) has a causal path to P2 that by-passes P1: (IV) has a path to P2 that runs via M1. (Or, if M1 is not distinct from P1 in the sense intended by Woodward's constraint on interventions, it seems plausible that the investigator will not be in a position to rule out such an independent path.) If so, then we have a top-down confounding correlation between M1 and P2.

However, it might yet seem that close multiple realisability would mitigate the concern. For if M1 is multiply realised at close worlds, then there are cases in which (IV) sets the value of P1 to 0 and the value of M1 is 1. In those cases, M1 is realised not by P1 but by P*. Now we are assuming a causal correlation between P1 and P2; such a case would therefore also include P2 with a value of 0. So we have cases in which the values of P1 and P2 are both 0, and the value of M1 is 1. In those cases, there is no (appearance of a) causal path from (IV) to

P2 via M1, since the values of M1 and P2 are not aligned. The question is, are these cases sufficient to prevent the confounding correlation between (IV) and P2?

We have reason to think not. To see this, we must bear in mind that the interventionist is only here concerned with the viability of (IV) as intervention variable for P1 with respect to outcome P2. She is not interested in whatever other intervention variables might be appropriate for M1 with respect to P2.[68] For this reason, the close multiple realisability of M1 will plausibly *not* count against the confounding correlations that concern us when considering (IV)'s viability.

If M1 is closely multiply realisable, then there are cases in which M1 takes a value of 1 and P1 takes a value 0. But these cases are not pertinent to the question of (IV)'s viability for P1 with respect to P2. This is for two reasons. First, we already know that (IV)'s assignation of 1 to P1 will always likewise assign a value of 1 to M1. Second, there are reasons to think that (IV) cannot both assign a value of 0 to P1 and a value of 1 to M1 in the same case. Suppose (IV) itself has a range of two possible – and mutually exclusive – values: 0 and 1. If (IV) has a value of 1, then it assigns a value of 1 to P1; if (IV) has a value of 0, then it assigns a value of 0. It cannot therefore take both values at once, since P1 cannot both instantiate and fail to instantiate in the same case. Now, we know that whenever P1 takes a value of 1, M1 also takes that value. It is therefore reasonable to think that (IV) assigns a value of 1 to M1 when it takes a value of 1 itself. We are assuming that when (IV) takes a value of 1, P1 also does, and when (IV) takes a value of 0, P1 takes that value. If the above is right, then (IV) cannot assign a value of 0 to P1 and a value of 1 to M1 in the same case. Because to do so, (IV) would have to be capable of taking the value of 0 and 1 simultaneously, which it cannot do.

[68] The interventionist therefore need *not* be concerned with establishing whether (IV) has a causal path to P2 via M1, independently of P1, by comparing cases of the following kind. In some cases, (IV) is correlated with P1, P2 and M1 – the cases in which (IV) assigns a value of 1 to P1. In multiply realised cases, (IV) – having assigned a value of 0 to P1 – is not correlated with the value of M1. If we were able to say that P2, in some cases where (IV) has set the value of P1 to 0 and in which M1 has a value of 1, has a value of 0, then we would have evidence against M1's causing P2 independently of P1; and if so, then we could potentially rule out (IV) from having an independent causal path to P2 via M1. There would be cases in which (IV) assigns a value of 0 to P1, P2 also takes a value of 0, and M1 takes a value of 1. Comparing this to the case in which (IV) assigns a value of 1 to P1, P2 takes the same value, and M1 does too, would enable the investigator to judge that (IV) has a causal path to P2 via P1 but not via M1. But we are not in that position: we do not know whether those cases where P1 is set to 0 and M1 has a value of 1 are cases in which P2 takes a value of 0. We are not entitled to assume that it does on the basis of P1 having that value, because that would be tantamount to P2's having *only* P1 as a causal variable, and there are no grounds for that assumption.

Other intervention variables might be capable of switching on M1 and switching off P1 in the same case; (IV) is not one such variable. If so, then the cases implied by close multiple realisability of M1 are irrelevant to the question of (IV)'s legitimacy as intervention variable for P1.

Summary

Thus far we have argued that two of the four specific forms of misleading correlation that might motivate, and warrant, application of Leanness do not arise in Exclusion contexts. The two that, due to close multiple realisability, will *not* arise are:

(i) Top-down pseudo-causal correlations from M1 to P1.

(ii) Bottom-up confounding correlations from P1 to M2.

Regarding the potential for top-down pseudo-causal correlations, M1's close multiple realisability implies that values of P1 will *not* correlate with those assigned under intervention upon M1. In light of this, there will be no bottom-up confounding correlation between the intervention variable (IV) (for M1) and M2 via P1. Therefore, Leanness cannot be justified by the need to preclude such correlations in Exclusion contexts.

Matters look different for the two other forms of correlation:

(iii) Bottom-up pseudo-causal correlations from P1 to M1.

(iv) Top-down confounding correlations from M1 to P2.

Close multiple realisability of M1 will not preclude the possibility of these. However, this need not deter us. As I argue below, we can make more fundamental objections to Leanness by appealing to Completeness and Supervenience. These will undercut all forms of worrisome correlation, and hence show that Leanness finds no warrant in Exclusion contexts.

Section 3c: Objections to Pluralism from Completeness & Supervenience

We have now seen that the close multiple realizability of M1 provides grounds for rejecting *some* of the worries that putatively motivate Leanness. Insofar as it does, and insofar as close multiple realizability is an assumption of Exclusion contexts, an application of Leanness is not warranted in those contexts.

But we also saw that some worries remain, because close multiple realizability of M1 is not capable of defusing the threat of either pseudo-causal correlations or confounding correlations produced by interventions upon P1. In this section, I therefore marshal arguments against those threats in Exclusion contexts by appealing not to close multiple realizability, but rather to other principles included in Exclusion set: Completeness and Supervenience.

The central target of my arguments will be the following Pluralist assumption: that if m-properties supervene upon p-properties, then the specific p-properties in Exclusion contexts are constitutively related to the specific m-properties implicated in those contexts. The constitutive relation between some p-properties and m-properties thus underpins an asymmetric determination relation that is integral to worries about misleading correlations between P1 and M1 (bottom-up pseudo-causal), and between M1 and P2 (top-down independent). But as I argue here, such worries do not arise in Exclusion contexts, for two main reasons. First, contrary to Campbell's assumption, the principles that form those contexts do *not* entail the kind of constitution relations that are required for those worries to get going. Second, correlations, constitutive or causal, between the relevant p- and m-properties in these contexts are inherently implausible. This being so, the application of Leanness would not be warranted in such contexts, and so the Pluralist solution to Exclusion would fail. I say 'would' because, as claimed in Section 3a, the Leanness principle is anyway methodologically redundant. But in addition to confirming the failure of Pluralism as a response to Exclusion, the arguments in the present section will also offer a positive message. Since the worries that might have recommended the application of Leanness do not arise in Exclusion contexts, the methodological redundancy of Leanness in those contexts need not prevent the Interventionist from explicating mental causal relations on their terms.

We want to see how a central assumption of Pluralism and of its motivation – that if M1 is constituted by P1, then P1 necessitates M1 – is unwarranted in light of the core Exclusion principles. We can approach this by first considering what variable P might be, given certain constraints imposed by these principles. We should recall that these are:

Distinctness: Mental properties are metaphysically distinct from physical properties.

Supervenience: Mental properties supervene upon physical properties.

Causation: Mental properties have – in their own right qua mental properties – genuine causal efficacy with respect to other folk-domain properties.

Completeness: All physical effects have sufficient physical causes.

Exclusion: No single effect can have more than one sufficient cause occurring at any given time—unless it is a genuine case of causal overdetermination.

The key principle here will be Completeness. Given Completeness, we should treat variable P as physical, since there are no reasons to think that the higher-level special sciences are complete in this sense. Indeed, there are well-known reasons for thinking that some of these, e.g. neurophysiology cannot be.[69] So insofar as we assume Completeness when considering the motivations and prospects for Leanness, we assume also that variable P corresponds to a physical property, that is, a property individuated by the theories and laws of physics.[70] [71]

[69] The causal generalisations of neuroscience are *ceteris paribus*. The obtaining of any causal relation at those levels will require background contexts. A causal relation that often obtains might fail to do so if, for instance, the physical environment of the subject is sufficiently different, e.g. an earthquake, or if other physiological factors are different, e.g. the subject is undergoing a heart attack.

[70] An alternative is to offer a relation by which instantiations of higher-level properties are necessarily implicated by instantiations of physical properties, such that every physical cause necessarily co-instantiates with some higher level property M (where M is a constant). In this way, one might secure a form of Completeness for higher level properties, too. Those properties would be, as it were, piggy-backing on the physical properties implicated by Completeness. However, to assume such a relation would be highly contentious. It would require either an implausibly local form of supervenience, or type identity relations between higher level and physical properties, or so-called bridge laws connecting tokens of higher level properties with those of physical ones. The first option is untenable; the second is here off the table since we are interested in non-reductive responses to Exclusion; and the third is problematic, if assuming multiple realizability of higher level properties (see Fodor, 1974).

[71] If, on the other hand, we drop Completeness then we achieve a swift solution to the Exclusion Problem, for without it, there will be no apparent implication of overdetermination by the remaining core propositions. However, though swift, such a solution is not theoretically cheap: Completeness enjoys widespread support across a variety of metaphysical perspectives, as well as from the scientific community. And those who reject it

So on grounds of Completeness, we treat variable P1 as corresponding to a physical property. (Variable P1 is just that variable taken above to be that which serves as constitutive of variable M1 for the purposes of considering motivations for Leanness.) Now, if P1 is physical, then it follows that:

(i) a plausible mental-physical supervenience relation does not imply misleading pseudo-causal or confounding correlations under interventions upon P1 and

(ii) such correlations – whether constitutive or causal – are inherently implausible.

(i) Plausible supervenience does not imply misleading P1 correlations

We'll start with the first implication: that a plausible mental-physical supervenience thesis will not imply misleading correlations under interventions upon P1. In the discussion above, I took supervenience to be formulated globally. On this formulation, the global distribution of mental properties supervenes upon that of the physical properties, entailing that any minimal physical duplicate of the actual world is a mental duplicate of the actual world.

As per my earlier exposition of the putative worries motivating Leanness, there are two forms of correlation that concern us here. First, bottom-up pseudo-causal correlations between P1 and M1, and second, top-down confounding correlations between M1 and P2. Let's now consider the prospects of these correlations manifesting under interventions upon P1, given a global supervenience thesis and our assumption that P1 is a physical variable.

If P1 is physical and supervenience is global, then there is no reason to expect interventions upon P1 to yield bottom-up pseudo-causal correlations between it and M1. This is because a global thesis only entails that interventions upon the *total set* of physical properties would correlate with changes in the *total set* of mental properties. But we are concerned not with the total sets of these properties but rather, specific physical P1 and specific psychological M1. Given that P1 is supposed, by the proponents of the Exclusion Problem, to be causally sufficient for P2, it must be (or include) a specific physical variable individuated by the theories / laws of physics. At any rate, it is not taken to be the complete set of physical properties.

– i.e. metaphysical emergentists – incur the responsibility of explaining and substantiating irreducible psychophysical laws governing causal relations between causally autonomous, irreducible properties and their physical effects. For these reasons, we have been assuming Completeness throughout.

Equally, M1 is supposed to be a specific psychological variable, not the complete set of such variables.

Nor can we infer anything of relevance here about P1 and M1 as specific variables from the global thesis, since changes in the total set of physical base properties might leave P1 unchanged, and changes in the total set of mental properties might leave M1 unchanged. This is logically consistent with the entailment outlined above. If so, then we can infer nothing about the status of M1 under interventions upon P1: global supervenience is consistent with interventions upon P1, assigning values of 1 or 0, having no impact on the value of M1. We thus have no reason to think that such interventions would produce bottom-up pseudo-causal correlations between P1 and M1, when taken as specific variables individuated by, respectively, the theories of physics and the theories of folk- and scientific psychology. Global supervenience is silent on the matter.

Regarding the second form of misleading correlation ostensibly generated by interventions upon P1, we obtain a similar result for similar reasons. Here, the concern was that interventions upon P1 would produce correlations between M1 and P2. The thought was that if M1 is constituted by P1, then P1 necessitates M1. If so, then any intervention assigning a value of 1 to P1 would also assign a value of 1 to M1. Now, assuming a causal correlation between P1 and P2, we would then have cases under intervention upon P1 where P1, P2 and M1 all have a value of 1. In which case, we have a top-down confounding correlation between M1 and P2.[72] However, since a global supervenience thesis does not entail that changes in the value of physical P1 must be accompanied by changes in the value of psychological M1, or vice versa, it offers no reason in favour of supposing interventions upon P1 to produce correlations between the values of M1 and P2. This worry was based upon the assumption that, if M1 is constituted by P1 then P1 necessitates M1. But my point here is just that global

[72] Such a correlation produced by interventions upon P1 would undermine the viability of those interventions; given Woodward's constraints on intervention variables, this independent correlation between M1 and P2 would rule out the candidate intervention variable responsible. According to that constraint, no intervention variable (IV) for causal candidate *c* with respect to outcome *o* can have a causal path to *o* that does not pass through *c.* In the present case, this means that (IV) cannot have a causal path to P2 that does not pass through P1. So if interventions upon P1 with respect to P2 did produce an independent correlation between M1 and P2, this constraint would be violated. (Or, if M1 were not distinct from P1, but the investigator did not know this, she would not be in a position to judge whether (IV) were in violation of the constraint or not. This would still be problematic.)

supervenience of m-properties upon p-properties will *not* warrant the claim that M1, as specific psychological variable, *is* constituted by P1, as specific physical variable.

So much for global supervenience. What are the prospects for endorsement of a more local thesis, such that m-properties and physical p-properties are plausibly constitutively correlated?

I do not think the prospects are promising. We have encountered the relevant problems before, in Chapter 2, Section 4b. There, we saw that applying the Determination principle in the Determination model of Exclusion required a *radically local* supervenient thesis. The same is true here: to set up constitutive correlations between specific physical and psychological variables, we require a thesis that entails necessitation of psychological properties by *specific* physical properties. But adopting this form of radically local supervenience involves a host of further substantive, controversial assumptions.

To briefly rehearse the points made in Chapter 2, Section 4b, these assumptions were as follows. To endorse radical local supervenience, one must deny:

(a) Mild content externalism
(b) Any form of causal theory of content

Previously, we saw that mild content externalism is the view that some content of some mental states is broad. Content that is broad is not wholly determined by intrinsic physical facts about the individual who instantiates those states. A radical local supervenience thesis rules this out because psychological properties are determined by specific physical properties, the instantiation of which would qualify as the obtaining of intrinsic physical facts about the relevant individual. But as we saw, there are plausible reasons to take mild content externalism seriously as an option.

Radical local supervenience also required ruling out any form of causal theory of content. Given that local necessitation of psychological properties by specific physical properties is synchronic, it follows that instantiation of physical P at time *t* is sufficient for instantiation of psychological M at time *t*. This view is incompatible with any view on which the content, and so constitution, of M depends upon the causal history of the subject instantiating M.

We also saw that radical local supervenience requires assumptions about standing states, such as intending. For if such standing states constitutively depend upon other psychological states, and those other background states themselves supervene upon specific physical properties, then it seems that radical local supervenience might require that the physical base property is conjunctive. That is, the physical base P for the standing state M needs to 'pack in' those subvening properties underlying the constitutive background conditions for M. A complex conjunctive P property seems to require that every conjunct has (confers) distinct causal dispositions because otherwise it's not clear in what respect they are *distinct* conjuncts. But in order to be conjuncts of the single conjunctive P, every conjunct would need to confer these distinct causal dispositions under the *same* laws. These are substantive assumptions.

For these reasons, I take radical local supervenience to be unappealing. One *could* bite the bullet and adopt it anyway; but it's not clear why one would want to. For, as we see in the next subsection, the thesis lacks positive motivation. The same considerations will also further support our central contention in this section: that there are no grounds for worrying about misleading P1 correlations in Exclusion contexts.

(ii) Misleading P1 correlations are inherently implausible

We turn now to the second implication of treating P as physical and supervenience as global: misleading correlations arising from interventions upon P_1 are inherently implausible.

The rough overall argument here is this. In Exclusion contexts, P_1 corresponds to a specific physical property; neither science nor metaphysics provide compelling grounds for taking specific physical properties as constitutive of psychological properties. But a minimal condition for either bottom-up pseudo-causal or top-down confounding correlations is that there obtains a constitutive correlation between P_1 and a specific psychological property.[73]

[73] This is because both forms of misleading correlation stem from a constitutive correlation between the p- and m-variables. In the case of pseudo-causal correlations, these arise on the basis of constitutive correlations and are mistakenly taken as causal correlations. In the case of confounding correlations, these arise again because of unacknowledged constitutive correlation. Because of the constitutive correlation between P and M,

So it follows that, without grounds for taking P_1 as constitutively related to psychological property M_1, there are no grounds for worrying about misleading pseudo-causal or confounding correlations that depend upon that constitutive relation.

There are two main upshots. In the preceding subsection, we pointed out that the proponent of radical local supervenience must provide cogent grounds for the thesis that are independent of the content internalism / externalism dispute, because internalism does not entail radical local supervenience. If there are no compelling scientific or metaphysical reasons for the non-reductive physicalist to hold that specific physical properties are constitutively correlated with psychological properties, then this also counts against radical local supervenience because that thesis entails such correlations. So our arguments below will shore up the points made in the preceding subsection against the plausibility of radical local supervenience. The second, and most important, upshot will be that there are no compelling grounds for worrying about the kinds of misleading correlation that supposedly motivate the Leanness principle, and hence motivate Pluralism.

Science: Physics and Psychology

Neither physics nor psychology furnish us with grounds to worry about constitutive relations between physical and psychological properties. Neither discipline formulates theories, descriptions, laws or generalisations that cover properties of the other. No law of physics explicitly connects physical properties with psychological properties; no generalisation in psychology explicitly connects psychological and physical properties. The respective theories of those scientific domains do not imply – absent extra-scientific theses – constitutive relations between the relevant properties. The properties of physics are absent from psychological theories and laws; the properties of psychology are absent from the laws and theories of physics. Neither field gives us reason to suspect constitutive correlations, under intervention, holding between its proprietary properties and those of the other domain.

interventions upon P are also interventions upon M. As a result, there is a confounding correlation running from the intervention variable to the effect of P via M.

This absence is especially important in respect of the physical laws. For if those laws are exceptionless, then were they to include psychological properties, Completeness might tempt us to postulate constitutive correlations between some physical and psychological properties. Suppose physics told us of an exceptionless law that implied a nomologically regular causal relation whereby a psychological property *M* caused a physical effect *E*. The occurrence of the psychological property nomologically necessitates an occurrence of the physical effect. Given that psychological laws are *ceteris paribus* generalisations, it seems no psychological law can explain the relation between *M* and *E*. And given that Completeness dictates that the physical effect has a sufficient physical cause (call it *P*), it would be tempting to explain the necessity of the causal relation by identifying the psychological cause, *C*, with the physical cause, *P*. In such a case, we would have grounds for a causal correlation, between *M* and *E*, *and* potentially a constitutive correlation, between P and M. But no such cases arise from psychological or physical laws in their present form.

Indeed, we have reason to suppose that physics would not countenance any such theories / laws. Physics is committed to generality of explanation. It is unconcerned with phenomena (e.g. psychological) that might be related to those within its purview in a non-systematic way. The experimental and observational practices of physicists presuppose that the psychological can impact the physical; the intentions, beliefs and desires of the scientist all play a role in directing measurement and analysis, for example. But physicists are unperturbed by this. A plausible explanation for the lack of concern is that psychological phenomena are not systematically related to the physical. Physicists do not take causal relations between the psychological and the physical to be a problem for their project because they are looking for relations that hold more generally. They are interested in relations that hold for all instantiations of certain properties within certain background contexts, rather than for just some, as is the case with psychological causes. Furthermore, a methodological commitment to the causal completeness of the physical also plausibly informs the sanguine attitudes of physicists towards apparent psychological causes of physical effects.

Historical evidence of this attitude can be found in physicists' response to counterevidence against determinism. The response to evidence suggesting that physical effects are not wholly determined by their causes – that is, not necessitated by those causes given the requisite background conditions – has been not to countenance causal factors from *outside* the

purview of physics or to drop commitment to exceptionless laws but rather to embrace a probabilistic notion of causation or lawhood. In other words, the response has been to stick with the causal phenomena as already defined and redefine the notion of causation or lawhood at stake. Here we see an illustration of physicists' commitments both to generality and to the potential of physical theories to sufficiently explain the physical domain, i.e. to Completeness.

If physicists are happy to overlook psychological causes for these reasons, then it is unlikely that they will figure in any physical laws or theories. Without theories or laws that connect physical and psychological properties, there are no scientific grounds for positing constitutive correlations between them.

Metaphysics: Supervenience, Completeness, Identity & Bridge Laws

I also see no compelling metaphysical grounds for suspecting constitutive correlations between physical and psychological properties.

We have already argued supervenience does not itself entail constitutive correlations between specific physical and specific psychological properties, unless formulated as a radically local thesis. But a radically local thesis is implausible, so supervenience does not imply constitutive correlations.

Completeness alone says nothing to suggest constitutive relations between levels. It is only concerned with physical effects and physical causes. However, it has been used in conjunction with other claims to establish constitutive correlations between the physical and psychological domains via type identity (Lewis, 1966).

But this kind of appeal to Completeness need not move the non-reductive physicalist: she will reject the identification of psychological properties and physical properties. She will endorse Distinctness.[74]

[74] We are especially concerned with the position of the non-reductive physicalist who endorses each principle of the Exclusion conjunction. Distinctness is one such principle. But, to reiterate a point made elsewhere, the non-reductive physicalist will typically endorse Distinctness, at least in part, because she accepts multiple realisability of mental properties. So both the Distinctness principle itself, and its motivation, count against accepting type identity.

If supervenience, completeness and type identity fail to recommend causal or constitutive physical-psychological correlations, is there any further metaphysical relation that might? Perhaps we might start with local constitutive correlations between some psychological properties and neurophysiological properties, since these are not entirely implausible.[75] From there, we then move to a further commitment to local psychological-physical correlations.

But how might one make this move? It seems that bridge laws, of the kind discussed originally by Fodor (1974), are needed. Let N be a neurophysiological predicate that picks out a property individuated by the descriptions and explanation within that domain. Furthermore, let's suppose that N is locally correlated with mental property M. Now let P be a physical predicate that picks out a physical property. In order to yield local correlations between M and P, we might connect N and P with the following:

$$N \leftrightarrow P$$

Initially, we can read this minimally as expressing co-extensionality of the two predicates, N and P. N applies to all and only those things to which P applies. But as Fodor pointed out, this shouldn't be all we take the law to saying, for co-extensionality could hold in virtue of N and P being nomologically correlated as distinct properties of different ontological kinds: one as physical and the other as non-physical. So read minimally, the bridge law will not imply a local correlation between N and P, and hence not provide the physicalist with what they want here. But in order to support this local, modal correlation between N and P (and hence bridge M and P), we would need to read the bridge law as expressing an identity relation between the properties that N and P pick out. In which case (and to paraphrase Fodor), we would read the bridge law as saying, 'Every property instantiation that satisfies N is identical to some property instantiation that satisfies P.'

But the problem is that the non-reductive physicalist is not at liberty to endorse such a reading; on this interpretation, the bridge law effectively identifies property types, thus contradicting Distinctness and multiple realisability. On this basis, such bridge laws could not

[75] I take it that multiple realisability of mental states is potentially compatible with some highly specific properties of mental states being identical to neurophysiological states. That is partly because I have taken multiple realisability to apply for physical properties, i.e. properties of physics covered by Completeness. But even if we take multiple realisability to apply for neurophysiological N-properties, we might coherently say that mental state M of a subject is multiply N-realisable and has some highly specific, perhaps phenomenal, property that is not.

be deployed by the non-reductive physicalist in attempting to establish, from local correlations between psychological and neurophysiological variables, local correlations between the relevant psychological and physical variables.

Conclusion

We have now seen how the worries that might have been thought to motivate the application of Leanness do not arise in Exclusion contexts. In those contexts, the inclusion of three principles in particular means there is no reason to suppose that misleading correlations – pseudo-causal or confounding – will obtain between physical and psychological variables. The first of these principles was Close Multiple Realisability. If psychological properties are multiple realisable at close worlds, then there is little reason to suppose that interventions upon m-variables will produce misleading correlations.

The other key Exclusion principles are Completeness and Supervenience. These shore up our considerations from multiple realisability and widen the scope of our argument to include the prospects of misleading correlations when intervening upon P1. Arguments from the conjunction of Completeness and Supervenience show that there is no reason to suppose that we will produce misleading correlations when intervening upon either m-variables or p-variables when the latter are physical. The Completeness principle implies physical p-properties as the relevant realising properties. But whilst a global formulation of mental-physical supervenience also quantifies over physical p-properties, it does *not* thereby imply *specific* physical properties as correlated with *specific* psychological properties. Supervenience is thus silent on the matter of misleading pseudo-causal or confounding correlations resulting from interventions upon P1. The Exclusion context itself provides no reason for worrying about such misleading correlations.

I have further argued that, if the relevant p-variables correspond to physical properties, then both constitutive and causal correlations between p-variables are psychological m-variables are inherently implausible; *a fortiori*, misleading correlations of the kind that worry proponents of Leanness are inherently implausible. There is little scientific or philosophical reason to think that specific physical properties either cause or constitute specific psychological properties.

I have elsewhere stated that our assumption here is that, as a methodological principle, the application of Leanness is warranted only insofar as it is required. Since the worries that putatively motivate – and so potentially warrant – the application of Leanness to causal variable sets do not arise in Exclusion contexts, it follows that Leanness is not warranted in those contexts. Campbell's solution to the Exclusion Problem hence relies upon unwarranted application of the principle. So even if Leanness were not methodologically redundant, it would have no place in responding to the Exclusion Problem because in contexts pertinent to the problem, Leanness is unwarranted. This is the negative result of this section.

The positive result is that the Interventionist is free to explicate psychological causal relations without worrying about misleading pseudo-causal or confounding correlations arising in Exclusion contexts. Whilst there might plausibly be potential for such correlations between psychological and neurophysiological variables (see below), the latter are not variables implied by the Exclusion principles. So when thinking about psychological causation, the Interventionist need not be concerned about screening out misleading correlations involving physical variables.

But the most significant result is yet to be articulated. In the next section, we will see that our rejection of Leanness in Exclusion contexts furnishes us with the grounds for an Interventionist response to the Exclusion Problem that has no need of Leanness. It provides, in other words, a means by which to respond to the problem that does not require Pluralism.

Section 4: Diagnosing Pluralism

Introduction

We have now seen that Pluralism fails as a solution to the Exclusion Problem. The task of the present section is to answer the question of how it could be that Pluralism might have seemed attractive; that is, how it might be that one could take Leanness to be applicable to Exclusion contexts. In considering how one might have thought Leanness relevant to the Exclusion Problem, and hence how one might have thought Pluralism an appropriate response to the problem, we can arrive at a diagnosis of the mistake. In so doing, we will also arrive at a potential diagnosis of broader mistakes concerning the Exclusion Problem from an

Interventionist perspective. From there, we can broaden our diagnosis still further, to suggest how mistakes regarding the problem have been made by those reading Exclusion under Dependence or Assimilation interpretations.

One potential route to worries about misleading correlations between m- and p-variables is that of Supervenience, as discussed above. Whilst we might sympathise with the rough intuitions that tempt us into inferring prospective correlations on that basis, these do not withstand scrutiny when P is characterised as physical. However, another route might start not from a general metaphysical thesis about mental-physical relations, but rather, from considerations about empirically plausible correlations of a much more local sort. These considerations would demand that we think of P not as physical but as neurophysiological.

Prima facie, there are some highly plausible cases of causal relations (and hence invariant correlations) between p-variables and psychological variables, when the former are taken as neurophysiological. One mundane case would that of the causal relation exploited by the anaesthetist when putting a patient under just prior to invasive surgery. In Interventionist terms, the anaesthetist intervenes upon a neurophysiological variable to produce the desired value of the psychological outcome variable, i.e. a value of 0 for a general state of consciousness.

Conversely, there are also cases in which, plausibly, mental variables are causally efficacious with respect to neurophysiological ones. A typical pharmaceutical trial will presuppose such efficacy in asking its participants to imbibe the relevant drugs. Requesting this action on the part of the participants only makes sense as part of the testing process if it is assumed that the psychological registering of the request and subsequent decision to comply, are *ceteris paribus* causally efficacious in bringing about the neurophysiological effects of imbibing the drug.

It is also not implausible to suggest that *some* neurophysiological variables might be constitutive of – because identical to – some psychological variables. For example, we might suppose that neuroscientists could establish that a specific phenomenal property of a pain state is identical to a particular neurophysiological configuration.

Given the foregoing, it follows that in the case of P-M constitution, our previous worries about misleading correlations, of either kind, might be vindicated. For both kinds of

prospective misleading correlation, it was interventions upon P1 that were thought most likely to produce difficulties. And where P1 is constitutive of M1, because identical to it, it will follow that any intervention whereby the value of P1 is set to 1 would also set the value of M1 to 1, and where P1's value is set to 0, so too is M1's. Indeed, in the case where P1 is identical to M1, multiple realisability will of course not hold, and so interventions upon M1 would also potentially yield the kind of problematic correlations that we earlier posited as motivations for Leanness. When we assumed multiple realisability of M1, this went some way in mitigating the worries of misleading correlations, but in our present case, it seems that interventions upon M1 would produce misleading correlations quite as much as interventions upon P1.

We should recall that the putative worries centred upon two kinds of misleading correlation: pseudo-causal correlations and confounding correlations. Pseudo-causal correlations are constitutive relations in disguise; a pseudo-causal correlation between M1 and P1 *looks* like it might be a causal relation in virtue of the correlation obtaining in virtue of P1's constituting M1. A confounding correlation is a correlation that metaphysically or epistemically threatens the legitimacy of an intervention variable. Let us suppose that M1 is constituted by P1 and P1 causes P2. We have a candidate intervention variable (IV) for assigning values to P1 in respect of correlations with P2. A confounding correlation obtains if there is a causal path from (IV) to P2 independently of a path via P1. And if such a correlation does obtain, then (IV) is in violation of one of Woodward's constraints on legitimate intervention variables.

If p-variables are neurophysiological, then it is plausible that both these forms of misleading correlation will arise. For both rest upon the thought that interventions upon P1 will also assign values to M1. In the case of pseudo-causal correlations, a correlation between values of P1 and M1 results; in that of confounding correlations, a correlation between the intervention variable for P1 and the values of P2 results. It therefore seems that there are some plausible cases, i.e. cases in which P is neurophysiological and constitutive of M, in which the worries that putatively motivate Leanness might arise.

It might be thought that the situation worsens when we consider whether Leanness is applicable in these cases. I argued above that it is methodologically redundant, because one is in no position to distinguish causal from constitutive relations prior to substantiating the kinds of interventionist counterfactuals required for articulation of bona fide causal relations.

That line of thought relied upon the possibility of the very same potential conflation of relations (causal and constitutive) that motivate the Leanness principle in the first place. And indeed, we do encounter the same problem in the present case. Suppose that M1 is identical to neurophysiological P1, and so P1 constitutes M1.[76] As I claimed above, interventions upon M1 will also necessarily be interventions upon P1. Consequently, changes in the values of M1 will be correlated with changes in the value of P1. Alternatively, suppose that M1 is *caused* by P1. Then, by stipulation, there are correlations under intervention between the values of P1 and those of M1. We would be unable to differentiate between cases in which M1 is correlated with P1 because P1 constitutes M1 and cases in which M1 is so correlated because P1 *causes* M1.

The prospects for clarity when P1 is neurophysiological might seem still bleaker when we consider that a substantial part of the protracted metaphysical dispute between property dualists and supervenience physicalists concerns this very question: are the apparent correlations between physical and mental properties grounded in causal or in constitutive relations? Furthermore, part of the problem of establishing neural correlates of consciousness may also reasonably be articulated in these terms (Chalmers, 2000)**.** Given that these debates are not straightforwardly settled – as testified by the continued disagreement between the relevant camps – we have further reason for pessimism regarding our capacity to distinguish between causal and constitutive relations.

Given the potential for misleading correlations between psychological and neurophysiological variables, it is reasonable to speculate that Leanness might be thought pertinent to Exclusion contexts on the basis of conflating neurophysiological variables and physical ones. If we took the relevant physical variables at issue in Exclusion contexts as neurophysiological, then it might well seem that the worries about misleading correlations demand a principle like Leanness to sort the causal from the constitutive.

This way of thinking about the situation opens up the possibility of framing the mistakes implicit in the Intervention reading of Exclusion, too. Just as Pluralism might plausibly rest upon the conflation of physical and neurophysiological properties, so might the Intervention reading arise from the conflation of two pictures, one involving physical properties and the

[76] Identity of variables / properties is sufficient for the constitution relation; it is not necessary.

other, neurophysiological properties. The conflation takes elements of the picture relevant to physical properties and mixes these with elements of the picture for neurophysiological ones. The result is a picture on which the Intervention reading of Exclusion emerges.

To appreciate how this conflation works, we should further consider the case of constitutive correlations between neurophysiological and physical properties. It is important to see that, whilst the potential for these renders plausible the worries that motivate Leanness, they do not contribute to the Exclusion Problem. For whilst the supposition that P1 is a neurophysiological variable opens up the potential for worries that might motivate (and undermine) the Leanness Principle, it also closes down worries pertaining to causal exclusion. If P1 is neurophysiological, then P1 is not a member of a property class to which we have any reason to think Completeness applies. There are well-known reasons for doubting that any special science is causally complete, e.g. causal generalisations individuated at the level of neurophysiology, for instance, are *ceteris paribus*, and the factors that might yield exceptions to those generalities can be factors not themselves individuated by the theories of that science – for example, an earthquake that prevents a neurophysiological cause from bringing about the effect typically associated with it in the relevant background context. Given considerations like this, the burden of proof lies with those who wish to assert that Completeness does cover the properties of the special sciences. For the purposes of our present discussion, I therefore adopt as a working assumption that these sciences are not causally complete / closed. On that basis, if P2 is neurophysiological, then there are no grounds for asserting that it must have a sufficient neurophysiological cause, and hence no grounds on the basis of the Exclusion principles for worrying that, for instance, where a mental variable M1 ostensibly bears a causal relation to a physical (i.e. neurophysiological) effect P2, M1 is in danger of causal exclusion by an underlying neurophysiological cause, P1. So if p-variables are neurophysiological, there is no reason to think that causal relations between m-variables imply widespread, systematic overdetermination.[77] Whilst there may plausibly be cases in which some neurophysiological variable is causally related to a psychological one, and such cases might be problematic for the Interventionist in terms of

[77] We should also note that our proposed hypothetical case above, involving identity relations between p- and m-variables, need not worry us from the point of view of Exclusion. Even if Completeness were somehow argued to apply to neurophysiological variables, for the case in which P1 is constitutive of M1 because identical to it Exclusion will not apply because the Distinctness principle will not.

distinguishing that relation from a constitutive one, these are not problematic in terms of Exclusion.

On this basis, I take it that neurophysiological p-variables do not figure in Exclusion contexts. So the picture whereupon p-variables are neurophysiological includes the following salient elements:

The Neuro Picture:

(i) PN-variables are not subject to Completeness.

(ii) PN-variables are not covered by Supervenience.

(iii) Specific pN-variables are plausibly constitutively correlated with specific psychological m-variables.

(iv) Specific pN-variables are plausibly causally correlated with specific psychological m-variables.

By contrast, we previously saw that Exclusion contexts *do* implicate physical p-variables because Completeness applies to the physical domain and Supervenience quantifies over that domain. However, we also saw that a global supervenience formulation does not entail constitutive correlations between specific p-variables and specific psychological m-variables. It is silent on which sets of m-variables supervene upon which sets of p-variables. Furthermore, constitutive relations between specific physical p-variables and specific psychological m-variables are independently implausible. The picture on which p-variables are physical includes the following:

The Micro Picture:

(i) PM-variables are subject to Completeness.

(ii) PM-variables are covered by Supervenience.

(iii) Specific pM-variables are not plausibly constitutively correlated with specific psychological m-variables[78].

[78] This is supported by both the absence of entailment of such correlations by Supervenience and their implausibility considered independently of Supervenience.

Neither picture is that of the Exclusion Problem under the Intervention reading. The Neuro Picture has p-variables (neurophysiological) that are not subject to Completeness and not covered by Supervenience. Given this, there is no ostensible entailment of systematic overdetermination.

The Intervention interpretation of Exclusion comprises the following principles[79]:

Distinctness: Mental properties are metaphysically distinct from physical properties.

Supervenience: Mental properties supervene upon physical properties.

Causation: Mental properties have – in their own right qua mental properties – genuine causal efficacy.

Completeness: All physical effects have sufficient physical causes.

Correlation: If m-properties supervene upon p-properties, then m-effects are correlated with p-causes and some p-effects are correlated with m-causes.

Intervention Causation: Variable *c* is a cause of variable *e* iff values of *e* are correlated with values of *c* under interventions upon *c.*

According to Intervention, there are two routes to systematic overdetermination.

First route: Assume that, as per Causation, psychological variable M1 causes psychological variable M2. Assume, as per Supervenience, that M2 supervenes upon physical P2. Assume, as per Completeness, that P2 has a sufficient physical cause, P1. Given Correlation, P2 is correlated with M1: values assigned to M1 are correlated with those taken by M2, and M2 supervenes upon P2, so those interventions will also be correlated with values taken by P2. By Intervention Causation, P2 is hence caused by M1. But P2 has a sufficient physical cause in P1. Hence, P2 is overdetermined by M1 and P1.

[79] The full list would also include the Exclusion principle itself (No single effect can have more than one sufficient cause occurring at any given time—unless it is a genuine case of causal overdetermination). But we are concerned here with those that ostensibly generate systematic diagonal causal relations.

Second route: The second route is similar to the first in many respects but differs in that M2 is the overdetermined variable. Assume that, as per Causation, variable M1 causes variable M2. Assume, as per Supervenience, that M2 supervenes upon physical P2. Assume, as per Completeness, that P2 has a sufficient physical cause, P1. Given Correlation, M2 is correlated with P1 because P2 is correlated with P1 and M2 supervenes upon P2. By Intervention Causation, M2 is causally correlated with P1. But, by hypothesis, M2 is caused by M1. Hence, M2 is overdetermined by M1 and P1.

Now, clearly the Neuro Picture is not the picture outlined by the Intervention model. On this model, the p-variables are physical, not neurophysiological. This means that neither Completeness nor Supervenience are relevant for the Neuro Picture. The absence of Completeness for neurophysiological p-variables means that, even if some mental variables do constitutively depend upon them, there is no reason for supposing that these p-variables have sufficient p-causes. Therefore, even if there are diagonal causal relations obtaining between M1 and P2, there is no reason to take P2 as overdetermined, since no reason to assume that P2 has a sufficient physical variable as its cause. Furthermore, even if there were some diagonal causal relations between m-variables and p-variables, there would be no grounds to assume systematic relations of this kind, since Supervenience has nothing to say about neurophysiological variables. The absence of Supervenience is the absence of any principle by which to infer systematic generation of diagonal causation.[80] Indeed, the absence of either Supervenience of Completeness on their own would be sufficient to block the assumption of systematic overdetermination. Even if those diagonal causal relations between psychological and neurophysiological variables did overdetermine effects shared with causes at the other level, there would be no reason to assume that such overdetermination is systematic: the absence of Supervenience means that there is no reason to suppose that *all* mental variables will be involved in such relations; the absence of Completeness means that there is no reason to think that *all* neurophysiological effects will have sufficient

[80] As noted in Section 1, the systematizing role in the Exclusion Problem is played by Supervenience together with Completeness. Supervenience, as a principle that quantifies over all mental properties, entails that all mental properties supervene upon physical ones. Completeness, as a principle that quantifies over all physical effects, has it that every physical property subvening a mental property will have a sufficient physical cause (assuming it has a cause).

neurophysiological causes. This is really by way of unpacking the implications of our stipulative notion of problematic overdetermination. In the Introduction, we said that overdetermination is problematic only if systematic, and systematic iff entailed by a modally strong, general principle, i.e. Supervenience, in conjunction with Completeness.

In terms of the two routes to ostensible systematic overdetermination, both of these require Supervenience, first to entail the relevant diagonal causal relation and then to entail that such causation is systematic. On both routes, the critical relation is the supervening of M2 upon P2. But on the Neuro Picture, P2 would be neurophysiological property and so Supervenience would not entail that M2 *does* supervene upon it. Clearly, then, the Neuro Picture does not coincide with the picture putatively generated by the Intervention model of Exclusion.

But nor does the Micro Picture. On that picture, with p-variables as physical, Supervenience and Completeness apply to p-variables. But the other component of that picture is the plausible absence of constitutive correlations between specific m-variables and specific p-variables. Given that both routes in the Intervention model invoke supervenience relations – i.e. constitutive dependence relations – between specific variables, M2 and P2, these routes are no more part of the Micro Picture than the Neuro Picture. The Micro Picture also fails to coincide with the Intervention model.

Our speculative suggestion is that the Intervention model is based upon a conflation of the Micro and Neuro pictures:

Micro-Neuro Picture

(i) P-variables are subject to Completeness.

(ii) P-variables are covered by Supervenience.

(iii) Specific p-variables are plausibly constitutively correlated with specific psychological m-variables.

(iv) Specific p-variables are plausibly causally correlated with specific psychological m-variables.

We should note that the Micro-Neuro picture omits the qualifiers on p-variables that were present in the distinct Micro and Neuro formulations. The p-variables in the Micro picture (principles (i) and (ii)) were designated as 'pM-variables'; those in the Neuro picture (principles (iii) and (iv)) designated as 'pN-variables'. That omission is deliberate on my part, because the contention here is that a sympathetic predisposition to the Intervention model of Exclusion is the result of mixing two plausible principles from the Micro picture and two from the Neuro, but without paying close enough attention to the differences between the variables designated by those two sets of principles. When the Completeness and Supervenience principles, (i) and (ii), are read with p-variables as physical, they are plausible; when the constitutive and causal correlation principles (iii) and (iv), are read with p-variables as neurophysiological, they are plausible. But they are not so when the relevant variables are switched. That is why any reading that fails to differentiate appropriately between the p-variables in principles (i) and (ii), and those in principles (iii) and (iv), will be flawed.

In contrast with the Micro and Neuro pictures taken separately, the Intervention model does coincide with the Micro-Neuro picture. Both routes to ostensible systematic overdetermination build upon the first two principles, taking p-variables to be quantified over by both Supervenience and Completeness. They then assume – via the Correlation principle – that this entails that specific m-variable M2 supervenes upon (i.e. is constitutively dependent upon) physical p-variable P2. It is in this assumption that we might potentially see the influence of the Micro-Neuro picture. The entailment fails; global supervenience does not imply the constitutive dependence of specific m-variables upon specific physical variables. And yet there is something tempting about the latter assertion. Our present contention is that it is tempting if one is, perhaps unconsciously, thinking of p-variables as neurophysiological.

The Micro-Neuro Picture: Generalising the Diagnosis

We are now in a position to see that the Micro-Neuro picture offers not only a diagnosis of why the Intervention model of Exclusion might appeal, but also of sympathies for the Dependence or Determination readings of Exclusion. Indeed, it has the potential to explain the temptation of any reading of Exclusion that depends upon the supervening of specific psychological variables upon specific physical variables.

We'll consider how the diagnosis might work for the Dependence and Determination readings of Exclusion. The Dependence model comprised the following:

Dependence Assumption (DA): If mental properties supervene upon physical properties, then M is counterfactually dependent upon P.

Dependence Chains: A dependence chain is any finite sequence of events a,...n such that b is counterfactually dependent upon a, c is counterfactually dependent upon b etc.

Causal Dependence: *c* causes *e* if there exists a dependence chain leading from *c* to *e.*

Supervenience: Mental properties supervene upon physical properties.

Distinctness: Mental properties are metaphysically distinct from physical properties.

Causation: Mental properties – qua mental properties – are causally efficacious.

Completeness: All physical effects have sufficient physical causes.

As with Intervention, a Dependence reading of the Exclusion Problem does not coincide with either the Micro or Neuro pictures taken separately. The crucial principle here is the Dependence Assumption. Given a global supervenience thesis, there is no reason to take it that a specific psychological property, M, will be counterfactually dependent upon specific physical property, P.[81] Global supervenience simply does not imply a supervenience relation between any specific psychological property, M, and any specific physical property, P. But the Neuro picture speaks only of neurophysiological properties, and denies both Completeness and Supervenience for such properties. So that picture is not commensurate with the principles cited by the Dependence model.

However, in view of our above diagnosis of the Intervention model, it is natural to speculate that a similar explanation applies in the case of Dependence. Our above suggestion was that the Micro-Neuro picture is behind the illicit move from general supervenience to constitutive relations between specific properties. That picture would also therefore account for the same

[81] I argued in Chapter 2, Section 2, that counterfactual dependence of M upon P does not follow from m-properties' supervening upon p-properties anyway. That was because of considerations concerning close multiple realizability of m-properties and wide physical determination bases for those properties. But in the present context, the salient point is that global supervenience does not entail such dependence of M upon P because these are specific properties and a global formulation only entails global dependence of m-properties upon p-properties.

move made in the Dependence model, a move implicit in the Dependence Assumption principle: If mental properties supervene upon physical properties, then M is counterfactually dependent upon P. The antecedent expresses a general supervenience claim; the consequent a dependence claim. The counterfactual dependence claim is supposed to follow from the antecedent *because* the supervenience relation between the family of mental properties and the family of physical properties is taken to imply constitutive dependence relations between *specific* members of those families. That implication is the implicit bridge between the antecedent and consequent in the Dependence Assumption: the general supervenience claim expressed by the antecedent is implicitly taken to imply the further claim that specific mental properties are constitutively dependent upon specific physical properties. The latter is then taken to imply the claim expressed by the consequent. The move that provides the bridge from global supervenience to specific counterfactual dependence in the Dependence Assumption is a move that fits with the Micro-Neuro picture. Global supervenience does not entail counterfactual dependence of specific mental properties upon specific *physical* properties. But as the Micro-Neuro picture has it, it is plausible that *some* mental properties are constitutively dependent upon *neurophysiological* properties. So an illegitimate conflation of global supervenience (physical) with specific constitution (neurophysiological) might account for the illicit manoeuvre within the Dependence Assumption principle. We start with global supervenience, move to highly local supervenience, and then to specific counterfactual dependence.

The same story applies in the case of the Determination model of Exclusion. Just as Intervention and Dependence rely upon a move from global supervenience upon physical properties to constitutive dependence of specific mental properties upon specific physical properties, so too does the Determination reading. The Determination reading comprises:

Supervenience: Mental properties supervene upon physical properties.

Distinctness: Mental properties are metaphysically distinct from physical properties.

Causation: Mental properties have – in their own right qua mental – genuine causal efficacy.

Completeness: All physical effects have sufficient physical causes.

Determination: If m-properties supervene upon p-properties, then P necessarily determines M.

Assimilation: If M is necessitated by P, M is a relatum (of the same type, i.e. cause or effect) in any causal relation where P is a relatum.

As before, we read the Supervenience principle as a global thesis. In the Determination model, we can see that the Determination principle plays a key role in moving from a general supervenience thesis to the antecedent of the Assimilation principle. M and P are here taken to be specific psychological and physical properties. We might say that Determination principle zooms in upon the specific physical and psychological properties quantified over by the Supervenience principle, rendering local the global necessitation implied by supervenience. In this way, the Determination principle has a function analogous to that of Dependence Assumption in the previous model: in terms of our taxonomy, it facilitates a transition from the Micro picture to the Neuro picture. But as we have already seen, that transition is unwarranted, and so might be explained by the unconscious conflation of the two pictures.

Originally, Determination served to articulate two potential routes to diagonal causation, and hence ostensible systematic overdetermination.

First Route: In accordance with Causation, M1 causes M2. In accordance with Supervenience, M2 supervenes upon P2 and M1 upon P1. In accordance with Completeness, P2 has a sufficient physical cause, which is posited as P1. But by Determination, if M1 supervenes upon P1, then P1 necessitates M1. And by Assimilation, if P1 necessitates M1, then M1 is drawn into the causal relation between P1 and P2. Therefore: P2 is caused by both P1 and M1; hence, P2 is overdetermined.

Second Route: Where the first route posits M1 as cause of P2, the second takes it that M2 is caused by P1. On the first route, the determination of M1 by P1 draws M1 into the causal relation between P1 and P2. On the second route, the determination of M2 by P2 draws M2 into the causal relation between P1 and P2.

In more detail: In accordance with Causation, M1 causes M2. In accordance with Supervenience, M1 supervenes upon P1 and M2, upon P2. By Completeness, P2 has a sufficient physical cause, which we posit as P1. By Determination P2 necessitates M2, and so by Assimilation, M2 is drawn into the causal relation between P1 and P2. M2 is caused by P1, in virtue of supervening upon P1's physical effect, P2. But M2 is caused by M1 also. So M2 is overdetermined by the conjunction of P1 and M1.

Neither route will work without the illicit move made by the Determination principle. Global supervenience does not entail the necessitation of specific psychological properties by specific physical properties. My suggestion is that the explanation for why the Determination principle might be thought plausible is the background assumption of the Micro-Neuro picture.

Summary

I have postulated a diagnosis of the error made in assuming Leanness warranted in Exclusion contexts: the illicit conflation of two pictures – the Micro and the Neuro. On the Micro picture, global supervenience entails constitutive correlations between global sets of psychological and physical properties. On the Neuro picture, we have constitutive and causal relations plausibly obtaining between specific psychological and specific neurophysiological properties. But the Micro picture does not entail constitutive relations between specific physical and specific psychological properties. To move from global supervenience to specific constitution relations is to mix the two pictures. To urge Pluralism is to unconsciously adopt the Micro-Neuro picture.

This diagnosis of Pluralism suggests a more general diagnosis of the erroneous moves made by the three models of the Exclusion Problem: Intervention, Dependence, and Determination. Each of these trade on principles which implicitly move from the global to the specific, a move that is unwarranted but – on our diagnosis – rendered plausible by the Micro-Neuro background picture**.** In the Intervention model, the move comes in the Correlation principle; supervenience is taken to entail that m-effects will be correlated with p-causes, but these are of course specific variables. Nothing in a global formulation entails that interventions upon specific physical variables will produce correlations with specific psychological variables or

vice versa. Yet the move might seem palatable given the Neuro picture. In the Dependence model, the Dependence Assumption moves from general supervenience to dependence relations between specific psychological and physical properties, and in Determination, the Determination principle makes a similar move from general supervenience to specific necessitation relations. Each of these principle enshrine moves that are unwarranted, but might seem plausible on the basis of a background assumption of the Micro-Neuro picture.

Conclusion

In this chapter, I have argued for three main claims:

(i) Campbell's Pluralism fails as a solution to the Exclusion Problem. This is partly because a key principle of that strategy, the Leanness principle, is methodologically redundant. But more significantly, that principle is unwarranted, because unmotivated. The misleading correlations that allegedly motivate the principle will not arise in Exclusion contexts, without controversial and likely unappealing assumptions. Some of those correlations will arise only if psychological properties are not closely multiply realisable. And insofar as the relevant p-variables are specific properties of physics, no misleading correlations will arise unless supervenience is radically local.

(ii) The motivation and ostensible warrant for Pluralism was predicated on an erroneous conflation of two pictures – the Neuro and Micro pictures. Such a conflation might render plausible the notion that supervenience implies the threat of misleading correlations between psychological and physical variables.

(iii) Our earlier Exclusion models – Dependence, Determination and Intervention – can also be seen as predicated upon the same conflation.

As such, the mistakes implicit in Pluralism have provided the basis for a diagnosis of Exclusion worries more generally. If so, then our non-reductive physicalist can guard against any resurging intuitions of specious overdetermination.

Chapter Five - Towards Causal Pluralism: Interventionism & Incompleteness

Introduction

In previous chapters, I have argued that implicit to exclusion-related concerns about mental causation are a host of oversimplifications. In Chapter 2, we saw that three Exclusion models imply systematic overdetermination on the basis of overly simple conceptions of the relation between supervenience and diagonal causation. In Chapter 3, I argued against oversimplifying the denial of systematic overdetermination, i.e. against thereby assuming that there are *no* diagonal causal relations from the mental to the physical. Chapter 4 was an extended warning against an oversimplified view of Exclusion contexts, particularly with respect to thinking about mental causation in Interventionist terms. In each case, the identification of potential oversimplification constitutes a kind of diagnosis. I want to say that where there may have seemed to be a problem, this was plausibly due to the assumptions sustaining these overly simple conceptions of how mental causation fits into a world construed in physicalist terms.

Furthermore, diagnosis in each offers a prophylactic: if we can acknowledge the assumptions at play in our thinking about mental causation, then we can set about avoiding them. The hope is that in so doing, we might see that philosophy need not worry about psychology interfering with physics, or physics with psychology. In avoiding nefarious assumptions about the relationship between mental causation and physical causation, we might make peace with both as autonomous causal domains that are nonetheless neither causally nor metaphysically isolated from each other.

One of the main assumptions targeted in previous chapters was that supervenience, taken together with three other Exclusion principles[82], implies diagonal causation and hence, systematic overdetermination. I then showed, in Chapter 3, that despite supervenience being benign in this respect, such diagonal causation is plausible in cases where mental properties supervene upon physical properties. In this way, I defused the threat that supervenience has typically been taken to pose to mental causation.

[82] Distinctness, Causation, and Completeness.

But perhaps some might find Exclusion worries lingering. Perhaps it is not only the systematicity of overdetermination that concerns them. Causal completeness might also be of concern, because it guarantees physical causes for physical effects and furthermore, guarantees the *sufficiency* of those causes.

We can bring this out by reflecting upon the Hadron Collider case. That case is a case of diagonal mental causation that is not supervenience-based. The mental cause is related to a physical effect but not in virtue of its supervening upon some specific physical property. It is what I have called, *direct* diagonal causation. As such, it is not a case that demonstrates (or entails) systematic overdetermination; on my use of the term, overdetermination is *systematic* only if the diagonal causal relation is entailed by a general, modally strong principle that guarantees such relations across worlds for all causes of the relevant kind (i.e. mental). The candidate principle in the Exclusion Problem is of course the supervenience thesis. So Hadron Collider is not a case of systematic overdetermination because the diagonal relation is not entailed by supervenience. But one might consider the case and still be unsure: how *could* the mental cause be efficacious given that the physical effect is guaranteed by Completeness to have a sufficient physical cause? How could my intentional flicking of the switch genuinely cause the collider to start given that there must be some physical cause sufficient to start the collider? The worry of exclusion might persist. Even if not systematic, perhaps Hadron is a case of overdetermination that still has the power to discombobulate. If so, then perhaps it is Completeness that lies at the root of that worry. For whilst we can think of cases which include non-supervenience-based causation, and which do not typically provoke concern, the Hadron case is plausibly different.

One such case is the famous firing squad case, in which several bullets, each alone sufficient to cause fatal injury, converge upon the same human target. Nothing in this case suggests that any of the gun-firing obtains in virtue of supervenience. It looks like a case of non-supervenience-based causation. But unlike, perhaps, Hadron Collider, the firing squad case does not typically arouse suspicion of *problematic* redundancy. We can recognize that, in a sense, the majority of shots are redundant; a single shot would have sufficed to kill. But we are not typically troubled by that. In Hadron, however, we might be worried that the redundancy of the mental cause does not look benign. One explanation for the difference is that Completeness plays no significant role in the firing squad case but does in Hadron

Collider. The presence of multiple sufficient causes is not alone enough to trouble us; but the presence of multiple sufficient causes when one guaranteed is by Completeness might be.[83]

This chapter will offer a broad, relatively speculative and schematic proposal for diagnosis and cure of such lingering worries. The proposal is that Interventionists committed to their conception of causation in psychology, and committed to causal Completeness, should endorse causal pluralism. On this picture, Interventionism in psychology sits beside a nomological regularity conception at the level of physics. The proposal would have the potential to both ameliorate residual Exclusion worries and provide the framework for a positive conception of interlevel causation and autonomy.

In Section 1, I present a reason to be sceptical of the prospects for deploying an Interventionist notion of causation in physics: the Incompleteness Argument. This argument seeks to demonstrate that there can be no complete set of explicit causal relations under Interventionism. This forms part of my case in support of the claim that Interventionism is limited as a conception of causation, and is compatible with causal pluralism.

In Section 2, I examine a potential objection to the Incompleteness Argument. We will see that the objection is not compelling, because it rests upon a problematic strategy for blocking the argument. The Incompleteness Problem persists.

In Section 3, I consider a strategy from Woodward in responding to the Incompleteness Problem. I argue here that whether or not the strategy is successful, there is a tension internal to Interventionism that could be relieved by adopting pluralism.

In Section 4, I extend my pluralist proposal to offer a diagnosis and cure of lingering Exclusion concerns.

[83] Why might Completeness render overdetermination apparently problematic, even in the absence of supervenience-based causation? We conjecture that Completeness involves an implicit appeal to exceptionless laws, and that the *implied lawful sufficiency* of physical causes is at the root of the trouble. But the point here is just that Completeness might be cause for concern in Hadron; for whatever is at the root of that, we will argue in this chapter that causal pluralism is a plausible diagnosis and prophylactic for the worry. The idea that Completeness involves implicit appeal to exceptionless laws will be further discussed below, in Section 3b.

Section 1: The Incompleteness Problem

In this section, I present my first reason for scepticism about the prospects for deploying an Interventionist notion of causation in physics: the Incompleteness Problem. The problem alleges that Interventionism is constitutively incapable of explicating a single, complete set of basic physical principles. Such a set would need to include "variables relative to which all causal relations can be made explicit" (Campbell, 2020, p.154), and this cannot be done. We will start by clarifying the claim, and then proceed to examining the reasons for it.

First, what do we mean by a single, complete variable set here? The single variable set is a set containing all variables relative to which causal relations can be articulated. But already, there are clarifications required. For one thing, we should not construe 'all causal relations' to mean all causal relations at all 'levels'. So, for example, we should not take this to include causal relations between kicks and goals, or between desires and fridge visits, or between smoking and yellow fingers. Why not? The single set at issue is supposed to be the set of principles constitutive of a 'theory of everything'. But popular misuse of the term aside, such a theory is properly taken as a theory not of fundamental forces, economic inflation, political change, cake baking, fashion trends and all else, but rather a theory that postulates a single set of principles uniting the forces of the Standard Model: gravity, electromagnetism, the strong and weak nuclear forces. Given that the Standard Model is a model in physics, and that any unification of the forces would consist also in physical principles, we should take a 'theory of everything' to be a physical theory, and thus treat any single variable set relative to that theory as comprising physical variables.[84]

So I do not interpret the single set of variables as including all variables relative to all causal relations at any level of description. On the contrary, I interpret the set as including only variables specified by physics.

[84] In addition, given that Campbell's presentation of the problem takes place in the context of a passage invoking the project of scientific unification, it is fair to take this hypothetical single set to include fundamental physical variables. But the kinds of 'higher level' variable mentioned above are taken to supervene upon physical variables (or whatever entity we take to correspond to variables, such as properties). Since Campbell rules out variables related by constitution from belonging to the same variable set relative to causal relations (i.e. he upholds Leanness as a constraint), presumably he thinks no further argument is required to exclude higher level, supervening variables from entering into the complete, unified, single set of variables under current consideration.

Second, the single set includes all variables relative to which causal relations can be articulated. I will call such variables, 'causally significant'. I therefore take the complete set of all causally significant variables to be *the set of all physical variables that are explicitly causally related under interventions*.

The Incompleteness Argument

Now to the argument for the claim. Here I offer a reconstruction of the argument in Campbell (2020):

1) A variable set is causally complete if and only if it includes all physical variables that are explicitly causally related under interventions.
2) For any causal variable set V, and any member endogenous variables X and Y, X is only explicitly causally related to Y if there is some exogenous variable I to intervene upon X.
3) For any causal variable set V, if I is an exogenous variable with respect to endogenous variables in V, then I is not itself a member of V.
4) For any causal variable set V, V is not a causally complete variable set.

We can think roughly of the problem as follows. Suppose we have a local system of causally related variables, such as a billiards table[85]. We can observe the behaviour of the balls in motion and, with assiduous care and attention, formulate causal hypotheses for certain trajectories of certain balls by observing correlations. We then check these hypotheses by intervening upon the system; that is, we devise ways of holding some balls in position whilst causing the causal candidates to follow their usual trajectory. Such interventions involve us reaching into the local system and having an influence upon variables within it. In principle, there seems no bar on this method.

But suppose that we are concerned not with a local system, but rather, a system that is global in the sense that causal variables operating within the system are the only operative variables for the system. There are no variables capable of causing anything within the system that are

[85] Hitchcock (2007) cites a similar example.

not themselves variables internal to the system. This contrasts with the billiard table system, where we could reach in and causally influence variables that were themselves hypothesized as causal within the system – e.g. balls on the table that were hypothesized as causally influential for the motion of other balls. The global system does not admit of such intervention. It is closed. Since any intervention upon the system has to be both causal and external to the system, it follows that no global system admits of interventions because the variables internal to the system are only causally related to other internal variables.

The thrust of the Incompleteness Argument is this: any system that is global in our present sense will be one for which causal relations cannot be explicated. Since, on an Interventionist view of causation, causal relations can only be made explicit with interventions, and since no such interventions are possible on a global system, any global system will lack explicit causal relations.

Let's consider the argument in more detail. Premise (1) just states our definition of a causally complete variable set. The justification for premise (2) is provided by the basic logic of Interventionism. Causation has to be relativised to variable sets in order to enable isolation and articulation of causal variables with respect to outcome variables. For any explicable causal relation there must be a possible intervention upon the causal variable, and that intervention requires an exogenous variable. It requires an exogenous variable because *interventions are themselves causal*: the intervention variable causally brings about changes in the value of the causal variable.

To see the basis for premise (3), we should first clarify what we mean by an 'exogenous' variable. Since interventions are deployed for the purpose of making explicit causal relations between candidate causes and outcome variables, the causal relation between the intervention variable and the causal candidate variable is excluded from the picture. In this sense, we might say that the causal relation between the intervention variable and its relative causal candidate is *implicit*. Such implicitly causal interventions are made using exogenous variables, i.e. variables that are outside of the variable set relative to the difference-making of the causal variable with respect to the outcome variable. In the billiard table example, our reaching onto the table and causing balls to move would be an exogenous variable. Our reaching in and pushing a ball is itself a causal matter, but for the purposes of explicating causal relations within the system, that causal relation remains implicit.

Now to the reasoning in premise (3), which turns upon the constraints on legitimate interventions. These constraints *rule out a variable from being both endogenous and exogenous with respect to the same set*. In Chapter 4, we saw Woodward's necessary conditions on legitimate intervention variables. Where variable I is the candidate exogenous variable for endogenous cause X with respect to outcome Y, these conditions are (see also Figure 13):

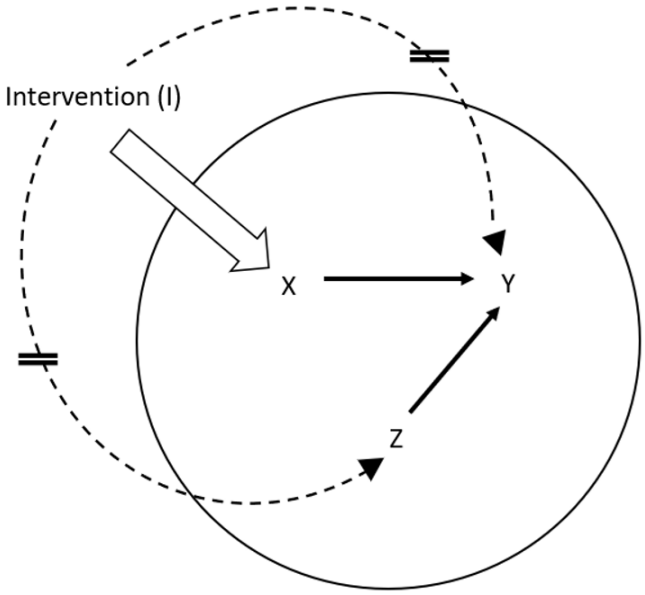
1. I causes X;

2. I acts as a switch for all the 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;

3. 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 − X − 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;

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

(Fig.13)



The most salient condition for the present point is (2), according to which exogenous variables must switch off the causal influence of any endogenous variables with respect to the candidate cause. If we permitted an endogenous variable to function as an exogenous variable with respect to its own set, then it would fail to satisfy this condition.

Suppose that variable I is both exogenous and endogenous relative to the same set. When we want to refer to I as exogenous, we call it I^{EX} and when we refer to it as endogenous, we call it I^{EN}. Suppose that the variable set also contains X as causal candidate for outcome variable Y. We assign a value to I^{EX} in order to assign a value to X with respect to Y. That is, we use I^{EX} to intervene upon X. But in assigning a value to I^{EX}, we are making use of an (implicit) causal relation between I^{EX} and X. In which case, I^{EX} fails constraint (2); if I^{EX} is an implicit cause of X then I^{EN} is an endogenous cause of X, and any such causes must be switched off by the exogenous intervention variable. So on pain of violating this constraint, an intervention variable cannot be both endogenous and exogenous with respect to the same set.[86]

We take claim (4) to hold for the following reasons. If intervention variable I is to be a legitimate intervention, I must be itself a causal variable (see Woodward's condition (1), above). But, as per premise (3), if I is exogenous, I cannot be a variable endogenous to any set that includes the candidate cause (X) to be intervened upon and the respective outcome variable (Y). And yet, if I is a causal variable with respect to X, then it must be explicable relative to some variable set containing I and X as endogenous variables. But the variable set V, containing causal candidate X with respect to outcome variable Y, is hypothesized as complete – understood here as meaning that V includes all variables related by physical causal relations. So there is no other variable set relative to which I can be characterized as causal variable for X, and hence no variable set relative to which I can function as causal intervention upon X. If so, then paradoxically, X can have no exogenous intervention variable operative

[86] We get the same result with slightly alternative reasoning. We start with the assumption that the set of endogenous variables is complete, and so includes all physical outcome variables and all related causal variables. We also assume that all causal variables in the set also have a cause, i.e. we assume that no causes are either self-causing or originate non-causally. It follows that every variable in the set is an outcome variable relative to some other endogenous cause. So if the set included all exogenous variables as endogenous, then it would include exogenous variables that were causally related to some endogenous outcome variables. But this would contravene what I'm calling the independence constraint on exogenous variables: if a variable is exogenous relative to a set, it must not be independently correlated with any endogenous outcome variable, and so not causally correlated with any such variable.

upon it, and therefore cannot be an explicit causal variable with respect to Y. The upshot is that, on the assumption that every causal variable must have an exogenous interventional variable causally acting upon it, no variable set can be complete with respect to a given causal domain (in this case, the physical domain). Any candidate set V will be incomplete either because some causal candidate lacks an exogenous variable by which to be explicated as a bona fide causal variable with respect to some outcome variable in the set, or because there is some exogenous (causally efficacious) variable which is not itself a member of the set.

Section 2: A Potential Objection

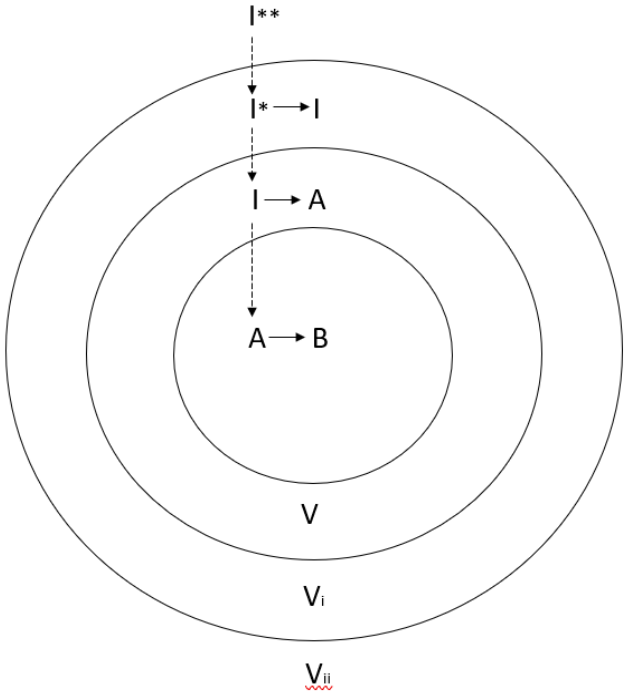
There is a potential objection to the Incompleteness Argument. The broad claim is that a single, complete set of explicit causal relations might be possible if the set is structured in the right way. The idea is that the argument might be evaded with the possibility of a structured set of proper subsets of variables. However, we will see that the prospects of a complete, structured set of explicit causal relations are not promising.

We'll start with a very simple example of a structured set. Here, set V is assumed to be the initial candidate for completion. Now, as we know, set V must exclude I to qualify as a set relative to which the causal correlation between A and B can be properly articulated. So it's right that V cannot include I. So we now make V a proper subset of an expanded set, V^i that includes I, A and B. Given that intervention variables must be exogenous, this would need to be a set that contains the set with just A and B as a proper subset, subset V. So V^i is a set that contains I, A and B as variables **(see Figure 1)**. The key idea of the proposal is that it might be possible to construct a structured set of proper subsets that includes all explicit causal relations by including all exogenous variables in distinct subsets from the endogenous subsets they serve.

An immediate challenge for such a model is the *regress problem*: given that exogenous variables must themselves be explicated as causal variables with respect to their endogenous targets, there is the threat of generating ever more variable subsets with which to explicate ever more exogenous intervention variables as causal.

Consider again our variable set V^i. For I to be causally efficacious with respect to A – that is, for I to be a legitimate intervention upon A – there must be some variable I* that acts as an intervention variable upon I with respect to A as outcome variable. And furthermore, I* must be exogenous to the variable set containing I and A as causal and outcome variables, respectively. So V^i , as a set comprising I, A and B cannot suffice: we need a set that includes I, A, B and I*. We require V^{ii} which contains (i) a proper subset containing A and B and excluding I, and (ii) a proper subset containing I and A, excluding I*. But because I* will also need to be causally efficacious with respect to I, there will need to be a further exogenous variable, I**, that functions as an intervention upon I*. So we also need (iii) a proper subset containing I* and I, excluding I**. And so on: an endless proliferation of causal subsets (Figure 14).

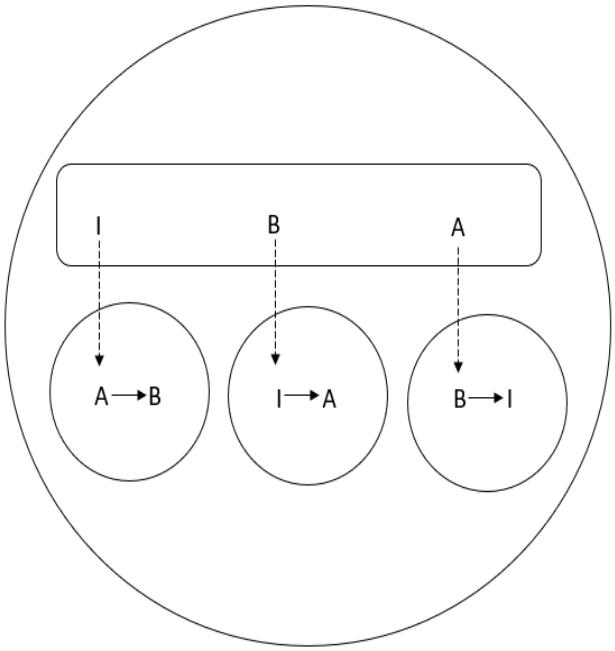
(Fig. 14)



Is there any way of avoiding an intervention regress? One potential strategy might be to postulate intervention variables that are already members of the set in virtue of being members of some subset wherein they are causal or outcome variables. The simplest model

of a single set organized in this way would consist of the following members: {A, B, I}. It would have as proper subsets the following: {A,B}, {I,A}, {B,I} and a subset of {I, B, A}. According to this structure, variable A is cause of B, and I functions as exogenous intervention variable with respect to A, which in turn requires a subset relative to which I is causal variable for A and which has B as exogenous intervention variable operating upon I. Finally, we need a subset with B as causal variable for I, with an exogenous variable operating upon B and this role is here played by A. The subset {I, B, A} is simply the set of all variables that function as exogenous variables with respect to one of the proper subsets. In this way, whilst every causal variable requires an exogenous intervention variable, and the latter in turn requires an exogenous intervention variable with respect to its status as causal variable for its intervention target, there is no vicious regress because exogenous variables with respect to a given causal subset can be themselves endogenous variables with respect to other subsets **(see Figure 15).**

(Fig.15)



Now, the above is only by way of illustration, because any viable candidate set will need to include confounding variables. The above is too simple, in that it only contains endogenous causes and effects, and exogenous variables.

For these reasons, we should consider what might happen if we were to include confounders in each subset. For the sake of simplicity, we will work with a model whereby each proper subset comprises three variables: a causal candidate variable, and outcome variable and one confounder. Our (unordered) variable set is: {A, B, C, I}. But as we saw in Section 1, there are constraints on how the proper subsets can be structured. These consist of the Explication Constraint, below, and Woodward constraints on legitimate interventions variables.

Explication Constraint

(EC) Every exogenous variable E, relative to some causal candidate C in subset S, must be an endogenous causal variable for C as outcome, relative to some other subset S^i.

The Explication constraint is required to ensure completeness of a set. Completeness here demands that every causal relation is made explicit, and so every implicit causal relation between an exogenous and endogenous variable must also be represented as an explicit relation between endogenous variables within a distinct subset.

In addition, we also have Woodward's constraints. Variable I is a legitimate exogenous variable for endogenous cause X with respect to outcome Y only if:

1. I causes X;

2. I acts as a switch for all the 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;

3. 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 – X – 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;

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

Given these constraints, we can start to construct our set of causal subsets. The main point to note throughout is just how narrow are the options for subset structure given the above constraints.

We could start with a subset consisting of endogenous variables, A, B, and C, and where A is the causal variable, B the outcome, and C the confounder. We represent this subset as an ordered triple: {A, C, B}. Given that our total set consists of A, B, C, I, and given our exogenous variable constraints, I is the only possible intervention variable for this subset. Given the membership, structure and exogenous variable for our first subset, some of our constraints immediately narrow the options for construction of a second subset.

In accordance with constraints (2), (3), and (4), if I is exogenous variable for our first subset, then I is not correlated with the outcome variable B (independently of a causal path via A) or the confounding variable C. Confounder C is a confounder either because it is a cause of B, independent of A, or because it is a cause of A. Woodward's constraint (4) rules out variable I from being independently correlated with C if C is a cause of B independent of A. Constraint (2) rules out variable I from being a cause of C if C is a cause of A.[87] Given that correlation is a necessary condition on causation, this implies that I is not a causal variable, relative to any subset, for B or for C. Furthermore, C, as confounding variable, is causally related to A or B. These implications, together with others derived from the construction of other subsets, will narrow the options for how further subsets can be populated and structured.

Because the principal concern of the subset model is to secure a set whereby all causal relations are explicit, we start constructing our second subset by ensuring that the causal relation between exogenous variable for the first subset, I, and causal variable for the first

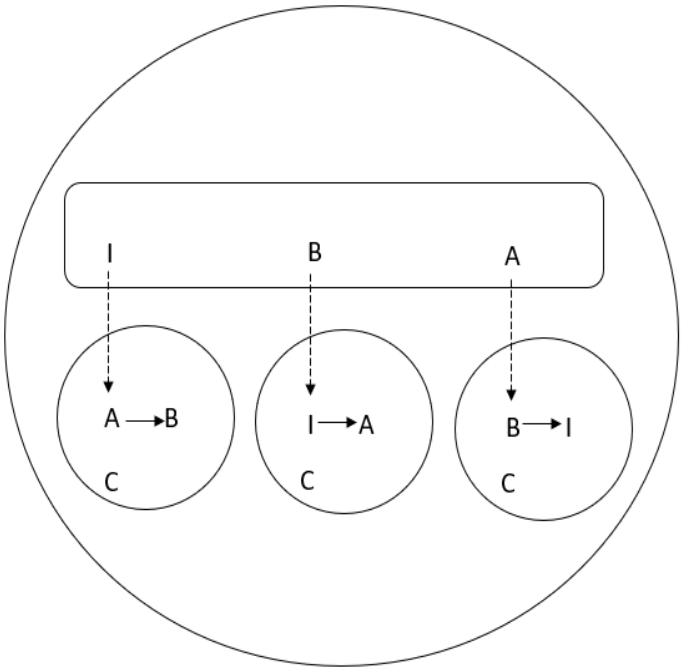
[87] There are broadly two ways in which a variable can qualify as a confounder relative to some candidate causal relation. The rough idea here is that if we have a causal candidate C and an outcome variable O as endogenous variables in a subset, then we can only articulate their causal relation if we hold fixed any other endogenous variables that are also causally related to the outcome variable (independently of the causal candidate), or that are themselves causal variables with respect to our causal candidate C. Any variable that is independently causally related to O will confound articulation of the relation between C and O, and any variable that functions as causal variable for C will confound the relation also. If we don't hold fixed a 'rival' cause of O, then we are in no position to explicate the way in which C makes a difference for O. Similarly for any causal variable with respect to C: if we do not effectively intervene upon C so that any common cause is held fixed, then cannot isolate the way in which O depends upon C as opposed to causes of C.

subset, A, is explicated here. Both I and A are endogenous to our second subset. Our total set of variables is {A, B, C, I}, so we must choose a confounding variable and exogenous variable from B and C. Nothing said so far rules out having C as confounder, and B as exogenous variable. So we can define our second subset as {I, C, A}, with B as intervention variable.

As before, populating the second subset in this way imposes certain constraints on permissible relations between variables for other subsets. As per constraints (2), (3), and (4), B cannot be correlated, independently of a path via I, with A or C. C, as confounder, is causally correlated with I or A. And again, given our constraint (EC), we must explicate the implicit causal relation from B to I, which necessitates another subset.

Our third subset must take B as causal variable and I as outcome. We arbitrarily assign C as our confounder and A as our exogenous variable, producing subset {B, C, I} (A). Once again, this structure in conjunction with our constraints (2), (3), and (4) implies that A cannot be independently causally correlated with I or C, and that C must be causally correlated with B or I. On the basis of our process above, it might look like we have a structured set of proper subsets where every causal relation has been made explicit. In particular, every implicit relation between exogenous variables and their relative endogenous causal candidates has been explicated **(see Figure 16)**. If so, then we have avoided the regress problem and evaded the Incompleteness Argument.

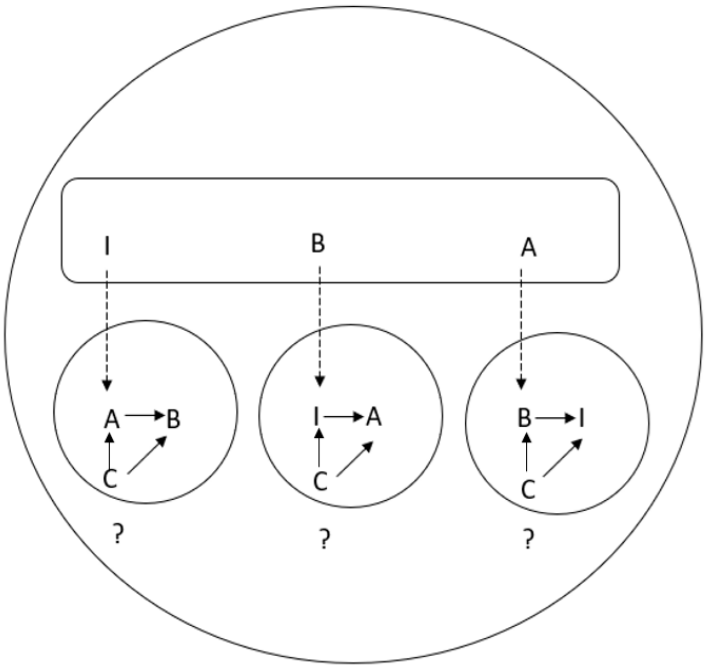
(Fig.16)



But this is all by way of showing just how challenging the construction of the requisite set would be. For the application of our constraints to each subset yielded further causal relations that have not been rendered explicit in our model above. Each confounder variable in each subset must either be a cause of the candidate cause in that subset or a cause of the outcome in that subset. These confounder causal relations have not been explicated in the above.

The options for explicating these are narrow. In the first subset, confounder C is a cause of A or B – this is just what it is to be a confounder. C is a cause of A or I in the second subset and in the third, C is a cause of I or B. So the possible combinations for minimal satisfaction of these disjunctions are: C causes A and B; or C causes A and I; or C causes I and B **(see Figure 17)**.

(Fig. 17)



We'll consider the prospects for modelling subsets, to be added to our existing variable set above, that explicate these causal relations. For a fourth subset that models C's causing A, we'll need either B or I to function as exogenous intervener. A cannot do it, because the exogenous variable cannot be independently causally correlated with the outcome variable. C cannot do it, because the exogenous variable clearly cannot be the same variable as the

endogenous causal candidate. But now we have a problem. B cannot be the intervention variable for this subset, because B is prohibited from causing C. That is due to B's being the intervention variable relative to the second subset; since B, in that capacity, cannot be directedly correlated with the confounder, C, in the second subset, it follows that B cannot be causally correlated with C. But if B cannot be causally related to C, then B cannot intervene upon C and so cannot act as intervention variable relative to C in our fourth subset. This leaves only variable I to function as intervener here. But if I is the intervention variable, then B must be the confounder, since I cannot be both intervener and confounder with respect to the same subset. But the trouble is that B is *also* ruled out from being the confounder. Our second subset had B as intervention variable, A as outcome and C as confounder. B is therefore ruled out from being independently causally correlated with A, and from being causally correlated with C. In which case, we have an incomplete subset – our provisional fourth subset, for explication of C's causing A, lacks a confounding variable.

Having ruled out a causal relation from C to A, two of our options above are also gone. The first option was that C causes A and B; the second, that C causes A – both of these are now eliminated. This leaves only C's causing I and B. As we'll see, this fares no better than the other two.

We will focus on a subset that aspires to explicate C's causing I. For now familiar reasons, A or B must be intervention variable. A cannot be the intervener, because A was the intervention variable relative to our third subset, in which C was the confounder. A cannot be a causal variable for C, and so cannot intervene upon C here. But now we have a problem: B cannot take the role of intervening variable either, for similar reasons to those that rule out A. B was the intervening variable relative to the second subset, where C was confounder. Since an intervener cannot be causally related to a confounder in the relevant subset, that means that B cannot be causally related to C. Consequently, our provisional fourth subset for explication of C's causing of I lacks a viable exogenous variable from our total set, {A, B, C, I}.

The upshot is that none of the options that would vindicate C's role as confounding variable in the first three subsets is explicable in terms of our limited set of variables, {A, B, C, I}. Given that these subsets are only legitimate if they contain a confounder, it follows that these fail to legitimately articulate their candidate causal relations.

The Subset Model: Summary

Two salient points emerge from our investigation of the subset strategy. First, prospects look bleak for successfully navigating the regress problem and hence blocking the Incompleteness Argument. The combined constraints associated with legitimacy of intervention and confounding variables block certain combinations of subsets whilst the need to explicate all causal relations, including those implied by assignment of exogenous interveners, and the causal implications of a variable's having confounding status, imposes a need for suitably rich combinations of subsets. Whilst I have not shown that no such combination is possible, nor even that no such combination is possible for a minimally populated set such as {A, B, C, I}, I take it that our discussion has suggested that the prospects are not promising.

The second lesson is that, even if some combination of subsets of some set of variables *is* able to explicate all causal relations, this is a purely formal achievement. The limitations on the membership and structure of our subsets were logically derived from the theoretic constraints together with the need to explicate all implicit causal relations between exogenous and endogenous variables. But this leaves as an open question whether such a set is viable as a structure for explicating the causal relations that in fact obtain. For although a coherent subset model will technically block the argument, it will only plausibly defeat the central worry insofar as it is a *plausible* account of a prospective set of causal principles. For whilst a coherent subset model will prevent the conclusion from following from the premises, unless the model is plausibly fit for purpose in capturing the structure of physical causal principles and their interrelationship, we lack independent motivation for accepting the model. If the logical space available for such models is small, then this places constraints on which variables can be causally related. So it remains to be seen whether such models are viable candidates for explication of bona fide causal relations.

Given the above, we should take the Incompleteness Argument seriously. The potential objection is not sufficiently compelling, either as a means of blocking the argument or as a means of addressing the underlying worry about an Interventionist complete causal set.

Section 3: The Limits of Interventionist Causation

However, there is available a further potential response to the argument, adapted from remarks by Woodward concerning another problem. Broadly, the response is to claim that a complete set of explicit causal variables is possible, because not all explicit causal relations require exogenous variables for their explication. We here want to show that *whether the response is successful or not, there is a tension internal to the Interventionist conception of causation as applied to physics.* Our speculative prophylactic suggestion is that the tension might be relieved by dropping Interventionism as a conception of causation for the physical domain. On the assumption that Interventionism is suitable for explication of psychological causal relations, this entails a recommendation of causal pluralism.

Section 3a: Woodward's Strategy

Woodward (2003, 127 – 133) suggests a strategy for dealing with the problem of physically impossible interventions, which might appear to provide the resources with which to defuse the Incompleteness Argument.

On Woodward's understanding, an intervention I upon a candidate causal variable is physically possible if and only if there is some set of possible initial conditions, perhaps different from those that actually obtain, such that the occurrence of I is consistent with those conditions and the laws of nature. By 'possible initial conditions', Woodward means conditions that are not inconsistent with the laws of nature. The thrust of Woodward's definition of physical possibility is that, even if determinism holds, an event can be physically possible just as long as it's occurrence would not violate any physical law. The clause, 'perhaps different from those that actually obtain', secures the compatibility of this notion of physical possibility with determinism: regardless of the *actual* preceding events, just as long as event E has some initial conditions in accordance with physical laws and is itself consistent with both those conditions and those laws, then E could occur. For example, in a pharmaceutical trial, and assuming determinism, it is physically possible to have intervened and administered treatments to subjects who did not actually receive it (Woodward, 2003, p.128).

Such a definition of course also delivers a sufficient condition for physical *impossibility*: an event E is physically impossible if: (i) it lacks preceding conditions such that those conditions

are consistent with physical laws and are consistent with an occurrence of E or (ii) it is itself inconsistent with physical laws. By way of illustration, acceleration of a particle from a velocity less than that of light to a velocity greater than that of light would be physically impossible in the above sense (Woodward, 2003, p.128).

The potential trouble for prospective interventions comes from two directions: (a) from the constraints of physical possibility, as defined above, and (b) from the constraints governing legitimate intervention variables, as defined in my earlier discussion. Both the physical possibility constraints and the intervention constraints narrow the options for bringing about the desired value in the causal candidate variable. Since I only raise this issue by way of introducing a potential strategy for assuaging the worries brought on by the Incompleteness Argument, I shall offer only a brief example (taken from Woodward, 2003, p.129). Consider the following statement:

'Changes in position of the moon with respect to the earth and corresponding changes in the gravitational attraction exerted by the moon on various points on the earth's surface cause changes in the moon's tides'.

We assume that this statement is true. But in cashing this out in interventionist terms, we need to specify an intervention variable for changing the position of the moon relative to the earth. As Woodward writes, "Here the constraint of physical possibility means, for example, that we are not allowed to imagine the moon moving from its present position to a new position at a superluminal velocity or for the moon to change its orbit except under the influence of some impressed force of appropriate direction and magnitude, where the force in question must obey the usual source laws and general laws of motion such as $F = ma$" (2003, p.128). But the problem is that there may be no such intervention that obeys both the physical constraint outlined here and the constraints given by the legitimacy conditions for intervention variables with respect to their relevant variable set. For example, one of the constraints on exogenous variables was that they are not independently correlated with the outcome variable of that subset. This means that we cannot bring about a change in the moon's orbit of the earth by positioning a massive body such that the gravitational force of this body changes the moon's position, because this force would also have a direct and independent effect on the tides. Other imagined processes appear to suffer from similar

problems; insofar as they are physically possible, and thus do not violate the laws of physics, they risk contravening one or more intervention constraints. The laws of physics seem to imply further, independent effects of the intervention variable upon either the outcome variable or upon confounding variables.

Our point here is not to establish that there are plausible cases of causation that lack physically possible intervention variables. Rather, we only want to draw attention to Woodward's response to this potential difficulty. Having reminded us that the technical notion of an intervention, expressed here via our associated constraints (I) through (IV) and including the constraint that the intervention must itself be a causal variable with respect to its target, is a *regulative ideal* for understanding what we mean and are interested in when talking of causal relations, Woodward (2003, p.130) says the following: "…as long as there is some basis for assessing the truth of counterfactual claims concerning what would happen if various interventions were to occur, it doesn't matter that it may not be physically possible for those interventions to occur."

We need, therefore, a means of assessing interventionist counterfactuals that does not require the positing of possible physical interventions upon the target causal variables. Woodward suggests that, in the case described above, Newtonian gravitational theory and mechanics enable assessment of the relevant counterfactuals. Although it might be physically impossible to engineer the position of the moon without thereby determining, independently, effects upon the tides we can, on this view, effectively 'subtract' the latter from the hypothetical process and hence arrive at a truth value for the appropriate counterfactuals. The mathematical precision afforded by Newtonian theory facilitates the postulation of values for the target candidate cause (the position of the moon relative to the earth) that affords us the precision required by our notion of an intervention: a means of assigning values for the causal variable along one dimension but not others, whilst holding fixed the values of confounding variables and not independently affecting any assignment of values for the outcome variable. And the correlations formalized by the relevant equations permit an assessment of the correlations 'under intervention' expressed by the necessary counterfactuals.

For the sake of argument, we will accept that this strategy may work for some cases. Our question is: can this strategy be applied so as to evade Campbell's argument, and to enable a complete set of explicit causal relations within an interventionist framework?

Let us remind ourselves of the argument:

1) A variable set is causally complete if and only if it includes all physical variables that are explicitly causally related under interventions.
2) For any causal variable set V, and any member endogenous variables X and Y, X is only explicitly causally related to Y if there is some exogenous variable I to intervene upon X.
3) For any causal variable set V, if I is an exogenous variable with respect to endogenous variables in V, then I is not itself a member of V.
4) For any causal variable set V, V is not a causally complete variable set.

I should clarify that the suggestion is *not* that the argument turns on the physical impossibility of a complete set in the sense of physical impossibility outlined above. But it seems reasonable to take Woodward's broad response – i.e. that interventions are a regulative *ideal* for our *conception* of causation and are not always required for actual assessment of the relevant counterfactuals – as a potential route to responding to the argument. Woodward's problem, to which this strategy was applied as solution, was that on the one hand, the claim concerning the causal relation between changes in the position of the moon relative to the earth and changes in the tides seems right, but on the other, there is no (physically) possible intervention variable by which to explicate the causal relation in interventionist terms. Analogously, here we have a case where, on the one hand, it seems intuitively plausible that there could be a complete set of explicit causal physical variables, but on the other, there seems no way that such a set could have the exogenous variables required to function as interventions. On the basis of this analogy between the problems, we might suspect that the strategy for dealing with one will be available for dealing with the other.

How might such a response look? It appears that the strategy outlined might work to block the argument by undermining premise (2). If there are permissible cases in which endogenous causal candidates lack intervention variables but are explicable as causal variables by some other means, then it seems that the availability of an exogenous intervention is not a necessary condition on causal relations. The strategy would be to undermine plausibility of premise (2) by showing the plausible means by which to assess the relevant interventionist

counterfactuals where no exogenous variable is available. Presumably, as with the moon and tides case, this would involve appealing to a formalized, precise expression of correlative dependency that allows us to assign values along one dimension and not others, in such a way as to avoid thereby assigning values to confounding variables, and where we can know the mathematical implications of that assignation for the outcome variable. The upshot would be that there would be no exogenous variables logically excluded from the set, since exogenous variables are not required for legitimate explication of causal relations between variables endogenous to the set. With no excluded but causally efficacious exogenous variables, there are no implied causal relations between them and their causal candidate targets to be articulated by placing them in a (necessarily distinct) set of endogenous variables.

Section 3b: Causal Completeness & Pluralism

I do not here offer direct assessment of the above strategy; I do not need to, because my argument takes the form of a dilemma for a certain kind of Interventionist. Recall that we are concerned with addressing Exclusion -related issues from the perspective of the non-reductive physicalist sympathetic to the Exclusion principles: Supervenience, Distinctness, Causation, Completeness, and Exclusion. Accordingly, we are here concerned with addressing Interventionists of the same kind.

The dilemma is this: If our Interventionist deploys Woodward's strategy to evade the Incompleteness problem, then she implies a tension internal to her conception of causation for physics; if she does not, and accepts the Incompleteness problem, then that internal tension is evident anyway.

The key principle in motivating these claims will be the Completeness principle: all physical effects have sufficient physical causes. To the extent that our Interventionist is committed to Completeness, I take it that she is liable to the above dilemma.

We will start by unpacking the first horn. Why think that Woodward's strategy implies a tension within Interventionism? The broad idea is that his strategy involves outsourcing the job of explicating causal relations to the laws of physics, and those laws – if causal Completeness holds – are exceptionless laws. But if the laws of physics are exceptionless, then this supports a notion of causation that is not Interventionist, but rather, a conception

of nomological regularity. So the proponent of causal Completeness deploys Woodward's strategy at risk of undermining Interventionism as a conception of causation for physics.

We saw above that Woodward's strategy, as applied to the Incompleteness problem, would be to use the equations of physics to plug in values for causal candidates and calculate the correlative values taken by the outcome variables. This would potentially enable evaluation of Interventionist counterfactuals required to answer the causal question: what would happen if we were to intervene upon the causal candidate with respect to the outcome variable? Crucially, it would potentially enable that evaluation in the absence of any available intervention variable. So in deploying the strategy, the Interventionist is relying upon the laws of physics for their evaluation of the counterfactuals.

Now, there are good grounds for taking the laws of physics to be exceptionless laws. A commitment to this is arguably evident in the practice of physicists, characterized as it is by, amongst other things, a process of attempted disconfirmation. But most importantly here, it is plausible that laws as exceptionless are entailed by the Completeness principle itself. For according to Completeness, no physical effect lacks a sufficient physical cause. The truth of this arguably requires a set of exceptionless laws. If the laws of physics are not exceptionless, then they are compatible with the physical effects individuated by those laws having causes that do not figure in those laws. This seems to be incompatible with causal Completeness; so Completeness plausibly entails that the laws of physics are exceptionless.

But if the laws of physics are exceptionless, then this suggests a conception of physical causation that is at odds with that of Interventionism. For it suggests that physical causal relations are causal in the sense of being regularly conjoined, where the regularity is nomological. It is the exceptionless nature of the laws that underpins the regularity of the causal relations covered by them. This is not the conception of causation at work in Interventionism, where such regularity is not necessary for a relation to be causal; on the contrary, what is required are that causal and outcome variables are correlated under intervention. But such correlation can be *ceteris paribus*: there is nothing internal to the Interventionist conception requiring causal correlations to be exceptionless. If so, then we have grounds for thinking that our Interventionist, committed to causal Completeness, deploys Woodward's strategy at the risk of undermining Interventionism as a theory of causation for physics. This is the first horn of the dilemma.

The second horn should already be evident. If our Interventionist accepts the Incompleteness Argument, then she accepts that Interventionism cannot provide a complete set of explicit causal relations for physics. One response to that is to bite the bullet and scale back the unificationist hopes of physics (see Campbell, 2020, Chapter Four). But this is, to say the least, an uncomfortable position for the Interventionist committed to causal Completeness, for that principle is about physical causation. It is not clear how one might reconcile that principle with the claim that there is no complete set of explicit causal relations. I therefore suggest that our Interventionist would find their conception of causation challenged as regards causal relations in physics. That is the second horn of the dilemma.

In the face of the dilemma, what should she do? Our speculative suggestion is this: endorse causal pluralism. Assuming that Interventionism is a suitable conception for psychology, but accepting that it is not suitable for physics – given a commitment to causal Completeness – one option for relieving the tension within an Interventionist notion of causation for physics is to drop that notion. In so doing, the Interventionist need not worry about the Incompleteness problem, and need not fret about Woodward's strategy. If pluralism is coherent, then she is free to accept nomological regularity as causation in physics, and interventionist correlation as causation in psychology.

Section 4: Exclusion – Diagnosis & Cure

We have now seen a tension that results from the combination of causal Completeness and Interventionism about physics. My speculative suggestion was that the Interventionist respond by embracing pluralism. But at the outset of this chapter, I said I would potentially shed light on why some might find Exclusion worries lingering even for cases of non-supervenience-based causation, such as Hadron Collider. I also said I would suggest a means of alleviating those worries.

I am now in a position to do so. My speculative diagnostic suggestion is that the same factors giving rise to the tension outlined above are responsible for lingering Exclusion worries. On my picture, those worries relate to cases of direct diagonal causation where causal Completeness entails sufficient physical causes for physical effects. It is therefore natural to suggest that those worries are predicated upon a single, over-extended conception of causation that yields homogenous diagonal and physical causal relations. Crane (1995) claims

that exclusion-based arguments for physicalism must assume homogeneity of mental and physical causal relations. I here borrow this line of thought. If one assumes that the causal relations obtaining between physical properties are the same kind of relation as those that obtain diagonally, and one assumes causal Completeness for the physical properties, then one might worry that the guaranteed sufficiency of the physical threatens to squeeze out mental causes. Such a worry might arise even in cases where systematic, supervenience-based causation is absent. But on a pluralistic picture, the threat loses its force. For if Completeness only applies to physical causation, and such causation is of a different kind to diagonal, mental causation, then any intuition of competing causes fades. Pluralism thus offers both potential diagnosis and cure for lingering Exclusion worries.

Conclusion

The central aim of this chapter has been to offer a rough, schematic recommendation to those non-reductive physicalists sympathetic to Interventionism: causal pluralism. Such pluralism involves conceiving of physical causal relations under a nomological regularity conception, and higher-level relations via the Interventionist conception. I provisionally recommend pluralism as a means of avoiding the dilemma posed by the Incompleteness Problem, and as a means of diagnosing and dispelling any lingering worries relating to diagonal causation.

I suggested that if the Interventionist accepts the Incompleteness Problem, then she thereby accepts a limitation on her conception of physical causation. But if she deploys Woodward's strategy, she introduces a tension into that conception. Pluralism acknowledges the limitation and scales back the intended scope of Interventionist causation. Alternatively, it ameliorates the tension.

I also suggested that some might feel uneasy, even in cases of non-supervenience-based causation, about mental diagonal causation. Insofar as that concern is based upon Completeness, i.e. about inevitably sufficient physical causes precluding mental efficacy, pluralism might serve as a diagnosis and cure. If diagonal mental causation is properly conceived under one conception of causation, and physical causation under another, then

perhaps a monistic notion of causation is to blame for an erroneous sense of competition between the two. Perhaps we might dispel such lingering unease with a pluralism that makes room for harmonious convergence.

Conclusion

With this thesis, I hope to have shown that mental causation, even mental causation of physical effects, is not undermined by the complete causal sufficiency of the physical. I have argued that our picture of psychology and physics as autonomous is well-grounded, despite challenges to that picture from philosophy.

The predominant challenge took the form of the Exclusion Problem. I hope to have persuaded that non-reductive physicalists can endorse the core principles of the Exclusion Problem whilst not thereby committing to systematic – and so problematic – causal overdetermination. For I have claimed that those core principles only entail systematic overdetermination if a host of other assumptions, e.g. the denial of close multiple realisability and the radical locality of supervenience, are also adopted. Such assumptions are problematic for non-reductive physicalists, and so the conventional view – that the Exclusion Problem presses such physicalists to respond with elaborate solutions – is mistaken. The crux of my claims here was that supervenience played an essential role in ostensibly generating cases of diagonal causation between levels. But supervenience, absent the assumptions above, does not do this. The alleged challenge to mental causal autonomy was based upon an overly general view of supervenience.

In the rest of the thesis, I have attempted to warn against overgeneralization in other respects. In Chapter 3, I showed that evading the Exclusion Problem should not push us into assuming that diagonal mental causation does not, or cannot, occur. Whilst we can consistently assert the core Exclusion principles – including the principle on which mental causes are genuinely efficacious – and deny systematic overdetermination, this is itself consistent with permitting cases of diagonal mental causation that do not demonstrate systematicity. I argued in support of this by focusing upon the Hadron Collider case, which I claimed to be a case of *direct* mental causation which thereby avoids intimating systematicity of overdetermination.

We also saw overgeneralization at work in Campbell's solution to the Exclusion Problem (2020). Here, a central principle of the solution, Leanness, was shown to be predicated upon worries that do not arise in Exclusion contexts. I offered a diagnosis of the picture on which

such worries might be thought to arise, which then extended to a diagnosis of the picture on which Exclusion worries more generally might originate. That picture was a conflation of what I called the 'micro-picture' and the 'neuro-picture', with physical properties unconsciously mixed with neurophysiological.

Finally, I have offered, if only schematically, a diagnosis and cure for any persistent worries about overdetermination in non-systematic cases. Interventionists about psychology should adopt causal pluralism with respect to physics, for which a nomological regularity conception of causation might plausibly be better suited. I claimed that adopting causal pluralism here perform a dual function: it offers a way out of a dilemma posed by Campbell's Incompleteness Argument (2020) and a means of ameliorating worries that causal Completeness might somehow push diagonal mental causes out of the running.

This last proposal was speculative, leaving much room for elaboration and argument. But I hope to have shown that the Interventionist, sympathetic to Completeness, does have a problem here, and that causal pluralism is a plausible solution.

In general, my aim has been to elucidate the ways in which philosophy has made problems for mental causation, and the ways in which philosophy can expel those problems. I have attempted to show that psychology, despite minds supervening upon the physical, retains its autonomy in the face of a causally complete physics.

Bibliography

Armstrong, David M. 1980. “The Nature of Mind.” In *The Nature of Mind and Other Essays,* edited by David M. Armstrong, 16-31. Itacha: Cornell University Press.

Baumgartner, Michael. 2009. “Interventionist Causal Exclusion and Non-reductive Physicalism.” *International Studies in the Philosophy of Science* 23, no.2: 161-178.

Bechtel, William and Jennifer Mundale. 1999. “Multiple Realizability Revisited: Linking Cognitive and Neural States.” *Philosophy of Science* 66, no.2: 175-207.

Bennett, Karen. 2003. “Why the Exclusion Problem Seems Intractable, and How, Just Maybe, to Tract It.” *Noûs* 37, no.3: 471-497.

Bernstein, Sara. 2016. “Overdetermination Underdetermined.” *Erkenntnis* 81, no.1: 17-40.

Boghossian, Paul A. 1997. “What the Externalist Can Know A Priori.” *Proceedings of the Aristotelian Society* 97, no.1: 161-176.

Burge, Tyler. 1979. “Individualism and the Mental.” *Midwest Studies in Philosophy* 4, no.1: 73–121.

Campbell, John. 2006. “An Interventionist Approach to Causation in Psychology.” In *Causal Learning: Psychology, Philosophy and Computation* edited by Alison Gopnik and Larry J. Schulz, 58-66. Oxford: Oxford University Press.

_____ 2020. *Causation in Psychology.* Harvard: Harvard University Press.

Chalmers, David J. 1996. *The Conscious Mind: In Search of a Fundamental Theory.* Oxford: Oxford University Press.

_____2000. “What is Neural Correlate of Consciousness?” In *Neural Correlates of Consciousness* edited by Thomas Metzinger, 17-39, Cambridge MA: MIT Press.

Cox, Edward T. 2008. “Crimson Brain, Red Mind: Yablo on Mental Causation.” *Dialectica* 62, no.1: 77-99.

Crane, Tim. 1991. “All The Difference in the World.” *The Philosophical Quarterly* 41, no.162: 1-25.

_____1995. “The Mental Causation Debate.” *Aristotelian Society Supplementary* 69, (Supplementary): 211-36.

Davidson, Donald. 1970. “Mental Events.” In *Experience and Theory* edited by Lawrence Foster and J. W. Swanson, 207-224, Oxford: Clarendon Press.

_____1987. “Knowing One’s Own Mind.” *Proceedings and Addresses of the American Philosophical Association* 60, no.3: 441–458.

Ehring, Douglas. 1996. “Mental Causation, Determinables and Property Instances.”*Noûs* 30, no.4: 461–480

Farkas, Katalin. 2008. “Phenomenal Intentionality Without Compromise.” *The Monist* 91, no.2: 273-93.

Fodor, Jerry. 1974. “The Disunity of Science as a Working Hypothesis.” *Synthese* 28, no.2: 97-115.

Funkhouser, Eric. 2002. “Three Varieties of Causal Determination.” *Pacific Philosophical Quarterly* 83, no. 4: 335-351.

_____2006. “The Determinable-Determinate Relation.” *Noûs* 40,no.3: 548–569.

Georgalis, Nicholas. 1999. “Rethinking Burge’s Thought Experiment.”*Synthese* 118, no.2: 145–164.

Gertler, Brie. 2012. “Understanding the Internalism-Externalism Debate: What Is the Boundary of the Thinker?” *Philosophical Perspectives* 26, no.1: 51–75.

Gibbons, John. 2006. “Mental Causation Without Downward Causation.” *The Philosophical Review* 115, no.1: 79-103.

Heil, John. 2003. “Level of Reality.” *Ratio* 16, no. 3: 205-221.

Hitchcock, Christopher. 2007. “What Russell got right.” In *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, edited by Huw Price and Richard Corry, 45-65, Oxford: Oxford University Press.

Jackson, Frank and Philip Pettit. 1988. “Functionalism and Broad Content.” *Mind* 97, no.387: 381-400.

Jackson, Frank and Philip Pettit. 1990.“Program Explanation: A General Perspective.” *Analysis* 50, no.2: 107–17.

Jackson, Frank. 1998. *From Metaphysics to Ethics: A Defence of Conceptual Analysis.* Oxford: Oxford University Press.

Kane, Robert. 1996. “Freedom, Responsibility, and Will-Setting.” *Philosophical Topics* 24, no.2: 67-90.

Kim, Jaegwon. 1976. “Events as Property Exemplifications.” In *Supervenience and Mind* edited by Ernest Sosa, 33-52, Cambridge: Cambridge University Press.

_____1987. “Strong and Global Supervenience Revisited.” *Philosophy and Phenomenological Research* 48, no. 2: 315-326.

_____1993a. *Supervenience and Mind: Selected Philosophical Essays.* Cambridge: Cambridge University Press.

_____2003. “Blocking Causal Drainage and Other Maintenance Chores with Mental Causation.” *Philosophical and Phenomenological Research* 67, no.1: 151-176.

_____2005. *Physicalism, or Something Near Enough*. Princeton: Princeton University Press.

Lewis, David. 1966. “An Argument for the Identity Theory.” *Journal of Philosophy* 63, no.1: 17-25.

Loar, Brian. 1988. “Two Kinds of Content.” In *Contents of Thought* edited by Robert H. Grimm and Daniel D. Merrill, 121-139, Arizona: University of Arizona Press.

_____2003. “Phenomenal Intentionality as the Basis of Mental Content.” In *Reflections and Replies: Essays on the Philosophy of Tyler Burge,* 229-258, MIT: MIT Press.

McLaughlin, Brian P. 1992. “On Davidson’s Response to the Charge of Epiphenomenalism.” In *Mental Causation* edited by John Heil and Alfred Mele, Oxford: Oxford University Press.

Marras, Ausonio. 1998. “Kim’s Principle of Explanatory Exclusion.” *Australasian Journal of Philosophy* 76, no.3: 439-451.

Ney, Alyssa. 2008. “Physicalism as an Attitude.” *Philosophical Studies* 138, no.1: 1-15.

Papineau, David. 2000. “The rise of physicalism.” In *Physicalism and Its Discontents* edited by Carl Gillett and Barry M. Loewer, 3-36, Cambridge: Cambridge University Press.

Polger, Thomas W. 2002. “Putnam's intuition.” *Philosophical Studies* 109, no. 2: 143-70.

Polger, Thomas W. and Lawrence A. Shapiro. 2016. *The Multiple Realization Book*. Oxford: Oxford University Press.

Post, John. 1991. *Metaphysics: A Contemporary Introduction.* New York: Paragon House.

Putnam, Hilary. 1967. “*Psychological* Predicates.” In *Art, Mind, and Religion* edited by W.H. Capitan and Daniel D. Merrill, 37-48, Pittsburgh: Pittsburgh University Press.

_____1975. *Mind, Language and Reality: Philosophical Papers*. Cambridge: Cambridge University Press.

Searle, John. 1983. *Intentionality: An Essay in the Philosophy of Mind.* Cambridge: Cambridge University Press.

Segal, Gabriel. 2000. *A Slim Book about Narrow Content*. Cambridge: MIT Press.

Shoemaker, Sydney. 2001. “Realization and mental causation.” In *The Proceedings of the Twentieth World Congress of Philosophy* edited by Carl Gillett and Barry M. Loewer, 23-33, Cambridge: Cambridge University Press.

_____2007. *Physical Realization.* Oxford: Oxford University Press.

Strawson, Peter F. 1962. “Freedom and Resentment.” *Proceedings of the British Academy* 48:187-211.

Thomasson, Amie. 1998. “A Nonreductivist Solution to Mental Causation.” *Philosophical Studies* 89, no.2-3: 181-195.

Van Frassen, Bas. 2002. *The Empirical Stance.* New Haven, CT: Yale University Press.

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

Woodward, James. 2015. “Interventionism and Causal Exclusion.” *Philosophy and Phenomenological Research* 91, no.2: 303-347.

Wilson, Jessica. 2009. “Determination, realization and mental causation.” *Philosophical Studies* 145, no.1: 149-169.

Wilson, Jessica. 2011. “Non-reductive realization and the powers-based subset strategy.” *The Monist* 94, no.1: 121-154.

Yablo, Stephen. 1992. “Mental causation.” *Philosophical Review* 101, no.2: 245-280.