



UMEÅ UNIVERSITY

Causal After All

A Model of Mental Causation for Dualists

Bram Vaassen

Department of Historical, Philosophical and Religious Studies

Umeå 2019

Umeå Studies in Philosophy 13

© Bram Vaassen

Series Editor: Pär Sundström
Department of Historical, Philosophical and Religious Studies
Umeå University
SE-901 87 Umeå, Sweden

Dissertation for PhD
This work is protected by the Swedish Copyright Legislation (Act 1960:729)
ISBN: 978-91-7855-098-2
ISSN: 1650-1748
Cover: *Objet Trouvé 3.0: Lascaux*, by Tom Swaak
Digital version available online at <http://umu.diva-portal.org/>
Printed by: UmU Print Service, Umeå University
Umeå, Sweden 2019

For Jasper

Abstract

In this dissertation, I develop and defend a model of causation that allows for dualist mental causation in worlds where the physical domain is physically complete.

In Part I, I present the dualist ontology that will be assumed throughout the thesis and identify two challenges for models of mental causation within such an ontology: the exclusion worry and the common cause worry. I also argue that a proper response to these challenges requires a thoroughly lightweight account of causation, i.e. an account that allows for causes to be metaphysically distinct from the phenomena that produce or physically necessitate their effects.

In Part II, I critically evaluate contemporary responses to these challenges from the philosophical literature. In particular, I discuss (i) List and Stoljar's criticism of exclusion worries, (ii) Kroedel's alternative dualist ontology, (iii) concerns about the notion of causal sufficiency, and (iv) Lowe's models of dualist mental causation. I argue that none of these proposals provide independent motivation for a thoroughly lightweight account of causation and therefore leave room for improvement.

In the first four chapters of Part III, I develop my thoroughly lightweight model of causation, which builds on interventionist approaches to causation. First, I explain how so-called 'holding fixed'-requirements in standard interventionist accounts stand in the way of dualist mental causation. I then argue that interventionist accounts should impose a robustness condition on causal correlations and that, with this condition in place, the 'holding fixed'-requirements can be weakened such that they do allow for dualist mental causation. I dub the interventionist model with such weakened 'holding fixed'-requirements 'insensitive interventionism', argue that it can counter the exclusion argument as well as the common cause worry, and explain under which circumstances it would predict there to be dualist mental causation. Importantly, these circumstances might, for all we know, hold in the actual world.

In the final three chapters of Part III, I defend insensitive interventionism against some objections. I consider the objection that causation must be productive, the objection that causes must (in some sense) physically necessitate

their effects, and the objection that insensitive interventionism is too permissive. I respond by drawing from the literature on causation by absences and on the relation between causation and fundamental physics. Overall, insensitive interventionism performs as well as standard interventionist accounts. I conclude that insensitive interventionism is a credible model of causation.

The upshot is that the standard position that dualists cannot have mental causation in worlds where the physical domain is complete is undermotivated, and perhaps even false.

Keywords: Mental Causation, Dualism, Non-Reductionism, Causal Exclusion, Causation, Interventionism, Negative Causation, Omissions, Neo-Russellianism, Causation and Physics

Contents

Acknowledgements	xi
1 Introduction	1
I The Set-Up	13
2 Nomic Naturalist Dualism	15
2.1 Dualism	15
2.2 Naturalist Dualism	21
2.3 Nomic Naturalist Dualism	25
3 Mental Causation	31
3.1 The exclusion worry	33
3.2 The common cause worry	40
3.3 The set-up	43
4 A Glimpse of an Answer	45
4.1 Heavyweight causation	46
4.2 (Thoroughly) lightweight causation	51
4.3 Dualist mental causation	55
II Contemporary Solutions	59
5 On Causal Exclusion	61
5.1 Exclusion and metaphysical distinctness	62
5.2 Metaphysically distinct co-causes	64

5.3	Exclusion worries persist	68
5.4	Weak Exclusion	70
6	Supernomological Dualism	77
6.1	Strengthening the laws	77
6.2	Against <i>Weak Exclusion</i>	81
6.3	The common cause worry	82
6.4	Evaluating Kroedel's proposal	84
7	On Causal Sufficiency	85
7.1	Causal exclusion and sufficiency	86
7.2	Against sufficient causes	87
7.3	Exclusion without causal sufficiency	90
7.4	Exclusion again	94
8	Lowe's Models of Mental Causation	99
8.1	Enabling causes	101
8.2	Diachronic mediation	105
8.3	Synchronic mediation	108
8.4	Evaluating Lowe's models	111
III	Dualist Mental Causation	113
9	Minimal Interventionism	115
9.1	Interventionist causation	116
9.2	Intervention variables	123
9.3	Interventions and actual manipulations	128
9.4	Summary	132
10	Interventionism and Non-Reductionism	133
10.1	Interventionist exclusion	133
10.2	Drainage	136
10.3	(M*) and (IV*)	141
10.4	Summary	146
11	Spurious Higher-Level Causation	149

11.1 The problem	150
11.2 The screening off pattern	153
11.3 The robustness solution	164
11.4 The upshot	169
12 Interventionism for Dualists	173
12.1 Insensitive interventionism	174
12.2 Dualist mental causation	177
12.3 Looking forward	186
13 Negative Causation	189
13.1 Negative causation and mental causation	190
13.2 Spurious negative causation	196
13.3 Replies to spurious negative causation	199
13.4 The upshot	217
14 Causation and Physics	223
14.1 Mismatches between causation and physics	224
14.2 The Neo-Russellian project	232
14.3 Spurious backwards causation	239
14.4 The upshot	246
15 Objections and Replies	249
15.1 Physical equivalence	249
15.2 Gerrymandering	256
15.3 On the Woodward-Baumgartner debate	257
15.4 Conclusion	259
16 Conclusion	261
Sammanfattning	289

Acknowledgements

Philosophy is good stuff, but writing a dissertation can be torture. Even so, I am happy that I decided to move to a place that I never heard of and write on a topic that I mistakenly considered to have understood. I would like to take the opportunity to thank all those who made this writing process so bearable.

First and foremost, I am immensely grateful and equally indebted to my supervisors. Pär Sundström was my main supervisor for this entire project and his support is what kept me going throughout these five years. The intellectual honesty that pervades his way of doing philosophy as well as his personality has been an invaluable resource to me. Among many other things, Pär has taught me that good philosophy requires much more patience and attention than I was used dedicating to anything, and that efforts to clean up one's thinking are never a waste of time. Gunnar Björnsson has co-supervised my development as a PhD student from day one and provided insightful comments on the many drafts of this thesis and on my attempts at writing articles. Andreas Stokke co-supervised the first two years of this project and taught me that my English is not as good as I thought it was and that one should not try to sound philosophical when writing philosophical texts. Torfinn Huvenes joined the Umeå department halfway through my PhD employment and I was lucky enough to have him assigned as my co-supervisor. Torfinn not only read through innumerable drafts of whatever I was writing, he was also a real time sounding board for many of my ideas as they developed. At several occasions Torfinn made it clear that there was serious trouble when I thought I was on solid ground, but just as often he explained why some philosophical problems are not mine to worry about.

Either way, talking to Torfinn always provided me with a clear outline of what I had to do. All in all, I have never met someone who was happier with their supervision team than I am.

Aside from my supervisors, there are several others who have commented on this text or have otherwise directly contributed to its development. Thomas Kroedel was my midseminar opponent and Mathias Frisch was my endseminar opponent. Both took effort to examine the text closely and their remarks have improved the text considerably. Alison Fernandes commented on the paper that laid the groundwork for Chapter 5 and Christian Loew commented on Chapter 14. Together with Thomas Blanchard, Alison and Christian answered a host of questions I had about the relation between causation and physics. I still feel dizzy when I think about it too much, but less so than I used to. Discussions with Ethan Nowak and Alexander Sandgren helped me polish Chapter 13. Jan Doumont sat through a lengthy Skype talk to patiently explain the secondary school physics I required for Section 15.1 and proofread that section afterwards. Further, I benefited greatly from all the thoughtful questions I received when presenting (parts of) this dissertation and from the informal discussions that followed or preceded such events. For fear of leaving anyone out, I cannot risk composing a list of those who corrected my thinking in the past five years.

It is one thing to formulate ideas and arguments and quite another thing to make them fit into a PDF file that can be printed into a book. Johan Junkka and Torfin Huvenes provided me with indispensable L^AT_EX formatting help. The cover is one of the many excellent designs by Tom Swaak. Tom also proofread well over 100 pages and continually provided feedback on language and layout issues. Other proofreaders whose benevolence I have shamelessly exploited in the past few weeks are Vera, Daniel, Moa, Jenny, Rowan and Wouter.

The Umeå department for historical, philosophical and religious studies proved to be an excellent working environment. I owe much to all those who contribute to such warm atmosphere. I am especially grateful to the philosophers whose time here overlapped with mine: Frans, Jan-Willem, Jessica, Kalle, Peter, Bouke, Alex, Ethan, Madeleine, Daniela, Lars S., Lars L., Julia, Erik, Sebastian, Petra, Emma, Bertil, and Jonas. The administrators

have patiently put up with my questions and misunderstandings. In particular, Linda, Fredrik, Kicki, Marie, Maria and Ulla-Stina have helped me out at several occasions. Thanks also to Peter Lindström, for making sure that all the formalia that come with a PhD education are taken care of. Before coming to Umeå, I started my studies at KU Leuven and I am still indebted to many philosophers who helped me in those early stages. In particular, I'd like to thank Markus Eronen and Harmen Ghijsen. Both went well beyond what was formally required of them to help me move forward. Without their help, I would not have gotten this position in the first place.

I spent the spring semester of 2016 at UC Berkeley. I wish to thank John Campbell for writing my invitation letter. John advised me throughout my stay and set aside a weekly time slot for inspiring discussions. His daring style of philosophizing and his encouragement incited me to double down on the seemingly ludicrous idea that dualists can allow for mental causation. I would also like to thank Hannah Ginsborg and John MacFarlane, for allowing me to participate in their seminars. Finally, I am grateful to Grayson Abed and Alberto Tassoni for making me feel welcome among the UC Berkeley graduate students.

One cannot focus on a project like this without a solid social network. I have been blessed with an amazingly supportive group of friends to divert my attention just the right amount. My fellow PhD students at the humanities faculty have made Umeå feel like my new home. Thanks to Claudia, Fredrik, André, Jenny, Emil, Gustav, Vala, Peter, Eva, and Johan. My friends in Belgium have made my old home feel like a great place to visit. Thanks to Ward, Faai, Stefan, Ruby, Hannah and all the others I will not list because the deadline is minutes away.

Neither my mother nor my father ever objected to my going into philosophy. What is more, they implanted in me the idea that education and understanding are of the utmost importance. I suspect they both underestimate how well they prepared me for a project such as this one. The ambition and energy of my brother Wouter set an excellent example for me. I dedicate this dissertation to my brother Jasper, who would have turned 25 on the day I finished the final draft. It continually pains me to know that there is no matter of fact as to what he is up to now.

Finally, I owe thanks to Moa. Getting to spend the past few years with someone of such empathy, wit and intelligence has been an exquisite gift.

Chapter 1

Introduction

In our everyday life, we take it as a given that mental phenomena can cause our behaviour. You eat because you feel hungry. I wince because I am in pain. Such cases of mental causation provide us with common sense explanations of ordinary behaviour. In scientific research as well, the possibility of mental phenomena causing our behaviour is widely accepted. Developmental psychologists might say that an infant's belief that the ball is in the red box causes it to crawl towards it. Economists might say that distrust in the government will cause tax evasion. In this dissertation, I will provide a model of mental causation.

More specifically, I will provide a model of causation that allows mental phenomena to be causes even if mental phenomena turn out to be distinct from physical phenomena. Philosophical theories according to which (at least some) mental phenomena are distinct from physical phenomena are traditionally called 'dualist' theories of the mental. Such theories typically posit two fundamental kinds of phenomena: mental phenomena and physical phenomena. According to the dualist, these two kinds of phenomena are radically different in nature and make up the fundamental building blocks of the universe. I will assume such a dualist theory for the better part of what follows. This dissertation will thus provide a model of *dualist* mental causation.

It is taken to be a substantial challenge for dualists to explain how mental phenomena can be causes of our behaviour if they are distinct from physical

phenomena. An early formulation of this challenge is due to Elisabeth of Bohemia, who argued that it is unclear how mental phenomena can interact with physical phenomena if they are so radically different in nature. Elisabeth's specific inquiries were directed to Descartes, whose dualist theory lies at the origin of contemporary dualism (Robinson, 2017, Sect. 1.2). According to Descartes, the mental and the physical are two different substances with radically different natures. The nature of the former is *thinking*, whereas the nature of the latter is *extension*. Elisabeth maintained that it is utterly mysterious how these two substances interact. She states (Descartes, 1970, p. 140):

I must admit that it would be easier for me to attribute matter and extension to the soul, than to attribute to an immaterial being the capacity to move and be moved by a body. (Elisabeth to Descartes, June 20, 1643)

Elisabeth's concern seemed to be mostly with *how* non-extended substances can bring about changes in extended substances. Descartes responded that it is a mistake to assume that causal interactions between the mental and the physical should happen in the same way as causal interactions between two physical substances (Descartes, 1970, p. 138). That is to say, it is wrong to conclude from familiar cases of causation between physical phenomena that all causation happens through local contact.

Descartes's conjecture that non-local causation is possible was soon redeemed. With the arrival of Newtonian physics, the idea that causation requires contact lost its credibility. Newtonian physics posits forces that could act at a distance, and therefore without requiring contact. However, further developments in physics posed a new challenge for dualist mental causation. These developments indicated that the physical domain is, in a significant sense, *complete*: the occurrence of any physical phenomenon is fully determined by its physical history (Papineau, 2001, p. 8–26). Even if we set aside worries of the kind proposed by Elisabeth of Bohemia, a proper understanding of physical phenomena appears to leave no room for their being caused by non-physical phenomena. The dualist is thus, once again, posed with a serious challenge. We take mental phenomena to cause our behaviour, and causing our behaviour appears to require causing changes in the physical do-

main (cf. Kim, 2005, Ch. 2): causing my wincing requires causing a change in the physical particles that make up my facial muscles, causing my reaching requires causing a change in the physical particles that make up my arm, etc. If dualists cannot allow for mental-to-physical causation, their position cannot accommodate the seemingly obvious fact that our mental life affects our behaviour.

Dualists have adjusted their theory accordingly. There is a variety of *non-interactionist* dualisms, according to which mental phenomena do not cause physical phenomena. The pervasive regularities between mental phenomena and behavioural phenomena are then explained without relying on mental causation. For example, Malebranche propounded an *occasionalist* explanation. He maintained that God creates the universe anew at every moment and, when doing so, ensures that the occurrences of mental phenomena are in harmony with the occurrences of physical phenomena.¹ Leibniz proposed instead that the harmony between physical and mental phenomena is pre-established. God created the universe such that phenomena in the mental domain and the physical domain would run in parallel with one another (Leibniz, 1898, (1714)); (Woolhouse, 1985). According to both views, there is no causal interaction between the mental and the physical. Instead, the regularities between mental and physical phenomena are God-given.

During the past few decades, a secular variety of non-interactionist dualism has been on the rise. Many contemporary dualists explain the psychophysical regularities by positing a one-way causal relation from the physical to the mental. Physical phenomena cause mental phenomena, but not the other way around. On such a view, mental phenomena are *epiphenomena*: they are caused, but they do not cause. In response to Elisabeth's inquiry, epiphenomenalists maintain that it is a brute fact of nature that physical phenomena cause mental phenomena. In response to the completeness of the physical, epiphenomenalists maintain that mental phenomena do not interfere with physical phenomena, and the completeness of the physical domain therefore remains unviolated. Such a picture has become quite popular among dualist philosophers, and defenses of epiphenomenalism can

¹In fact, occasionalism was thought to solve a variety of problems, but the mental causation problem for dualists was one of them. See Lee (2016).

be found in Huxley (1874), K. Campbell (1970), Jackson (1982),² Chalmers (1996, Sect. 4.4), Kim (2005), and Robinson (1982, 1988, 2006, 2015, 2018). Today, it still is standardly accepted that, if the physical domain is complete, dualists cannot allow for mental causation. Here is an incomplete list of philosophers who have claimed this to be the case, or are otherwise committed to its being the case:

Armstrong (1968); Bieri (1992); Bennett (2007, 2008); Chalmers (2010, 2013); Dennett (1978, 1991a); Gibb (2014); Goff (2017a); Horgan (1987); Kim (1989, 1998, 2003, 2007); Lepore and Loewer (2011); Levine (2001); Lewis (1966); Loewer (2001, 2007b, 2015); McLaughlin (2015); Ney (2009, 2012); Peacocke (1979); Papineau (2002, 2013); Smart (1959) and Woodward (2015).³

Given that the completeness of the physical domain is standardly accepted as well, the default position on dualism is that it will entail the denial of mental causation.

Denying mental causation elicits some strong reactions. For example, Fodor formulates his view on the matter as follows (1989, p. 156):

if it isn't literally true that my wanting is causally responsible for my reaching, and my itching is causally responsible for my scratching, and my believing is causally responsible for my saying... If none of that is literally true, then practically everything I believe about anything is false and it's the end of the world.

Even if most philosophers express their views more temperately in print, many tend to agree with Fodor's underlying point. Not allowing for mental causation is taken to be a major strike against any theory of the mental (Bennett, 2007, p. 316); (Bontly, 2005, p. 331); (McLaughlin, 2015, p. 83); (Stoljar, 2008a, p. 271). Consequently, the reigning conviction that dualism cannot allow for mental causation incites many to reject dualism out of hand.

²Although Jackson famously switched camps and defended a non-dualist position later on (Jackson, 2006).

³Some of these authors, like Ney (2009, 2012), are committed to this claim in virtue of endorsing a similar claim about non-reductionist physicalism. Based on interactions about my dissertation project, I have some anecdotal evidence that this conviction about dualism is widespread among philosophers who do not write on the topic as well.

For example, Goff (2017a), spends around 100 pages diligently providing arguments against the view that all mental phenomena are at bottom physical (p. 23–135), but devotes only two paragraphs to his rejection of dualism on the basis of mental causation worries (p. 158). With these two alternatives out of the way, he goes on to defend the view that all physical phenomena are at bottom mental — a view that he elsewhere labels ‘crazy’ (2017b). Levine (2001, p. 16) acknowledges the pull of dualism, but, with an alacrity similar to Goff’s, concludes that mental causation worries exert a stronger pull towards non-dualism. Its difficulties in allowing for mental causation appear to impair the dualist position considerably.

Given that the situation is so pressing, remarkably little attention has been paid to the question what causation actually *is* in these debates. For example, Goff does not discuss what it requires for a phenomenon to be a cause. Similarly, those dualists who accept that their view results in epiphenomenalism rarely dig very deep into the question what it means to be a cause.⁴ It is typically accepted that, whatever the requirements on causes are, dualist mental phenomena will not meet them. If the arguments and considerations put forward in this dissertation are adequate, this assumption is undermotivated and perhaps even false.

The central thesis of this dissertation is that there is a credible model of causation according to which dualist mental phenomena can be causes in a world where the physical domain is complete. I dub this model *insensitive interventionism*. Very roughly, it states that what it is for phenomenon *A* to cause phenomenon *B* is for it to be possible to manipulate the occurrence of *B* by manipulating the occurrence of *A* in a wide variety of natural scenarios. This model allows for dualist mental causation. In a wide variety of natural scenarios, I can make you wear a coat by making you believe that it is raining. In a wide variety of natural scenarios, you can make me wince by making me experience pain. Consequently, if insensitive interventionism is an adequate account of causation, it is plausible that there is mental causation, regardless of whether or not mental phenomena are distinct from physical

⁴One could be tempted to call Kim an exception, as he is the driving force behind the so-called causal exclusion arguments (Kim, 1989, 1998, 2003, 2005, 2007). For all that, Kim has had very little to say about the nature of causation, except that it is *productive* and that counterfactual dependence or any other ‘lightweight’ definition of causation is unsatisfactory (Kim, 2005, p. 17–18); (Kim, 2007, p. 235–236).

phenomena. If this insensitive interventionism is a credible account of causation, the assumption that dualists cannot allow for mental causation in a physically complete world would indeed be undermotivated, and perhaps even false.

I will argue for the credibility of insensitive interventionism by drawing on recent developments in philosophy of causation. The model builds on the increasingly popular interventionist models of (mental) causation (e.g. Campbell, 2008, 2010; List and Menzies, 2009; Woodward, 2008, 2015), as well as investigations into the relation between causation and physics (e.g. Albert, 2000, 2015), and investigations into the causal role of absences (e.g. Schaffer, 2004; Russo, 2016). I will argue that these developments contravene the reigning conviction that the completeness of the physical domain excludes the possibility of dualist mental causes. This reigning conviction is rooted in more heavyweight notions of causation, according to which causes must be tightly related to either physically sufficient conditions of their effect or to phenomena that transfer energy on their effect. A proper investigation into the nature of causation casts doubt on such requirements and can thus vindicate dualist mental causation by motivating insensitive interventionism. Or so I will argue.

Perhaps the resulting model of dualist mental causation is still flawed. It might be the case that, even given the increasing tendencies towards lightweight accounts of causation, dualist mental phenomena simply cannot be causes in a world where the physical domain is complete. I am not wedded to the idea that there is dualist mental causation. (I am not, in fact, a dualist). My efforts to provide a dualist model of mental causation are driven by the conviction that such a model deserves an honest attempt and that the outstanding models can be improved upon. Borrowing a phrase from Lycan (2009), one could say that I aim to give dualism its due. I aim to do so by developing the most promising model of dualist mental causation that is currently available. Even if this model fails, it is instructive to see *why* it fails, and to understand whether or not its failure is as dramatic as some make it out to be. I think a proper investigation into the nature of causation indicates that the failure of mental causation would not mean the end of the world. However, I postpone that worry until the very end of the dissertation

(i.e. Chapter 16). First, let us investigate whether there is a plausible model of dualist mental causation in a world where the physical domain is complete. Afterwards, we can briefly look at what it would mean if there isn't.

Aim and Outline of the Dissertation

The central thesis of this dissertation is that there is a credible model of causation according to which dualist mental phenomena can be causes in a world where the physical domain is complete. I will develop and defend such a model of causation in what follows. I maintain that this model is *credible* in the sense that it is worth taking seriously; there are no strong reasons to reject it and it is as plausible as some of the models that are currently popular in the literature on causation and mental causation. Dualist mental phenomena *can* be causes according to this model in the sense that, if the sufficient conditions on causation provided by this model are adequate, dualist mental phenomena are causes of physical effects in worlds where dualism is true, the physical domain is physically complete, and the correlational facts are in a way that I will spell out. Importantly, these correlational facts might, for all we know, hold in the actual world. If my central thesis is correct, the standard position that dualists cannot have mental causation in worlds where the physical domain is complete is undermotivated, and perhaps even false.

In order to support my central thesis, I will (i) present a dualist ontology and identify the challenges for models of mental causation within such an ontology, (ii) critically evaluate responses to these challenges from the philosophical literature, and (iii) present and defend a model of causation based on recent developments in philosophy of causation that allows for mental causation within the dualist ontology we started from. The dissertation is divided into three parts, each of which addresses one of these points.

In Part I of this dissertation, I set up the mental causation problem as it arises for a popular variety of dualism. First, I define the dualist theory of mind that will be assumed throughout this dissertation (Chapter 2). Second, I provide a more precise account of what I mean by 'mental causation', and explain why it poses a problem for this dualist theory of mind (Chapter 3). I do so by distinguishing two worries for dualist models of mental causation:

the exclusion worry and the common cause worry. According to the exclusion worry, the conjunction of dualism and some well-accepted theses about causation logically entails that there is no mental causation. According to the common cause worry, models of causation that allow for dualist mental causation will fail to distinguish between correlations that are merely due to two phenomena having a cause in common on the one hand, and genuine causal relations on the other. Finally, I provide a brief preview of my preferred solution (Chapter 4). I distinguish between heavyweight and lightweight accounts of causation and propose to solve the dualist's mental causation problem by adhering to a *thoroughly* lightweight account of causation.

In Part II, I investigate recent proposals to solve the dualist's mental causation problem. In Chapter 5, I discuss List and Stoljar's (2017) objections to the exclusion principle in exclusion arguments that target dualism. In Chapter 6, I discuss Kroedel's (2015; 2020) proposal to avoid mental causation problems by adopting *supernomological* dualism. In Chapter 7, I discuss the recent criticism on the notion of *causal sufficiency* at work in the causal exclusion argument. In Chapter 8, I discuss Lowe's three models of dualist mental causation (Lowe, 1992, 1996, 1999, 2000, 2008). I argue that all of these proposals fail to provide a non-*ad hoc* solution to the problems faced by dualist mental causation, and hence there is still room for improvement.

In Part III, I present and defend my own model of dualist mental causation, which builds on *interventionist* accounts of causation. First, I present a minimal version of interventionism based on the standard interventionist definitions provided in Woodward (2003) (Chapter 9). I then investigate how this minimal interventionism interacts with the question of non-reductionist mental causation (Chapter 10). It will appear that, according to this minimal model, so-called 'holding fixed'-requirements on the physical phenomena underlying mental phenomena stand in the way of dualist mental causation. In Chapter 11, I present a problem for the minimal interventionist model: it allows for too much higher-level causation. I propose to solve this problem by restricting causal correlations to *robust* correlations and briefly compare this solution to other solutions from the literature.

In Chapter 12, I present my model for dualist mental causation. I argue that the robustness condition added in the previous chapter renders the 'hold-

ing fixed'-requirements on the physical phenomena underlying dualist mental phenomena redundant. Hence, an interventionist account that includes a robustness condition can let go of those specific 'holding fixed'-requirements. I dub the resulting account of causation 'insensitive interventionism' and demonstrate how it allows for dualist mental causation and provides a principled reply to both the exclusion worry and the common cause worry. The three remaining chapters of Part III are dedicated to anticipating and countering some objections to insensitive interventionism.

In Chapter 13, I rely on causation by absences to respond to the objection that causation must be productive. I also consider whether or not such cases require us to adjust insensitive interventionism further. In Chapter 14, I rely on recent developments in the literature on causation and fundamental physics to respond to the objection that causes must physically necessitate their effects. I also consider the problem of spurious backwards causation and survey some replies from the literature. In Chapter 15, I briefly reply to some remaining objections. Finally, I will make some concluding remarks in Chapter 16.

Notes on terminology

In order to avoid strain in the following expositions and arguments, I have made some simplifying terminological choices. It is worth making note of these in advance.

First of all, I will refer to causal relata as 'phenomena'. Philosophers disagree on the nature of causal relata. Some philosophers follow Lewis (1973a) and hold that causal relata are events,⁵ others propose to think of them as facts (e.g. Mellor, 1995; Bennett, 1988), states of affairs (Armstrong, 1996), or aspects (Paul, 2000). The recent rise of interventionist theories of causation comes with a tendency to use the term 'variables' in the case of type causation, and 'values' in the case of token causation (see for example Woodward (2003) and Section 9.1). I will not engage with debates on the nature of causal relata. I take it that the term 'phenomena' is sufficiently generic to act as a placeholder. I thereby assume that the outcome of these debates

⁵Although Lewis does not claim that only events can be causes, but rather delimits his account of causation to causation by events (Lewis, 1973a, p. 558).

will not affect the plausibility of what follows.

Second, in the interest of continuity, I will characterize theories about the mental, like dualism and physicalism, in terms of ‘phenomena’ as well. Philosophers use different terms when formulating their theories of the mental. For example, Chalmers (1996) talks of mental ‘properties’, whereas Lowe (1996) follows Descartes (2003, (1643)) by talking of ‘substances’. Others talk of ‘events’ (e.g. Davidson, 1970), ‘states’ (e.g. Putnam, 1975), or ‘processes’ (e.g. Smart, 2017). Given that we are interested here in the causal status of the mental, and we decided to name causal relata ‘phenomena’, I will use the same term when characterizing the central positions on the ontological status of the mental. I assume that translations in the reader’s preferred terminology are straightforward.

Third, my use of the word ‘mental’ will be somewhat atypical. In philosophy of mind, there is a popular distinction between *phenomenally conscious* mental phenomena and *intentional* mental phenomena. Phenomenally conscious mental phenomena are those mental phenomena that constitute *what it is like to be* for someone (cf. Nagel, 1974). Pain is a prime example of such a phenomenally conscious phenomenon. If pain occurs, it constitutes what it is like to be the person that is in pain. Intentional mental phenomena are those mental phenomena that are *about* something. For example, my belief that I am in pain is a belief about my pain. To use another popular formulation, that belief ‘represents’ my pain to me and is therefore ‘directed at’ my pain. The categories of phenomenally conscious and intentional are by no means mutually exclusive. Just like an object can be spherical and red, a phenomenon can be phenomenally conscious and intentional. For example, my desire to get rid of my pain is *about* my pain and there is plausibly something it is like to have that desire. However, many think we can still distinguish between my desire *qua* phenomenally conscious phenomenon, i.e. its being like thus and so to have that desire, and that desire *qua* intentional phenomenon, i.e. its being about my pain. It is therefore customary in debates on the mind to clearly distinguish between those two kinds of mental phenomena.⁶

In this text we will take Chalmers’s (1996) dualism as a starting point,

⁶Though many oppose to the distinction as well. See for example Horgan and Tienson (2002).

and this dualism is a position on the relation between phenomenally conscious mental phenomena and physical phenomena. This position thus maintains that all mental phenomena, *in as far as they are phenomenally conscious*, are distinct from physical phenomena. Consequently, mental causation problems, as they pertain to dualism specifically, are problems about the causal status of mental phenomena in as far as they are phenomenally conscious. It will be easier to drop the explicit ‘in as far as they are phenomenally conscious’-qualifier and assume it as implied throughout the text. Therefore, I will stipulate that ‘mental’ means *phenomenally conscious* in this text. For similar reasons, I will assume that phenomenally conscious mental phenomena are *homogeneous* in their relation to the physical: either all are distinct from the physical, or none are. This assumption too, is not essential to my arguments and expositions.

Moreover, note that the general question we are investigating is whether or not something that is distinct from the physical can be a cause in a world where the physical domain is complete. Whether this concerns an intentional phenomenon, a phenomenally conscious phenomenon or something entirely non-mental, is of no essential concern. We can consider phenomenally conscious mental phenomena as conceived of by the dualist as a case-study, and in order to make that case study run smoothly, we restrict the meaning of ‘mental’ to phenomenally conscious. This is, of course, not to recommend this use more generally. I leave conceptual engineering to the Norwegians.

With these remarks taken care of, we can turn to the set-up of the mental causation problem for dualists.

Part I

The Set-Up

Chapter 2

Nomic Naturalist Dualism

In the first part of this dissertation, I expound the mental causation problem as it arises for a popular variety of dualism. In this chapter, I define the dualist theory of mind that will be assumed throughout this dissertation. I define dualism as the ontological position that denies physicalism about the mental. I add two further qualifications. First, I qualify the dualism we will consider as *naturalist*, because it maintains that all non-mental phenomena are metaphysically necessitated by the physical and because it respects the completeness of the physical domain. Second, I qualify the dualism we will consider here as *nomic*, because it explains psychophysical regularities by positing nomic necessitation relations between physical phenomena and mental phenomena. This nomic naturalist dualism will serve as a starting point for the dissertation.

In the next chapter, I provide a more precise account of what I mean by ‘mental causation’, and explain why it poses a problem for this dualist theory of mind. In the final chapter of this set-up, I provide a glimpse of my proposed solution to the mental causation problem for dualists.

2.1 Dualism

Dualism is an ontological position. It is a philosophical position on which fundamental phenomena make up reality. According to dualism, there are at

least two fundamental kinds of phenomena in our world: mental phenomena and physical phenomena. Physical phenomena are those phenomena that can be exhaustively described in a complete theory of physics, like leptons and quarks having mass and spin.¹ Mental phenomena are those phenomena that are *phenomenally conscious*, like pains and depressions. According to dualism, the fundamental building blocks of the universe consist at least of physical phenomena and mental phenomena. In this section, I give a more precise characterization of dualism by distinguishing it from its main opposing view: physicalism about the mental.

To be a physicalist about X is to maintain that X is at bottom physical. Physicalism about the mental is driven by the idea that all mental phenomena in the actual world are at bottom physical. For example, my pain when I stub my toe is, in a significant sense, nothing over and above a certain physical phenomenon. We can give this idea a more precise formulation by saying that such mental phenomena are *metaphysically necessitated* by actual physical phenomena, where metaphysical necessitation is understood as follows:

Metaphysical Necessitation For any two phenomena A and B , A metaphysically necessitates B if and only if all possible worlds that contain A also contain B .

For example, a piece of cloth being auburn metaphysically necessitates its being red, because there is no possible world in which the piece of cloth is auburn but not red. Similarly, according to the physicalist, all actual mental phenomena are metaphysically necessitated by some actual physical phenomenon or by a collection of actual physical phenomena. For convenience, let us call the collection of physical phenomena that metaphysically necessitates a given mental phenomenon the *metaphysical base* of that mental phenomenon. Physicalism about the mental is the thesis that, for any mental phenomenon in the actual world, there is a physical phenomenon or collection of physical phenomena that metaphysically necessitates the occurrence of that phenomenon.

Building on a definition provided by Stoljar (2010, p. 57), we can characterize the physicalist position on the mental as follows:

¹There are further difficulties with defining ‘physical’ that I will not go into here. See Ney (2008) and Stoljar (2010) for comprehensive discussions of these issues.

Physicalism about the Mental In the actual world, all mental facts are metaphysically necessitated by physical facts.²

Here, physical facts are understood to be facts about which physical phenomena occur when and where.³ Correspondingly, mental facts are those facts about which mental phenomena occur when and where. Given that we are mostly interested in the ontological status of the mental in what follows, we will use ‘physicalism’ as a shorthand for ‘*Physicalism about the Mental*’ from here on.

Physicalists disagree on *why* this metaphysical necessitation relation holds. *Reductionist* physicalists maintain that it holds because mental phenomena are identical to physical phenomena. One can be a reductionist physicalist in two ways. One can maintain that all *types* of mental phenomena are identical to types of physical phenomena. Just like water is identical to H_2O , so pain is identical to a certain physical type (e.g. Gozzano and Hill, 2015; McLaughlin, 2010; Smart, 1959). Alternatively, one can maintain that all *tokens* of mental phenomena are identical to tokens of physical phenomena. Just like Clark Kent is identical to Superman, so my pain is identical to a certain token physical phenomenon (e.g. Fodor, 1974; Peacocke, 1979). Needless to say, either identity claim is compatible with physicalism.⁴ If all mental phenomena are identical to physical phenomena, then all mental phenomena are metaphysically necessitated by physical phenomena.

Other physicalists deny the identity of physical phenomena and mental phenomena. Such *non-reductionist* physicalists maintain that the metaphysical necessitation relation between the mental phenomena and physical phenomena holds in virtue of another kind of intimate relation. Some hold that mental phenomena are *realized* by physical phenomena (e.g. Shoemaker, 2007;

²There are without a doubt problems with this definition, as defining the physicalist position has proven to be quite challenging (cf. Stoljar, 2010). However, most of the challenges appear to concern the *sufficiency* of the imposed requirements (cf. Stoljar, 2016, Section 9). For our purposes the *necessity* of these requirements are more important. Even though some would deny that the above definition imposes adequate necessary requirements (e.g. Montero and Brown, 2018), such views have yet to gain traction in the debate. Overall, I take the above definition to be sufficiently clear and precise for our current project.

³Depending on one’s conceptions of laws, one might want to include facts about the fundamental laws of physics as well. For simplicity’s sake, I ignore that issue here.

⁴Stoljar (2010, Ch. 2) argues that token identity does not suffice for physicalism. Either way those who maintain a token identity of mental and physical phenomena tend to embrace physicalism as well.

Wilson, 2009), others hold that mental phenomena are *grounded* in physical phenomena (e.g. Kroedel and Schulz, 2016), and still others propose that a ‘determinate-determinable’ relation holds between the two (e.g. Yablo, 1992). For our purposes, we can ignore the differences between these varieties. The crucial point is that all these views entail that all the mental phenomena in the actual world are metaphysically necessitated by the physical phenomena in the actual world. On that point, all physicalists are in agreement.

By contrast, the dualist denies physicalism. According to dualism, there is a possible world that contains the same physical phenomena as the actual world, but lacks mental phenomena altogether. Following Chalmers (1996, p. 94–99), we can call such worlds lacking mental phenomena ‘zombie worlds’. A zombie world that is identical in physical facts to the actual world would contain human-like bodies, or ‘zombies’, that behave exactly like we do when we are in pain, at exactly the times and places where there are humans in pain in our world. However, these zombies do not have pain experiences. There is nothing it is *like* to be such a zombie. Very much like there is nothing it is like to be a carburetor or a rock in our world.

The dualist thus maintains that for any collection of physical facts about the actual world, i.e. facts about the distribution of purely physical phenomena in the actual world, there is a possible world where all these facts hold and there are no mental phenomena whatsoever. In other words, no physical fact or collection of physical facts can metaphysically necessitate a mental phenomenon. Using the terminology introduced above, we can say that mental phenomena lack physical metaphysical bases: there are no physical phenomena that metaphysically necessitate mental phenomena.

It will simplify the following discussions if we assume that this lack of metaphysical necessitation holds in both directions. By doing so, we can follow List and Stoljar (2017, p. 103) and characterize dualism as follows:

Dualism For any mental phenomenon M and any purely physical phenomenon or any collection of purely physical phenomena P , it is the case that M is metaphysically distinct from P .

Here, metaphysical distinctness is understood as follows (cf. List and Stoljar (2017, p. 98) and Stoljar (2008a, p. 266)):

Metaphysical Distinctness For any two phenomena A and B , A and B are *metaphysically* distinct if and only if it is metaphysically possible for A to occur without B occurring, *and* vice versa.

For example, a piece of cloth being red and its being square are metaphysically distinct, because it is metaphysically possible for a piece of cloth to be red but not square, and vice versa. That is to say, there are possible worlds in which the piece of cloth is red without its being square and there are possible worlds in which the piece of cloth is square without its being red. By contrast, that same piece of cloth being red and its being auburn are *not* metaphysically distinct, because it is not metaphysically possible for it to be auburn but not red. According to the dualist, mental phenomena stand to physical phenomena as the cloth's being red stands to its being square: they are metaphysically distinct.

Some may find our simplifying assumption controversial. Indeed, there are reasons to believe that some mental phenomena metaphysically necessitate physical phenomena. For example, some philosophers maintain that my believing that there is water in the well requires that I stand in the right kind of causal connection to H_2O (e.g. Putnam, 1973; Burge, 1979, 1986). Consequently, my believing that there is water in the well metaphysically necessitates my having been in contact with water, a phenomenon that at least requires some physical phenomena to have occurred. Moreover, it has recently been argued that knowledge is a mental phenomenon (Williamson, 2000, Ch. 1), and, given that knowledge is factive, knowledge of physical phenomena certainly metaphysically necessitates those phenomena. Allow me to make two notes in response to these worries.

First, recall that we stipulated 'the mental' to mean *the mental in as far as it is phenomenally conscious*. The externalist considerations that lead us to believe that mental phenomena can metaphysically necessitate physical phenomena are considerations about *representation* and *factivity*, and these are typically not individuated in terms of phenomenal consciousness.⁵ I take

⁵Though see Kriegel (2013); Mendelovici (2018) and Montague (2016) for proposals to ground representation in phenomenality. See Clutton and Sandgren (2019) for criticism on such proposals. Note also that most of those who ground intentionality in phenomenality tend not to be externalists about content.

it that it is less plausible that mental phenomena *qua being phenomenally conscious* metaphysically necessitate physical phenomena.

Second, defining dualism in terms of distinctness as done above allows for a cleaner formulation of the upcoming exclusion argument against dualist mental causation. The exclusion argument is often the focal point of discussions on mental causation, and it frequently relies on a notion of distinctness that is analogous to our *Metaphysical Distinctness* (e.g. Bennett, 2008; List and Stoljar, 2017; Stoljar, 2008a). It is thus in the interest of continuity with the ongoing mental causation debates to use this distinctness definition of dualism. If worst comes to worst, and the definition turns out to be untenable, I am confident that the expositions and arguments in this text can straightforwardly be translated into a more precise vocabulary.

With these notes taken care of, we can summarize our characterization of dualism as follows: mental phenomena are metaphysically distinct from physical phenomena. Given that we are looking for a model of dualist mental causation, I will consider this a non-negotiable part of our starting position. Note however, that not everyone agrees with this characterization of dualism. Some philosophers use the term ‘dualism’ to denote a position that is not committed to the metaphysical distinctness of mental and physical phenomena, but instead is closer to the non-reductionist physicalist position discussed above. Examples of such uses can be found in Bogardus (2013); Garrett (2000); Pietroski (1994); White (2018) and Yablo (1992). Brown (forthcoming) explicitly argues that such a theory of mind deserves to be called ‘dualism’ *and* can respond to at least some versions of the causal exclusion argument. Even so, I stick to my characterization in terms of metaphysical distinctness. It is far less controversial that theories of mind that do not posit such a distinctness between the mental and the physical can have mental causation (e.g. Bennett, 2008; Stoljar, 2008a). If we give up on the metaphysical distinctness of the mental and the physical, we risk getting involved in a different debate altogether.

In order to understand why *Dualism* has such problems allowing for mental causation, we need to understand some further assumptions that are standard in contemporary philosophy of mind.

2.2 Naturalist Dualism

There are two central assumptions about the physical that dualists and physicalists tend to agree upon. First, they tend to agree that all non-mental phenomena are metaphysically necessitated by physical phenomena. Second, they tend to agree that the physical domain is, in a significant sense, *complete*. The second assumption plays an important role in our current investigation and, like *Dualism*, it will be granted a non-negotiable status within the confines of this project. The first assumption is not central to our project, but it will simplify our discussions considerably if we assume it nonetheless. By adopting these assumptions, the dualist can brand her position as *naturalist* in that it respects current advances in science. I discuss these assumptions in the named order.

The dualism that we will be considering here assumes that mental phenomena are unique in their independence from physical phenomena.⁶ That is to say, all actual non-mental phenomena *are* metaphysically necessitated by the totality of actual physical phenomena according to this dualism. This holds for biological phenomena, economical phenomena, meteorological phenomena, political phenomena, medical phenomena, etc. Such a physicalism about the non-mental seems plausible. Try imagining a world that is physically identical to ours, but where the 2008 banking crisis did not occur, where napping cures cancer, or where global warming has not taken place. It is highly implausible that such worlds exist. It thus follows that all of these non-mental phenomena are metaphysically necessitated by the totality of physical phenomena.

We can give a more precise formulation of this idea as follows:

Physicalism about the Non-Mental In the actual world, all non-mental facts are metaphysically necessitated by physical facts.

If we add this claim to our current definition of dualism, we arrive at the following position:

Naturalist Dualism–I For any mental phenomenon M and any

⁶See for example Chalmers (1996, Section 2.5) and Stoljar (2010, p. 45).

purely physical phenomenon or any collection of purely physical phenomena P , it is the case that M is metaphysically distinct from P . Moreover, in the actual world, all non-mental facts are metaphysically necessitated by physical facts.

According to the naturalist dualist, a world that is identical to the actual world in all its physical facts must contain the same non-mental phenomena as the actual world, but might lack mental phenomena.

Physicalism about the Non-Mental is not essential to dualism. In principle, the dualist could maintain that some non-mental phenomena in our world are metaphysically distinct from the physical as well. Moreover, some philosophers hold that this is indeed the case. For example, Hattiangadi (2018) maintains that moral phenomena are not metaphysically necessitated by physical phenomena. One might worry that, by adding *Physicalism about the Non-Mental* to our theory of mind, we are needlessly burdening it with extra commitments.

However, it is worth it to assume the more demanding *Naturalist Dualism-I* in the context of our project. Suppose that it is demonstrated that mental phenomena cannot be causes on the assumption that they are the only phenomena in our world that are not metaphysically necessitated by physical phenomena. If there should then turn out to be another exception, like moral phenomena, it will indeed be the case that we cannot conclude that conscious phenomena cannot cause moral phenomena. However, it would still be the case that conscious phenomena can *only* cause moral phenomena, and this is still a significant consequence of our theory. In everyday life, we certainly assume that conscious phenomena *can* cause non-moral phenomena, like my wincing or your wearing a coat. More generally, the phenomena we typically take to be caused by mental phenomena, like behavioural phenomena, do appear to be metaphysically necessitated by the physical. I therefore ignore possible exceptions to *Physicalism about the Non-Mental*. Let us now turn to the second aspect of the naturalism we will assume.

Advances in science indicate that the physical domain is *physically complete*.⁷ That is to say, advances in science indicate that every occurrence of a

⁷See Papineau (2001) for a concise summary of the improvements in science that motivate this thesis.

physical phenomenon is fully determined by its physical history.⁸ For example, my wincing is fully determined by the physical phenomena that precede it. We can give the completeness of the physical as it is assumed in these debates a more precise formulation as follows:

Physical Completeness For any actual physical phenomenon P and any time t , there is a purely physical phenomenon that occurs at t and physically necessitates the occurrence of P .

where physical necessitation is understood as follows:

Physical Necessitation For any two phenomena A and B , A physically necessitates B if and only if all physically possible worlds that contain A also contain B .

By ‘physically possible world’, I mean any possible world in which the same fundamental laws of physics as in our world hold. This necessitation relation is thus modally weaker than the metaphysical necessitation relation, as that relation applied to *all* possible worlds, even those with different laws of physics than ours. For example, even in worlds where particles can accelerate from a speed below the speed of light to a speed above the speed of light — something that our laws of physics do not allow⁹ — all red pieces of cloth are coloured pieces of cloth. That is why we can say that a cloth’s being red metaphysically necessitates the cloth’s being coloured. By contrast, a physical phenomenon that physically necessitates my wincing might occur in a world with different laws of nature *without* my wincing. (Perhaps a rogue particle accelerates above the speed of light threshold and kills me right before I get a chance to wince.) Such physical phenomena are only sufficient for my wincing on the assumption that the laws of physics remain constant.

⁸This is only true on the assumption of determinism in the physical domain. It is customary to assume this in debates on mental causation and physical completeness, because it simplifies formulations and translations to indeterministic-friendly formulations are straightforward (e.g. Bennett, 2003; Loewer, 2001; Papineau, 2001). I will follow suit here and assume determinism as well. See Bourget (2019) for an argument to the extent that indeterminism *does* make difference for mental causation debates.

⁹I found mention of this point in Fernandes (2016, p. 8), who refers to Maudlin (2002, Ch. 4).

Even so, *Physical Necessitation* delivers a significant completeness claim. *Physical Completeness* entails that any change in a physical phenomenon requires a change in its physical history. For example, to bring about a change in my wincing, one needs to bring about a change in the physical history of my wincing. After all, this physical history physically necessitates my wincing and the laws of physics remain constant in our world, and therefore there is a physically *sufficient* phenomenon for my wincing at any time in its history.¹⁰

By adding *Physical Completeness* to our current definition of dualism we arrive at the following dualist position:

Naturalist Dualism–II For any mental phenomenon M and any purely physical phenomenon or any collection of purely physical phenomena P , it is the case that M is metaphysically distinct from P . Moreover,

- (i) In the actual world, all non-mental facts are metaphysically necessitated by physical facts.
- (ii) The physical domain is physically complete.

We now start to see how mental causation worries are bound to arise for this position. According to (i), any non-mental change metaphysically requires a change in the physical domain, and according to (ii), the physical domain appears to leave no room for non-physical causes. After all, all physical phenomena are already necessitated by their physical histories and the fundamental laws that hold in our world.

Indeed, Chalmers (2010, p. 130–133) even equates the combination of dualism and *Physical Completeness* with a denial of mental causation, and Bennett (2007, p. 328) states that whether or not dualists should worry about mental causation “[...] depends entirely on whether or not they are obligated to accept the completeness of physics.”¹¹ One of the central

¹⁰In fact, *Physical Completeness* entails more than that. It entails that there is a physical phenomenon that physically necessitates my wincing at *any* time. This means that changing my wincing requires a change in physical phenomena at *any* point in time, both before and after my wincing. We will address the time-symmetry of physical necessitation more elaborately in Chapter 14.

¹¹To be more precise, she states that the dualist should only be bothered by the exclusion argument if she is obligated to accept the completeness of physics. Given that Bennett does not seem impressed by other arguments against dualist mental causation, I take it this amounts to the same thing for her.

aims of this dissertation is to investigate whether or not the dualists can have mental causation *and* adhere to *Physical Completeness*. Consequently, I will treat the acceptance of *Physical Completeness* as non-negotiable.

This is not to say that *Physical Completeness* is in fact true. Indeed, some philosophers cast doubt on *Physical Completeness* for reasons that have nothing to do with mental causation (e.g. Cartwright, 2010).¹² However, those dualists who deny *Physical Completeness*, tend to avoid mental causation problems. The main challenge for such positions is to motivate the denial of *Physical Completeness*. This makes for an entirely different project. Proposals for such a view can be found in Chalmers and McQueen (2014); Foster (1991); Hasker (2010); Hodgson (1991); Molenaar (2006); O'Connor (2000); Popper and Eccles (1977); Sellars (1981); Stapp (1993, 2001, 2013, 2014) and Swinburne (1986).¹³ Before providing a more precise formulation of why accepting *Physical Completeness* creates such pressing mental causation worries for the dualist, we add one more qualification to our dualist position.

2.3 Nomic Naturalist Dualism

Our current dualist position leaves a striking fact about our world unexplained. There are pervasive correlations between mental phenomena and physical phenomena: pains are often followed by wincings, beliefs and desires often precede actions, some neural disorders come with mental disorders, etc. If there is *no* modal connection between mental phenomena and physical phenomena, these psychophysical correlations have the status of a massive coincidence. The dualist can solve this lack of explanation by positing a weaker modal relation than metaphysical necessitation between the mental and the physical.

According to some contemporary dualists, psychophysical correlations are explained by psychophysical laws. Psychophysical laws state that certain occurrences of physical phenomena give rise to certain occurrences of mental

¹²See also Lycan (2009, p. 559), who cautiously suggests that the relevant conservation laws might allow for the kind of interactionism that Descartes had in mind.

¹³See also Gibb (2013a,b, 2014, 2015a,b) for a view that denies *Physical Completeness* but according to which there is no mental causation either. She argues that the incompleteness of the physical allows for mental phenomena to be causally *relevant* in some significant sense, without their being proper causes.

phenomena. For example, there could be some physical phenomenon, call it ‘phys’, such that all occurrences of phys give rise to pain. It is thus no surprise that there are pervasive correlations between pain and physical phenomena. All actual occurrences of phys are accompanied by occurrences of pain. So if it is the case that occurrences of phys tend to be followed by wincings, it will also be the case that occurrences of pain tend to be followed by wincings, and if scorching my thumb is bound to result in phys, it is bound to be followed by pain. Similar explanations will be available for other correlations between mental phenomena and physical phenomena.

It is important to qualify the psychophysical laws carefully. The existence of these laws should NOT interfere with the metaphysical distinctness of mental and physical phenomena, but they should still be able to explain the psychophysical correlations. Here I outline what I take to be the standard dualist characterization of psychophysical laws and add this characterization to our dualist starting position.

Dualists can explain psychophysical correlations without interfering with the metaphysical distinctness of the mental and the physical by modelling psychophysical laws on the fundamental laws of physics. Typically, dualist psychophysical laws are put on a par with the fundamental laws of physics in two significant ways. They are taken to be on a par in modal status, because they are contingent, and in ontological status, because they are fundamental. I discuss these two features in the named order.

First, the psychophysical laws are contingent. That is to say, there is at least one possible world where these laws fail to hold. After all, the dualist maintains that there are possible worlds where phys does not give rise to pain — zombie worlds being the prime example. If she were to maintain that these laws hold in all possible worlds, like the law that all red things are coloured things and the law that all squares are rectangles, she would be committed to physicalism. It is therefore essential to the dualist position that psychophysical laws are *not* laws that hold in all possible worlds.

Psychophysical laws share this feature with the fundamental laws of physics. These laws too, fail to hold in other possible worlds. For example, there are possible worlds where particles accelerate above the speed of light threshold. Such worlds are not *physically* possible, but they are *metaphysically* possi-

ble. Qua modal status, the psychophysical laws are thus on a par with the fundamental laws of physics.

Second, psychophysical laws are *fundamental*. They are fundamental in the sense that they are not explained by any further laws or facts, but rather explain other facts, like the fact that I am in pain when phys occurs in my body, and plausibly some *ceteris paribus* laws, like the law that states that, all else equal, scorching your thumb results in pain. The psychophysical laws are thus granted the same status as the physical laws: they are part of the fundamental laws that govern the universe. Qua ontological status, psychophysical laws are on a par with the fundamental laws of physics as well.¹⁴

From now on, I shall refer to such fundamental, contingent laws as *nomic laws*. Consequently, we can characterize the modal relation between the mental and the physical as a relation of *nomic necessitation*, where nomic necessitation is understood in analogy with metaphysical necessitation, but the metaphysical modality involved is replaced with a nomic modality:¹⁵

Nomic Necessitation For any two phenomena *A* and *B*, *A* nomicallly necessitates *B* if and only if all nomicallly possible worlds that contain *A* also contain *B*.

For example, it might be nomicallly impossible for a certain physical phenomenon phys to occur without pain occurring. Just like it is nomicallly impossible for a particle to accelerate across the speed of light threshold. Extending on our terminology, we can say that phys is the *nomic base* for pain in such cases: it underlies the mental phenomenon and *nomicallly* necessitates its occurrence. To simplify discussions, we will assume that mental phenomena and their physical bases occur simultaneously in what follows. Consequently, the nomic necessitation that holds between them can be called ‘synchronic’ or

¹⁴The above paragraph makes some non-trivial assumptions about laws *governing* matters of fact, rather than laws merely metaphysically supervening on, and being derived from, the matters of fact (cf. Loewer, 1996). I see no reason why the dualist could not adopt the latter, Humean view as well, but discussing the metaphysics of laws would take us too far afield. I take translations into a more Humean-friendly dualism to be straightforward.

¹⁵See also Chalmers (1996, Ch. 4 & Ch. 6), who speaks of ‘nomological supervenience’ to denote the modal relation between the mental and the physical. Talk of supervenience has lost some of its popularity since then, and I avoid using the term here. However, I take my characterization to capture the position he presents.

‘vertical’. Unless stated otherwise, I have such synchronic nomic necessitation in mind when talking of nomic necessitation. In semi-formal environments like definitions or arguments, I will use the relevant reminder.

As opposed to the adherence to *Physical Completeness*, I take this characterization of psychophysical laws to be a negotiable aspect of the dualist position in the current context. Some dualists argue that the psychophysical laws could be modally weaker than nomic laws (e.g. Lavazza and Robinson, 2014, p. 3–4). Others propose to stipulate that they are modally stronger than nomic laws (e.g. Kroedel, 2015, 2020). I am open to revising the characterization of psychophysical laws — and will explicitly discuss this possibility in Chapter 6 — as long as it does not threaten *Dualism*. The nomic characterization serves mostly as a placeholder in the dualist position that we take as a starting point here.

If we add this final qualification to our current definition, we arrive at the following characterization of dualism:

Nomic Naturalist Dualism For any mental phenomenon M and any purely physical phenomenon or any collection of purely physical phenomena P , it is the case that M is metaphysically distinct from P . Moreover,

- (i) In the actual world, all non-mental facts are metaphysically necessitated by physical facts.
- (ii) The physical domain is physically complete.
- (iii) For any occurrence of a mental phenomenon M in the actual world, there is a physical phenomenon or set of physical phenomena P such that the occurrence of P synchronically nomically necessitates the occurrence M .

We can summarize the current dualist ontology as follows. There are physical phenomena and there are mental phenomena. There are psychophysical laws that govern the correlations between these two, and there are physical laws that govern the correlations between physical phenomena in such a way that the physical domain is physically complete. And that’s it. All the rest is a ‘free lunch’ in the sense that it is metaphysically necessitated by the

physical and the mental.

I take this to be a relatively standard dualist ontology. Chalmers (1996, ch. 4), Koons and Bealer (2010, p. xvi), and Pautz (2010, Section 12) provide a similar ontology.¹⁶ Stoljar considers such an ontology the ‘standard dualism’ (Stoljar, 2008a, p. 270), and it appears to be the default dualist position in debates on dualism and mental causation (e.g. Bennett, 2008; Kroedel, 2015, 2020; Woodward, 2015). It is also this position that is typically denominated as ‘epiphenomenalism’ by its own proponents (e.g. Huxley, 1874; Campbell, 1970; Jackson, 1982; Robinson, 1982, 1988, 2006, 2015, 2018; Chalmers, 1996; Kim, 2005). This denomination is based on the common assumption that, by combining *Physical Completeness* and *Dualism*, such a view can only allow for mental phenomena to be caused by physical phenomena, but can never allow for mental phenomena to be causes themselves (e.g. Chalmers (2010, p. 230–233) and Bennett (2008, p. 283)). Hence, such a view is taken to entail that mental phenomena are *epiphenomena*.

I aim to challenge that common assumption in this dissertation. I will thus take *Nomic Naturalist Dualism* as a starting point and argue for a model of causation according to which that ontology can allow for mental causation. With that purpose in mind, I will use ‘dualism’ as a shorthand for *Nomic Naturalist Dualism*. Now we can take a look at why allowing for mental causation is taken to be such a challenge for this brand of dualism.

¹⁶See also Hasker (2014, p. 215–216), who seems to propose a similar view when he states that the mind ‘naturally emerges’ from the brain and when he talks of the brain as ‘generating’ the mind (Hasker, 2010). However, Hasker does not embrace the completeness of the physical.

Chapter 3

Mental Causation

What would it mean for a model of causation to allow for (dualist) mental causation? We can get a grip on the answer by looking at what philosophers take to have lost if there were no mental causation. Here again is Fodor's inspiring quote (1989, p. 156, italics added):

if it isn't literally true that my wanting is causally responsible for my reaching, and my itching is causally responsible for my scratching, and my believing is causally responsible for my saying... If none of that is literally true, *then practically everything I believe about anything is false* and it's the end of the world.

If mental phenomena cannot be causes, a substantial subset of our causal judgments are false. This subset is substantial both in the sense that it is sizeable and in the sense that it is important. It is not just once in a while that we take mental phenomena to be causes. We rely on this possibility of mental causation all the time. Moreover, we rely on that possibility in crucial situations. We try to keep our superiors from feeling angry and take care not to cause distress or pain in our loved ones. Similarly, doctors and psychiatrists prescribe anaesthetics and antidepressants to minimize unpleasant experiences in their patients. Such practices presuppose that experiences have effects. In legal contexts, it matters whether the crime was committed due to an irresistible urge or as a result of a premeditated plan. This

distinction too assumes that urges can cause behaviour.¹ Contradicting this subset of causal judgments would contradict leading theories in psychology, as well as most legal systems. This might not be the end of the world, but it *would* mean that both our common sense causal judgments and plenty of special science judgments about mental phenomena are systematically mistaken. That subset is too central to our understanding of the world to let go off. So the main line of argument against denying mental causation goes.

What it means to allow for mental causation is thus to allow for that subset of causal judgments to be true. We can give a rough characterization of these causal judgments as follows:

Mental Causation Mental phenomena systematically cause physical phenomena.

We thus need a plausible account of causation according to which mental phenomena can systematically cause physical phenomena. In particular, we require an account of causation according to which mental phenomena can cause our behaviour. The goal of this dissertation is to provide a model of mental causation according to which both *Nomic Naturalist Dualism* and *Mental Causation* can be true.

There are two central challenges for such a model. First, many philosophers agree that the combination of *Nomic Naturalist Dualism* and *Mental Causation* will, together with some further plausible propositions, result in contradiction. That is the challenge posed by so-called ‘exclusion arguments’ (e.g. Bennett, 2008; Kim, 2005). Second, it is sometimes suggested that any account of causation that *does* allow for both *Nomic Naturalist Dualism* and *Mental Causation* to be true, will have implausible consequences (e.g. Bennett, 2008; Kim, 2007; Lewis, 1966). More specifically, such an account runs the risk of predicting that two phenomena stand in a genuine causal relation merely in virtue of their having a cause in common. I will call this second challenge the ‘common cause worry’.

Before addressing these two particular challenges to dualist mental causation, we should distinguish mental causation problems from a closely related

¹Note that urges are distinctively phenomenal; consider the absurdity of saying ‘I felt an irresistible urge to strike the referee, but it was not like anything to have that urge’.

objection to dualism. Many object to dualism because it threatens to reduce mental phenomena to *nomic danglers*: they are nomically necessitated, but do not nomically necessitate. The mental and the physical would therefore be very poorly integrated with one another and the resulting ontological pictures seems, in a word, *inelegant*. Even though these complaints are often mixed with complaints about the counterintuitiveness of not allowing for mental causation,² it is important to keep them separate. Our model of dualist mental causation is supposed to solve the mental causation problem, *not* the nomic danglers problem. This should absolve dualism from its counterintuitiveness, but it is unlikely to help against its inelegance.

3.1 The exclusion worry

The exclusion worry is by far the most discussed challenge to dualist mental causation in the contemporary literature. This is in a large part because the underlying exclusion argument was developed to affect non-reductionist physicalism as well as dualism at a time when non-reductionist physicalism had become a popular view on the mental. The argument originates in work by Malcolm (1968) and was further developed by Peacocke (1979), Kim (1989, 1998, 2005, 2007), Ney (2009, 2012), Papineau (2002), and others. In this section, I provide a formulation of the argument that specifically targets dualism and discuss the motivations of its premises. This exclusion argument will serve as a touchstone for accounts of dualist mental causation throughout the rest of the dissertation.

Causal exclusion arguments have many formulations. All of them consist of an inconsistent set of propositions. One proposition states the targeted non-reductionist position, one states some variation of *Mental Causation* and the other three state supposed truisms about causation. We can take the following formulation, loosely based on Bennett (2008), as a starting point:

Mental Causation Mental phenomena systematically cause physical phenomena.

²See for example, Smart (1959, p. 155–156), Dennett (1991a, p. 402), and Papineau (2002, p. 23).

Dualism Mental phenomena are metaphysically distinct from physical phenomena.

Causal Closure Every physical phenomenon has a sufficient physical cause at any given time t (if it has a cause at all at t).

Causal Exclusion For any three phenomena A , B and C : if A occurs at t and is a sufficient cause for B 's occurrence at $t + x$, no phenomenon C occurring at t that is metaphysically distinct from A and is metaphysically distinct from all of A 's parts is a cause of B , unless it is a case of genuine overdetermination.³

Non-Overdetermination There is no systematic genuine overdetermination of physical effects with mental causes.

Mental Causation and *Dualism* together entail that physical phenomena are systematically caused by phenomena that are metaphysically distinct from physical phenomena. *Causal Closure* and *Causal Exclusion* together entail that there are no phenomena that are metaphysically distinct from physical phenomena and cause physical phenomena, *unless* it is a case of genuine overdetermination. Finally, *Non-Overdetermination* states that genuine overdetermination is not an option. One of these five has to go.

Exclusionists can exploit the internal inconsistency of this set of propositions to devise an argument against dualist mental causation. In particular, one can use the inconsistency to argue that the denial of *Mental Causation* logically follows from the assumption of *Dualism* in conjunction with the three truisms about causation.

Dualism Mental phenomena are metaphysically distinct from physical phenomena.

Causal Closure Every physical phenomenon has a sufficient physical cause at any given time t (if it has a cause at all at t).

Causal Exclusion For any three phenomena A , B and C : if A occurs at t and is a sufficient cause for B 's occurrence at

³For the sake of simplicity, I assume throughout the rest of the text that, for any phenomenon X , being metaphysically distinct from X requires being metaphysically distinct from all of X 's parts. I will use reminders like 'and is distinct from all of X 's parts' in definitions.

$t + x$, no phenomenon C occurring at t that is metaphysically distinct from A and is metaphysically distinct from all of A 's parts is a cause of B , unless it is a case of genuine overdetermination.

Non-Overdetermination There is no systematic genuine overdetermination of physical effects with mental causes.

No Mental Causation Mental phenomena do not systematically cause physical phenomena in the actual world.

The argument is valid. It follows from *Causal Closure* and *Causal Exclusion* that no cause of a physical effect can be metaphysically distinct from its sufficient physical cause, unless it is a case of genuine overdetermination. *Dualism* states that mental phenomena are metaphysically distinct from all physical phenomena. It follows that they can only cause physical effects in cases of genuine overdetermination. *Non-Overdetermination* states that it is not the case that physical effects are systematically genuinely overdetermined by mental causes. It follows that mental phenomena do not systematically cause physical phenomena in the actual world. If the four premises are true, *No Mental Causation* must be true as well. The four premises thus provide a valid argument that straightforwardly contradicts the possibility of dualist mental causation. In order for our project to succeed, we will thus have to deny one of the (supposed) truisms about causation.

In the previous chapter, we gave a precise characterization of *Dualism*. In the rest of this section, I take a closer look at the other four propositions in the exclusion argument. Afterwards, I turn to another worry for models of dualist mental causation: the common cause worry.

Mental Causation

Mental Causation is essential to our current project. We will thus *not* be interested in responses to the exclusion arguments that require a flat-out denial of *Mental Causation*. Even so, it is important to be clear on the motivation for *Mental Causation*, as this motivation forms the demands on admissible models of (mental) causation.

Mental Causation is motivated by our causal judgments. As discussed earlier, denying *Mental Causation* is tantamount to denying a substantial subset of our common sense causal judgments, as well as plenty of causal judgments in respected special sciences like psychology. More specifically, it is tantamount to denying a substantial subset of our judgments about what causes what. These can be judgments about specific instances, like when my pain causes my wincing, or causal generalizations, like when fear causes trembling. I take such judgments to be the primary motivation of *Mental Causation*.

We should distinguish these causal judgments about what causes what from intuitions about the nature of causation. Just like we have strong intuitions about what causes what, we have some strong intuitions about the nature of causation. For example, Hall (2004) remarks that we strongly feel that causation is transitive: we readily accept that, if A causes B and B causes C, then A causes C. For example, if a hurricane causes the crops to fail and the failure of the crops causes a famine, the hurricane causes a famine. It is important to distinguish such intuitions about the nature of causation from our causal judgments about what causes what because the latter, but not the former, motivate *Mental Causation*. We do not think, for example, that mental phenomena systematically cause physical phenomena because we think that causation is transitive, intrinsic, or local. Instead, we are convinced that *Mental Causation* is true because we take ourselves to know of many important instances of mental phenomena causing physical phenomena. Consequently, I will reserve the term ‘causal judgments’ for our judgments about what causes what in the rest of the dissertation.

The central role of such causal judgments forms the demands on responses to the exclusion argument in two important ways. First, responses to the argument should not rely on models of mental causation that systematically contradict our causal judgments. If we were to rely on such a model, we would, at the very least, require an explanation of why our causal judgments in cases of mental causation should take precedence over the causal judgments that the model systematically contradicts. It would be particularly problematic if our model contradicted our causal judgments by systematically overascribing causation. For example, a model that counts tar-stained

teeth as a cause of lung cancer or smoke as a cause of fire cannot be trusted to deliver the right result in cases of pains and winces. As we shall see later in this chapter and in Chapters 13 and 14, there is a real risk that models of dualist mental causation spuriously ascribe causation and this restriction on models of causation will pose a recalcitrant challenge throughout our project.

Second, demands on causation that systematically contradict our causal judgments are of only limited concern. This is because such demands would undercut our motivation for maintaining that there is mental causation. If our causal judgments are systematically unreliable, it is at least less clear what motivates *Mental Causation*, and it is therefore at least less clear why it would be such a decisive blow against dualism if it cannot allow for mental causation. Our opponent would owe us an explanation for why we should still believe there is mental causation if our causal judgments are systematically mistaken.⁴ As we shall see in Chapters 13 and 14 these considerations will be of central importance for our defense of dualist mental causation.

With these remarks on *Mental Causation* taken care of, we can turn to the first truism about causation: *Non-Overdetermination*.

Non-Overdetermination

Non-Overdetermination is perhaps the least controversial of the premises in the exclusion argument. Or at least, it is fairly uncontroversial if one distinguishes *genuine* overdetermination from *trivial* overdetermination.

By *genuine* overdetermination, I mean a situation where one effect has two simultaneously occurring causes which cause the effect *independently* of one another. That is to say both of these causes *could* have occurred in the absence of the other and, in such a case, would still have caused the same effect. For example, two bullets simultaneously piercing a victim's heart is a case of genuine overdetermination. Both Bullet 1 piercing the heart and Bullet 2 piercing the heart are causes of the victim's death, and both Bullet 1 piercing the heart and Bullet 2 piercing the heart would have caused his

⁴This is not to say that requirements on causation that systematically contradict our causal judgments do not have a place in philosophy of causation. As noted by Ney (2009) and others, there are several goals worth pursuing with a model of causation. However, such requirements would not pick out the relation that is at issue in the mental causation debate.

death in the absence of the other phenomenon's occurrence. I call this form of overdetermination 'genuine', because in such cases the effect is genuinely *overdetermined* in the sense that each of the two causes is *redundant* given the occurrence of the other. There is one cause too many. This distinguishes genuine overdetermination from trivial overdetermination.⁵

By *trivial* overdetermination, I mean a situation where one effect has two simultaneously occurring causes but one of these causes could not have occurred in the absence of the other because these causes stand in a metaphysically intimate relation to one another. For example, the hurricane occurring and the microphysical phenomenon underlying the hurricane occurring both destroying my lemon tree is a case of trivial overdetermination. Both of these phenomena caused the destruction of my lemon tree, and the hurricane is arguably non-identical to the underlying physical phenomenon. After all, the hurricane could have occurred even if the underlying phenomenon had been different by containing a few electrons more or a few electrons less. But it is *not* the case that the microphysical phenomenon underlying the hurricane could have occurred in the absence of the hurricane. The hurricane is metaphysically necessitated by the the underlying microphysical phenomenon.⁶ I call this form of overdetermination 'trivial', because it appears to be omnipresent (cf. Sider, 2003; Schaffer, 2003). A traffic light's being scarlet and its being red can cause me to hit the brake, a painting's having certain microphysical features and its having certain aesthetic features can cause me to like it, etc.⁷ Consequently, there seems to be nothing objectionable or suspicious about such overdetermination.

Many philosophers maintain that physical effects with mental causes are systematically *trivially* overdetermined. In fact, most non-reductionist physicalists endorse a model of mental causation that relies on trivial overdetermination (see Chapters 5 and 10).⁸ To my knowledge, nobody maintains

⁵This is without a doubt a defective characterization of what I mean by 'genuine overdetermination'. As is demonstrated in Won (2014), it is particularly difficult to provide a good characterization of the phenomenon. However, I take the difference with trivial overdetermination to be sufficiently clear for our purposes.

⁶Try conceiving of a world where that same microphysical phenomenon occurs but there is no hurricane.

⁷There are also cases of trivial overdetermination where the two causes stand in a metaphysically intimate relation that is modally weaker than metaphysical necessitation. I reserve discussion of such cases for Chapter 5.

⁸Though see also Bernstein (2016), who argues that trivial overdetermination cannot

that they are systematically *genuinely* overdetermined.⁹ The consensus is that positions that deny *Non-Overdetermination* would be both *ad hoc* and unparsimonious (e.g. Papineau, 2002, section 1.5). They would be *ad hoc* because we do not appear to have any reason for believing that these effects are genuinely overdetermined apart from an inclination to safeguard dualist mental causation. They would be unparsimonious because they posit a redundant causal relation for any effect of a mental phenomenon. The resulting ontological picture is not inconsistent, but is bound to offend the naturalist sensibilities of the dualism we are currently considering. All else equal, one would prefer an answer to the exclusion argument that does not require such an *ad hoc* and unparsimonious component as the denial of *Non-Overdetermination*.

The main motivations for *Non-Overdetermination* thus appear to be (i) a total lack of non-*ad hoc* reasons to deny it and (ii) the unparsimonious picture that results from denying it. Consequently, *Non-Overdetermination* is not incontestable, but we still require independent motivation to contest it. I follow the consensus in the literature by ignoring blank denials of *Non-Overdetermination*.

Closure and exclusion

The two remaining truisms about causation are *Causal Closure* and *Causal Exclusion*.

Proponents of exclusion arguments take *Causal Exclusion* to be an *a priori* principle about causation (e.g. Kim, 2005, p. 17–18).¹⁰ The underlying idea is that once an effect is ‘sufficiently caused’ by a certain phenomenon, there is no causing left to do by another phenomenon that is metaphysically distinct from that sufficient cause — unless the effect is caused twice over.

deliver the solution sought by the non-reductionist.

⁹The closest example I came across in the literature is Meixner (2004). See also Lowe’s (2005) discussion of that text. Note however, that it is unclear whether or not Meixner is a dualist according to our use of the term, as he takes the claim that there are no token-identities between the mental and the physical to suffice for the denial of physicalism (Meixner, 2014, p. 18). Consequently, it is unclear whether the overdetermination he has in mind in cases of mental causation is a brand of *genuine* overdetermination according to our categorization.

¹⁰In fact, Kim defends an even stronger exclusion principle than the one currently under consideration. Given that he takes this stronger principle to be *a priori* true, it is safe to assume that he takes *Causal Exclusion* to be so as well.

The upshot is that all non-overdetermining causes of any given effect must be metaphysically necessitated by the sufficient cause of that effect.

Causal Closure states that there are sufficient *physical* causes for any physical effect. This proposition is taken to follow directly from *Physical Completeness*, which states that for any physical phenomenon P , there is a physical phenomenon at any time t that is physically sufficient for the occurrence of P . *Prima facie*, these physically sufficient physical phenomena mentioned in *Physical Completeness* will qualify as sufficient *causes* for E as well. Consequently, *Physical Completeness* and *Causal Closure* are often used interchangeably, and arguments to support the one are often assumed to directly support the other (e.g. Papineau, 2002; Chalmers, 2010). As we shall see in Chapter 7, matters are not quite as simple, but for now we will follow suit. Given that we committed ourselves to *Physical Completeness*, it is *prima facie* plausible that we are committed to *Causal Closure* as well.

Together, *Causal Closure* and *Causal Sufficiency* entail that all causes are metaphysically necessitated by the sufficient physical causes of their target effects (unless it is a case of genuine overdetermination). With *Non-Overdetermination* tentatively out of bounds, it seems that the dualist has no place left to go. After all, the two remaining truisms about causation are a principle that is taken to be *a priori* true and a principle that is taken to follow directly from one of her central commitments. Denying any of these three truisms seems like a hopeless task. The situation is *that* bad. Even so, I believe there is room for dualist mental causation. My reasons for believing so will have to wait for now. Let us first turn to another worry.

3.2 The common cause worry

According to the dualist picture, it seems plausible that mental phenomena and their purported effects always have a cause in common: the physical nomic base of the mental phenomenon. This poses an extra worry for dualist mental causation. Correlations between effects of the same cause are often due to their relation to that common cause, rather than due to a causal relation between the two effects. A proper model of dualist mental causation has to motivate treating dualist mental phenomena as causes of their

target effects, despite their exhibiting such a common cause structure. This constitutes the common cause worry.

The correlation between phenomena with a common cause is often due to this common cause playing the role of *confounder*. For example, there is a strong correlation between the occurrences of tar-stained teeth and the occurrences of lung cancer in most populations, because both of these phenomena are caused by smoking behaviour. In such cases, it can often *seem* that two phenomena are causally related due to the strong correlation, when in fact this strong correlation is fully explained by the presence of the confounder, as is represented in Figure 3.1. Consequently, the effects have no causal role left

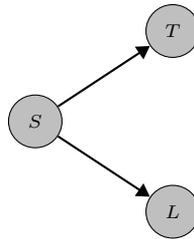


Figure 3.1: Smoking (S) causes both tar-stained teeth (T) and lung cancer (L).

to play relative to one another. Tar-stained teeth do not cause lung cancer, nor does lung cancer cause tar-stained teeth.

However, it is *not* excluded that effects of a common cause stand in a genuine causal relation. For example, it could be the case that going to a private school causes academic success, even though both going to private school and academic success are caused by the financial status of one's parents. The situation would thus look as depicted in Figure 3.2. Both the financial situation of one's parents and private school attendance cause academic success, even though the financial situation also causes private school attendance. In practice, this would mean two things. First, private school attendance will affect one's chances of academic success, even if one attends private school despite the family's poor financial situation — perhaps by means of a stipend or extreme frugality on part of the parents. Second, the financial situation of the parents will affect one's academic success, even if one does not attend private

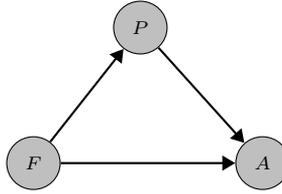


Figure 3.2: The financial situation (F) causes both private school attendance (P) and academic success (A). P also causes A .

school. In such a case, effects of a common cause do stand in a genuinely causal relation.

The challenge for the dualist is to argue that the case of dualist mental phenomena and their target effects is (in the relevant sense) like the second case rather than the first. *Prima facie*, this seems unlikely. It seems plausible that the correlation between conscious phenomena and their purported effects is fully explained by the presence of a confounding cause; namely the physical nomic base of the conscious phenomenon. The *appearance* of mental causation would thus be explained by a common cause structure, and adding extra causal relations from the mental phenomenon to the purported effect would be unwarranted. Our eventual model of dualist mental causation must motivate treating the correlation between mental phenomena and their target effects as causal despite these appearances. That is the first task imposed by the common cause worry.

Common cause scenarios pose a further challenge for models of dualist mental causation. If one attempts to adjust an account of causation such that it allows for dualist mental causation, one runs the risk of making it consider confounded common cause correlations as genuinely causal as well. Consequently, the resulting account of causation will be inadmissible. Similar challenges for accounts of dualist mental causation can be found in the literature. For example, when considering a proposal for dualist mental causation in terms of a mere regularity theory of causation, Lewis objects (1966, p. 25):

The position exploits a flaw in the standard regularity theory of cause. We know on other grounds that the theory must be cor-

rected to discriminate between genuine causes and the spurious causes which are their epiphenomenal correlates. (The “power on” light does not cause the motor to go, even if it is a lawfully perfect correlate of the electric current that really causes the motor to go.) Given a satisfactory correction, the nonphysical correlate will be evicted from its spurious causal role [...]

If a theory of causation counts confounded correlations, like the one between “power on” lights and revving engines, as causal, we should not trust that theory to deliver the right result in purported cases of mental causation either. In a similar vein, Kim (2007) and Bennett (2008, sect. 6.2) argue that accounts of causation that allow for dualist mental causation will count confounded correlations as causal. If the dualist is to provide a convincing account of mental causation, that model should provide a principled distinction between such cases of confounding and cases of dualist mental causation. That is the second task imposed by the common cause worry.

It is worth noting that the common cause worry does not rely on the same assumptions as the exclusion argument. Even if *Physical Completeness*, *Nomic Exclusion* and *Non-Overdetermination* turn out to be false, it might still be the case that the correlations between mental phenomena and their target effects are confounded by the nomic bases of mental phenomena. Common cause worries thus constitute a separate worry for dualist mental causation. A successful model of mental causation must strike the right balance in treating dualist mental phenomena as causes whilst still treating confounded correlations as non-causal. By properly motivating such a balance, one can respond to the common cause worry.

3.3 The set-up

The past two chapters presented the set-up of the mental causation problem for dualists.

We characterized the dualist position that is standard in mental causation debates, i.e. *Nomic Naturalist Dualism*. This position is *dualist* in that it states that the mental and the physical are metaphysically distinct. It is *naturalist* in that it states that the physical domain is complete and all non-

mental phenomena are metaphysically necessitated by physical phenomena. It is *nomio* in that it explains the pervasive psychophysical regularities correlations by positing fundamental laws of nature that relate occurrences of mental phenomena to occurrences with physical phenomena. For the purpose our investigation, we decided to treat the completeness of physics and the metaphysical distinctness of the mental and the physical as non-negotiable.

Once one considers these two features as non-negotiable, there appears to be no room for mental causation. In order for there to be mental causation in such an ontology, we require an account of causation according to which at least one of the following alleged truisms about causation is false: *Causal Closure*, *Causal Exclusion*, and *Non-Overdetermination*. Moreover, this account of causation should be able to provide a principled distinction between mental-to-physical correlations and spurious correlations that are confounded by a common cause, like the correlation between tar-stained teeth and lung cancer. Given this set-up the prospects look bleak.

In the next part of the dissertation, I will make matters worse. In particular, I will argue that the recent attempts to secure dualist mental causation against the backdrop of these assumptions are unconvincing. If prospects look bleak now, they should thus look worse at the end of Part II. Even so, I think there is still hope for dualist mental causation. To keep this hope alive, I will spend the final chapter of Part I providing a brief glimpse of my proposed solution. Afterwards, we will turn to Part II, in which we critically evaluate alternative proposals.

Chapter 4

A Glimpse of an Answer

My proposal is to argue for dualist mental causation by turning our attention to the notion of causation. When doing so, we should bear in mind that the notion of causation that is relevant to the mental causation debate must be one that respects our causal judgments (cf. Section 3.1). I will argue that a proper study of some of the problems facing general philosophy of causation indicates that the relevant notion of causation must therefore be *lightweight*, in that it does not impose a production or physical necessitation requirement on causation. Further, I will argue that such a study even allows for a notion of causation that is *thoroughly* lightweight, in that it does not even require causes to be tightly related to the phenomena that *do* produce or physically necessitate their target effects. This thoroughly lightweight account of causation provides an answer to the exclusion worry as well as the common cause worry, and thereby allows for dualist mental causation. In Part III of this dissertation, I will argue that objections to such a thoroughly lightweight account are undermotivated. Consequently, the dualist has a safe haven to retreat to in the mental causation debate. She can adopt this thoroughly lightweight account of causation and claim that her ontology allows for mental causation as well.

In this chapter, I draw some important distinctions between accounts of causation and provide a rough characterization of my eventual thoroughly lightweight model of dualist mental causation. First, I provide three individually sufficient conditions for an account of causation to be substantive

or heavyweight. Second, I qualify as lightweight those accounts of causation that are not heavyweight and remark that there are gradations in how an account of causation can be lightweight. Third, I provide a rough outline of how a *thoroughly* lightweight account of causation can solve the mental causation problems for dualism.

4.1 Heavyweight causation

Borrowing a metaphor from Hume, Mackie named his influential book on causation: ‘The Cement of the Universe’ (1974). I think the popularity of this metaphor is due to our deep-rooted conviction that causation plays a substantive role in the universe. Causation, it seems, holds the universe together across time. At the same time, we take causation to be the driving force of the universe: causes make their effects happen. If causation would stop working today, there would be no tomorrow. Or at least, the occurrence of tomorrow would have to be due to some cosmic coincidence, rather than being a natural consequence of today’s occurrence. All in all, causation plays a substantive role in our intuitive picture of the world: it drives one phenomenon to follow from the other, and it thereby holds the universe together across time.

The philosophical literature is rich in formulations that express this intuitive picture. Terms like ‘oomph’ (e.g. Schaffer, 2016, Section 1.1), ‘meaty’ (e.g. Baron and Miller, 2014, fn. 2), and ‘biff’ (e.g. Handfield et al., 2008) are used to characterize causation as we conceive of it pre-theoretically.¹ Less colloquial terms like ‘causal process’ (e.g. Dowe, 2009), ‘causal contribution’ (e.g. Ney, 2012) and ‘causal power’ (e.g. Harré and Madden, 1975), bear a similar connotation among those familiar with the literature. Furthermore, philosophers will often say that effects ‘derive their existence’ from their causes (e.g. Anscombe, 1993, p. 92), or that causes ‘produce’, ‘generate’ or ‘determine’ their effects (e.g. Kim, 2005, p. 18). I take all of these formulations to express the intuition that causation is a substantive or heavyweight relation that holds the universe together.

¹As it happens, *Oumph* — pronounced as one would pronounce ‘oomph’ — is a popular brand of vegan meat substitutes in Sweden. I take it this is a coincidence.

Proponents of what I will call ‘heavyweight’ accounts of causation aim to respect these deep-rooted intuitions. I will not canvas the philosophical work on substantive notions of causation here. Instead, I would like to distinguish three ways in which an account of causation can characterize causation as a substantive relation. One way is to focus on the locutions in terms of ‘production’ and ‘generation’ and subsequently require that causation is *productive*. Another way is to focus on locutions in terms of ‘determination’ and subsequently require that causes, in some sense, physically necessitate their effects. A third way is to require that causes either produce or physically necessitate their effects. In order for our project to succeed, it is important that none of these three characterizations is correct. Let us take a brief look at these options.

First, one can require that causes produce their effects. Here, production is understood as a transfer of physical energy or a physical quantity, as proposed by Dowe (2000, 2001), Fair (1979) and Salmon (1984).² If one imposes such a requirement on causation, it should come as no surprise that there can be no dualist mental causation. As Bennett notes (2008, fn. 19), Princess Elisabeth’s objection to dualism seems to have real bite if one assumes such a view of causation: it is very hard to see how dualist mental phenomena could transfer energy onto physical phenomena (cf. Chapter 1).³ Furthermore, the exclusion premise in exclusion arguments seems particularly convincing if one assumes such an account of causation. To see this, suppose that effect E is sufficiently caused (and thus sufficiently produced) by cause C_1 . Now suppose that C_2 also causes E . Whatever energy C_2 transfers onto E in order to cause it, would have to be energy that is redundant for E to come about. At least in as far as producing the effect goes, it seems that whatever causal work C_2 does is redundant for the occurrence of E .

Second, one can require that causes, in some sense, physically necessitate their effects. One straightforward way of doing so is by imposing the following

²Terms like ‘production’ and ‘produce’ are used liberally in the causation literature to denote relations that do *not* require any transfer of energy (e.g. Andreas and Günther, 2019; Woodward, 2015). It should be clear that this is not how the term is used in this text.

³However, the objection still loses some of its appeal if one is forced to accept action at a distance, as one would if Newtonian physics were true. The earth transferring energy onto the moon from a distance seems as mysterious to me as my dualist mental phenomena transferring energy onto my brain cells.

requirement on causation:

The Physical Necessitation Requirement For any two phenomena A and B , A causes B only if all physically possible worlds that contain A also contain B .

This is, in effect, to demand that causes physically necessitate their effects in the sense that we have outlined in Section 2.2.

Such requirements on causation were predominant for quite some time. In her 1970 inaugural lecture at the university of Cambridge, Anscombe reports that, since Hume's work on causation, nobody (with the exception of C.S. Peirce) "[...] called into question the equation of causality with necessitation" (1993, p. 90). The relevant notion of necessitation is typically taken to be underpinned by the fundamental laws of nature that govern the past-to-future development of the universe — i.e. those laws that we have assumed to be the fundamental laws of physics (cf. Chapter 2). Some notable philosophers who thought of causation in roughly these terms are Mill (1843), Russell (1912),⁴ Davidson (1967, 1970),⁵ and Armstrong (1996, 1999). However, much has changed since 1970. For one thing, Anscombe herself denounced such a view of causation in the very same lecture. Moreover, Lewis presented his influential counterfactual account of causation, which also entails a rejection of the physical necessitation view, only three years later (1973a). Both events have been critical to the development of philosophy of causation. By now there are many accounts of causation that do not impose such a physical necessitation requirement.

Even so, claims about causation often assume something in the neighbourhood. For example, the idea that causes must determine their effects (e.g. Kim, 2005, p. 18), or that causes are sufficient for their effects, seems to require that causation and necessitation are closely related. Moreover, the notions of necessitation and sufficiency are central to exclusion arguments.

⁴Russell went on to conclude that there is no such thing as causation. See Chapter 14 for a discussion.

⁵Although Heil (2013) argues that this reading of Davidson relies on a crucial misunderstanding. He states that Davidson required there to be some necessitation relation supporting every causal relation. In order for A to cause B , there must be some necessitation relation underlying the relation between A and B , but this does not require that A necessitates B . I am in no place to answer exegetical questions about Davidson's work, but I do not think it matters much here.

We will discuss the notions of sufficiency that are relevant to the exclusion argument at length in Chapter 7, but it is easy to see how it plays an important role in standard formulations of the argument. Recall the exclusion principle in our formulation:

Causal Exclusion For any three phenomena A , B and C : if A occurs at t and is a sufficient cause for B 's occurrence at $t + x$, no phenomenon C occurring at t that is metaphysically distinct from A and is metaphysically distinct from all of A 's parts is a cause of B , unless it is a case of genuine overdetermination.

In order for this principle to have any credibility, the cause that does the excluding must be sufficient for the effect in some important sense. Otherwise, how is its mere presence going to render any further cause an *overdetermining* cause? And if a cause is sufficient for an effect, it must, in some sense, necessitate that effect.

One way to hold on to this underlying idea that causes are sufficient for their effects without imposing a full-blown physical necessitation requirement on causation is to impose the following, more permissive, requirement on causation:⁶

The Sophisticated Physical Necessitation Requirement For any two phenomena A and B , A causes B only if there is some set of background conditions c that includes neither B nor anything that physically necessitates B , such that A and c together physically necessitate B .

The sophisticated physical necessitation requirement does not require that causes *on their own* physically suffice for their effects, but does require that they do so together with some fixed set of background conditions.

It should be clear that such a requirement on causation would exclude dualist mental phenomena from causing physical phenomena in worlds where *Causal Completeness* is true. According to nomic naturalist dualism, all

⁶I base this principle on Bennett's (2017, p. 60) 'necessitating-in-the-circumstance'-condition on building relations. Note that Bennett takes causation to be a building relation as well, I thus take it that she adheres to something that is at least close to this requirement on causation.

the sets of background conditions that, together with a given mental phenomenon, physically necessitate a physical effect also physically necessitate that physical effect in the absence of that mental phenomenon. For example, if nomic naturalist dualism is true, any set of physical phenomena that, together with my pain, physically necessitates my wincing, physically necessitates my wincing in the absence of my pain as well. Consequently, there is no set of background conditions such that it does not physically necessitate my wincing on its own *and* physically necessitates my wincing when combined with my pain. Dualist mental phenomena therefore cannot meet the sophisticated physical necessitation requirement on causation. We shall consider accounts of causation that require causes to physically necessitate their effects, against background conditions or not, to be heavyweight accounts of causation.

The third way for an account to be heavyweight is to impose the following disjunctive requirement on causation: every cause must either produce or physically necessitate its effect (against background conditions). I do not know of anyone who defends such a disjunctive view. Nonetheless, it is important for our project that this view is mistaken as well. Given that dualist mental phenomena cannot meet either the production criterion or the physical necessitation criterion for causation, there can be no dualist mental causation according to this disjunctive heavyweight account. I will argue against these heavyweight requirements on causation in Chapters 13 and 14.

By lumping together production and physical necessitation accounts of causation under the same denominator, I do not mean to suggest that production and physical necessitation are in some sense intimately related. Perhaps phenomenon A could produce phenomenon B indeterministically. Conversely, it might be possible that phenomena can physically necessitate other phenomena without producing them; it is up to the physicists to figure out whether or not this is actually possible.⁷ However, I take both the production requirement and the physical necessitation requirement to give voice to our conviction that causation is an ‘oomphy’, ‘biffy’, or ‘meaty’ relation. In fact, both requirements are often imposed with that motivation. For example,

⁷If I understand Ney (2009, 2012) correctly, the entirety of the productive causes will in fact be physically sufficient for their effects, and physically sufficient causes will also produce their effects. However, I do not want to put weight on this being the case here.

Strawson (1987, p. 255) takes both of them to be necessary conditions on any relation that deserves to be called Causation with a capital ‘C’. Moreover, exclusion arguments are often said to rely on either or both of these requirements on causation.⁸

Even so, I do not want to put much weight on the viability of this classification here. My main motivation to carve up accounts of causation as either heavyweight or non-heavyweight is ease of exposition. A significant part of our project will be to reject these heavyweight accounts of causation. In order to secure dualist mental causation one has to go lightweight on causation. However, *just* going lightweight will not do. As it turns out, we will have to go *thoroughly* lightweight.

4.2 (Thoroughly) lightweight causation

Lightweight accounts of causation are those accounts of causation that are not heavyweight. Rejecting heavyweight accounts of causation still allows for a variety of views on causation, and most of these still pay lip service to heavyweight accounts of causation — and thereby to the beefy intuitions underlying those accounts. In order to secure dualist mental causation, we have to distance ourselves from heavyweight accounts further than most standard lightweight accounts do. The subsequent account of causation will thus have to be *thoroughly* lightweight. In this section, I explain what I mean by ‘thoroughly lightweight’, as opposed to ‘just’ lightweight.

Lightweight accounts of causation do not impose a production or necessitation requirement on causation. Instead, lightweight accounts typically take causation to consist of the patterns of regularities or correlations themselves, rather than whatever process underlies it. A central challenge for such lightweight accounts of causation is to isolate those correlations that are in fact causal, like the correlation between smoking and lung cancer, from those that are *spurious* or non-causal, like the correlation between tar-stained teeth

⁸For some critical remarks on the exclusion argument’s reliance on productive accounts of causation, see Crane (1995, p. 18), List and Menzies (2009, p. 489), Kroedel (2020), Loewer (2002), Russo (2016), Woodward (2008, p. 264) and Chapter 13 in this dissertation. Kim explicitly acknowledges that he relies on a productive account of causation for his exclusion argument (e.g. Kim, 2002, p. 675). For a discussion on the exclusion argument’s reliance on the physical necessitation account, see Chapter 7.

and lung cancer. Heavyweight accounts can maintain that in the former case, but not in the latter case, there is a production or physical necessitation relation between the two phenomena. Lightweight accounts cannot. Or at least, they cannot maintain that this explains the difference between causation and spurious correlation in every possible case. They require a different explanation for the difference between causal correlations and spurious correlations.

I would like to make three further remarks on the current way of drawing the line between lightweight and heavyweight accounts of causation in this way. First, although standard lightweight accounts analyze causation in terms of patterns of correlation, it is not essential to lightweight accounts that they do so. Recall that we called all non-heavyweight accounts lightweight. One can therefore cook up views of causation that are lightweight but do not analyze causation in patterns of correlation. For example, if my view says that A causes B if and only if the common French word for A , starts with a ‘c’ and the common French word for B starts with an ‘e’, that view qualifies as lightweight. However, the view says nothing about patterns of correlations between occurrences of A and occurrences of B . The common French word account of causation is lightweight, but it does not analyze causation in patterns of correlation. Having said that, we will focus on more standard lightweight views on causation, and those do analyze causation in terms of patterns of correlation.

Second, lightweight accounts can allow for production or physical necessitation to be *sufficient* for causation. According to my categorization, lightweight accounts are only required to deny that production or necessitation is *necessary* for causation. One could for example maintain that a cause must either exhibit the right pattern of correlations with its effect *or* produce its effect.⁹ My motivation for allowing such sufficiency conditions in lightweight accounts of causation is mostly strategic. What is at issue in this project is whether or not dualist mental phenomena *can* be causes. Consequently, we should scrutinize the necessary conditions imposed by accounts of causation, but we can be liberal about sufficiency conditions. As long as

⁹Hall (2004) can be read as suggesting such an account of causation. However, note that his main aim in that paper is to distinguish two *concepts* of causation, rather than proposing a full-blown account of causation. Other proposals in this neighbourhood are made by Schaffer (2001b) and Loew (2019).

causation does not *require* production or physical necessitation, there is still hope for dualist mental phenomena to qualify as causes in virtue of standing in some other relation to the target effect.

Third, lightweight views do not entail that nothing holds the universe together. The lightweight theorist about causation is plausibly committed to saying that A causing B does not *require* that A is related to B by the cement-like relation that holds the universe together, if there is such a relation.¹⁰ However, this is equally compatible with there being a cement of the universe as with there not being a cement of the universe. In fact, for reasons addressed in the previous remark, the lightweight theorist can even allow that the cement of the universe is causal. She just has to allow that there can be other causal relations as well. To put it another way, the typical lightweight theorist is not committed to the universe consisting of patterns of correlations all the way down. She is just committed to causal relations being able to hold in virtue of some particular kind of correlation pattern holding, *regardless* of what is going on all the way down (cf. Beebe, 2006, p. 518).¹¹

There are several popular lightweight accounts of causation: counterfactual accounts (e.g. Lewis, 1973a, 1979, 2000), probability raising accounts (e.g. Mellor, 1995), agency accounts (e.g. Price and Weslake, 2009; von Wright, 1971), interventionist accounts (e.g. Woodward, 2003), etc. What unifies these theories is that they all take causation to be some pattern of correlations, rather than production or physical necessitation. What distinguishes them is what they consider to be the right kind of patterns of correlations. Each account imposes different requirements on causal correlations to distinguish them from spurious correlations. For example, counterfactual accounts are driven by the idea that causes make a difference to their effects. That is to say, if phenomenon A is a cause of phenomenon B, it has to be the case that, in nearby scenarios were A is lacking, B is lacking as well. Con-

¹⁰‘Plausibly’, because one could maintain that the cement of the universe consists of patterns of correlation. However, let us assume that this would not qualify as cement.

¹¹Opinions differ on what lightweight theorists should commit themselves to when it comes to cement. Beebe (2006, Section 2.3) argues that they should deny the existence of a fundamental cement even though their position on causation does not entail its absence. Demarest (2015, Section 1.2) emphasizes that questions about causation and what holds the universe together should be held separate. My sympathies happen to lie with the latter view, but I cannot argue for this here.

sequently, causation is analyzed in terms of patterns of correlations across scenarios that are very much like the actual one, but where the purported cause is lacking. The hallmark of counterfactual accounts is that exhibiting such a pattern of correlations suffices for causation.¹² Causation does not require that there is something over and above this pattern, like production or physical necessitation, in virtue of which the pattern holds. Similarly, other popular lightweight accounts of causation will maintain that phenomenon A's causing phenomenon B does not require anything over and above their exhibiting the pattern of correlations that these accounts propose to pick out.

We can thus characterize the central tenet of the typical lightweight accounts as follows: exhibiting the right kind of correlation patterns is sufficient for causation. This characterization might instill some hope for the dualist. Certainly, mental phenomena exhibit patterns of correlations with physical phenomena. In particular, our mental phenomena exhibit patterns of correlations with our behaviour. Given that we are mostly interested in establishing that mental phenomena cause behavioural phenomena, the popularity of the lightweight accounts of causation must be good news for the dualist.

This hope will be short-lived for most standard lightweight accounts of causation. When specifying what it means for a correlation pattern to be “of the right kind”, standard lightweight accounts of causation build in requirements that effectively eliminate the possibility of dualist mental causation. In particular, these accounts *de facto* require that causes stand in a tight relation, such as metaphysical necessitation or some slightly weaker variety thereof, to physical phenomena. Or at least, they do so in worlds where *Physical Completeness* is true. In Chapters 5 and 10, I will argue at length that two prominent lightweight accounts of causation, counterfactualism and interventionism, impose such a requirement. Given that dualists cannot allow mental phenomena to stand in a metaphysical necessitation relation to physical phenomena, and on the plausible assumption — to be motivated in Chapter 5 as well — that they cannot allow for them to stand in the relevant slightly weaker varieties of that relation to physical phenomena either, these standard lightweight accounts of causation do not allow for mental causation in a nomic naturalist dualist ontology. More detailed arguments will be

¹²They have a notoriously hard time providing necessary conditions in terms of such patterns. See for example Lewis (2000) and Schaffer (2001a,b).

provided (relatively) soon. For now, this summary will have to do.

If this summary is correct, *just* adhering to such standard lightweight accounts of causation will not do. In order to provide a credible model of dualist mental causation, we require an account of causation that can convincingly characterize “the right kind” of correlations *without* relying on requirements to the effect that causes must be tightly related to the phenomena that heavyweight accounts would deem causes. I will call such accounts of causation *thoroughly* lightweight accounts of causation, because they do not even require that causes stand in tight relations to phenomena that produce or physically necessitate their effects. Such thoroughly lightweight accounts of causation are not popular in mental causation debates, but they can be motivated by findings in recent philosophy of causation. That is what I argue in Part III of this dissertation. In the final section of this chapter, I provide a rough outline of how this proposal works.

4.3 Dualist mental causation

Here is how I propose to think of causation. Those correlations that can be exploited to reliably manipulate future phenomena are causal. That is to say, manipulations of causes reliably result in changes in their effects. On such a view, it seems plausible that dualist mental phenomena can be causes in a world where *Physical Completeness* is true. At the very least, we have some *prima facie* evidence that we can reliably manipulate behaviour by manipulating mental phenomena, *regardless* of whether or not nomic naturalist dualism is true. That is, more or less, what incites us to believe that mental phenomena are causes in the first place.

This line of defense has not been properly tested by the dualist. Perhaps somewhat surprisingly, the clearest summary and defense of this strategy against epiphenomenalism objections in philosophy of mind is by Dennett, whose writings are oftentimes not of the dualist-friendly kind (e.g. Dennett, 1978, 1988, 1991a, 2012).¹³ Dennett defends his mild realism about beliefs from the charge of epiphenomenalism by endorsing the kind of lightweight

¹³Perhaps Burge (1993) and Rudder Baker (1993) come close to suggesting a similarly lightweight view in response to mental causation worries, but they are less explicit about the importance of patterns of correlations and manipulability.

view I have in mind. According to Dennett, beliefs should be considered real only in virtue of exhibiting pervasive patterns of correlation with other phenomena, such as behaviour. Because these patterns are available for reliable manipulation across a broad variety of scenarios, these patterns, and the beliefs that feature in them, deserve to be called ‘real’. Faced with the objection that just exhibiting such a pattern with their targets effects would amount to to an epiphenomenalism about beliefs, Dennett states that exhibiting such real patterns with the target effects *just is* being a cause of these effects (1991b, fn. 22):

Several interpreters of a draft of this article have supposed that the conclusion I am urging here is that beliefs (or their contents) are epiphenomena having no causal powers, but this is a misinterpretation traceable to a simplistic notion of causation. If one finds a predictive pattern of the sort just described one has *ipso facto* discovered a causal power: a difference in the world that makes a subsequent difference testable by standard empirical methods of variable manipulation.¹⁴

I propose that the dualist takes a similar line of defense. She should maintain that, if we are right that we can reliably manipulate physical phenomena by manipulating mental phenomena, we are *ipso facto* right about there being mental causation, *regardless* of the status of dualism or physicalism. Pre-theoretically, one might have thought that there must be more to causation than just exhibiting such a pattern. For example, one might have thought that causation is the productive relation that drives these patterns, rather than consisting of the patterns themselves. However, a proper philosophical study of causation shows that these pre-theoretic intuitions are mistaken. Or so the dualist should argue.

The bulk of Part III will be dedicated to developing and defending such a thoroughly lightweight account of causation. In particular, I will motivate my proposed account with considerations about higher-level causation (Chapter 11), causation by omissions (Chapter 13) and the relation between causation and physics (Chapter 14). Once such an account of causation is defended,

¹⁴Ladyman and Ross (2007) further develop such a ‘real patterns’ account of causation, but it has yet to gain traction in either philosophy of mind or philosophy of causation.

the dualist can provide a powerful response to the mental causation worries. I will flesh out this response in Chapter 12, but one can see the rough outlines of such a proposal without too much effort. If it suffices that I can reliably manipulate B by manipulating A in order for A to cause B, then the exclusion premise in exclusion arguments is likely to be false. For it might be the case that A is metaphysically distinct from the co-occurrent sufficient physical cause of B *and* that A still causes B. Dualist mental phenomena in a physically complete world might very well be a case in point: putting me in pain certainly seems like a reliable way to make me wince. Moreover, spurious correlations due to a common cause are standardly *not* reliably exploitable for this kind of manipulation: getting an insurance at an agency that mostly insures healthy people is not a reliable way of increasing your life expectancy (cf. Cartwright, 1979). So the thoroughly lightweight account does suggest a way to distinguish between dualist mental causation and correlations that are confounded by a common cause. If this proposal can be made to work, the future looks brighter for the dualist.

Even so, it is important to be clear on what such an account does deliver and what it does not deliver. The account *allows* for dualist mental causation in worlds where *Physical Completeness* is true. By doing so, it provides a response to the exclusion worry and to the common cause worry. Given the current state of the debate about dualist mental causation, providing this much is a significant assignment. As we have discussed in the introduction, it is standardly accepted by dualists and non-dualists alike that dualism of the kind we are considering must result in epiphenomenalism. Once this assignment is completed, the future looks brighter for the dualist, but that does not mean that my thoroughly lightweight account delivers dualist mental causation.

Here is why the account does not deliver dualist mental causation. First of all, in order for there to be dualist mental causation, dualism has to be true. Our thoroughly lightweight account only provides a response to a central argument *against* dualism, namely that it results in epiphenomenalism. It does *not* provide us with a direct argument for the truth of dualism. Second, even if dualism is true and this account of causation is adequate, the truth of *Mental Causation* will still depend on how the correlation patterns

between mental phenomena and behavioural phenomena turn out. They need to be *reliable* in the right way. A central step in developing our thoroughly lightweight account of causation will be to pin down the right kind of reliability. I will dub that kind of reliability ‘robustness’, and, as we shall see in Chapter 12, it is an open empirical question whether or not the correlations between mental phenomena and behaviour are robust. The answer might very well differ across different psychophysical correlations. Perhaps the correlations between pains and wincings are not robust, but the correlations between hunger and eating are. Empirical research will have to tell.

Note however, that this outstanding challenge to mental causation is not specific to dualism. Even though the focus in mental causation debates is often on philosophical problems like the exclusion argument, it is acknowledged that empirical results will still have bearing on the issue as well. For example, Woodward (2008, Section 7) argues that the plausibility of non-reductionist physicalist mental causation depends on whether or not the psychophysical correlations are reliable in more or less the same way as my proposed account of causation requires causal correlations to be reliable.¹⁵ According to my proposed thoroughly lightweight account of causation, the plausibility of *Mental Causation* thus boils down to the same empirical issues for dualists and non-reductionist physicalists. This would mean that the dualist recovered a substantial piece of ground in the mental causation debate. If the thoroughly lightweight account of causation can deliver that much, it is well worth it for the dualist to investigate the viability of this account.

But first, let us take a look at some alternative approaches. If the dualist’s mental causation problems are already convincingly solved, it is perhaps not worth it to delve into philosophy of causation for her benefit. I will argue that the outstanding proposals for dualist mental causation can be improved upon.

¹⁵List and Menzies (2009) also argue that it is an empirical question whether or not there non-reductionist physicalist mental phenomena are causally excluded by their underlying physical phenomena.

Part II

Contemporary Solutions

Chapter 5

On Causal Exclusion

In this part of the dissertation, I critically evaluate recently proposed solutions to the mental causation problems for dualism. First, I discuss List and Stoljar's criticism on the exclusion principle in exclusion arguments that target dualism. Second, I discuss Kroedel's proposal to avoid mental causation problems by adopting *supernomological* dualism. Third, I discuss recent criticism on the notion of causal sufficiency. Finally, I discuss Lowe's three models of dualist mental causation. These proposals share one goal: to safeguard *Mental Causation* whilst respecting both *Dualism* and *Physical Completeness*.¹ I will argue that these proposals fail to do so convincingly. Let us start with List and Stoljar's proposal.

It is customary to maintain that causal exclusion worries pose a more serious threat to dualist theories of mind than to non-reductionist physicalist theories of mind. Some dualists oppose this custom and maintain that, in as far as non-reductionist physicalists have a response to causal exclusion arguments, dualists must have one as well (e.g. Koons and Bealer (2010, p. xix–xx); Pautz (2010, p. 65)).² Recently, List and Stoljar have provided a concrete formulation of this contention by challenging the *Causal Exclusion* premise in a way that mirrors a popular non-reductionist physicalist strategy. That is to say, they argue that metaphysically distinct phenomena *can*

¹Although, as we shall see, Lowe puts some stretch on what it means to respect *Physical Completeness*.

²Horgan (2010, sect. 4) seems at least open to this possibility.

non-overdeterministically cause the same target effect. They conclude that dualists can convincingly respond to the exclusion argument.

In this chapter, I outline List and Stoljar's proposal and argue that it ultimately fails.³ Metaphysically distinct phenomena *can* non-overdeterministically cause the same effects. However, for all List and Stoljar say, metaphysically distinct phenomena can only do so if they stand in a sufficiently tight relation to one another. It is unlikely that nomic naturalist dualism can allow mental phenomena to stand in such a tight relation to their physical nomic bases. One can exploit this difference in tightness to formulate an exclusion argument that excludes dualist mental phenomena from being causes without excluding patently causal properties from being causes. I propose such a formulation and conclude that exclusion worries for dualism persist.

5.1 Exclusion and metaphysical distinctness

Recall the exclusion argument as we presented it earlier:

Dualism Mental phenomena are metaphysically distinct from physical phenomena.

Causal Closure Every physical phenomenon has a sufficient physical cause at any given time t (if it has a cause at all at t).

Causal Exclusion For any three phenomena A , B and C : if A occurs at t and is a sufficient cause for B 's occurrence at $t + x$, no phenomenon C occurring at t that is metaphysically distinct from A and is metaphysically distinct from all of A 's parts is a cause of B , unless it is a case of genuine overdetermination.

Non-Overdetermination There is no systematic genuine overdetermination of physical effects with mental causes.

No Mental Causation Mental phenomena do not systematically cause physical phenomena in the actual world.

³My exposition and criticism here expand on my earlier discussion of List and Stoljar's proposal in Vaassen (2019).

A similar exclusion argument has been taken to threaten non-reductionist physicalist theories of the mind. As we have seen in Chapter 2, non-reductionist physicalists maintain that mental phenomena, whilst metaphysically necessitated by physical phenomena, are not identical to physical phenomena. For any mental phenomenon and any physical phenomenon, the non-reductionist physicalist maintains that these phenomena are numerically distinct.⁴ The following exclusion argument can thus be levied against such non-reductionist physicalism (cf. Kim, 2005; Ney, 2009, 2012; Papineau, 2002):

Non-Reductionism Mental phenomena are numerically distinct from physical phenomena.

Causal Closure Every physical phenomenon has a sufficient physical cause at any given time t (if it has a cause at all at t).

Numeric Causal Exclusion For any three phenomena A , B and C : if A occurs at t and is a sufficient cause for B 's occurrence at $t + x$, no phenomenon C occurring at t that is numerically distinct from A and is numerically distinct from all of A 's parts is a cause of B , unless it is a case of genuine overdetermination.

Non-Overdetermination There is no systematic genuine overdetermination of physical effects with mental causes.

No Mental Causation Mental phenomena do not systematically cause physical phenomena in the actual world.

In response, many non-reductionist physicalists have argued that the required exclusion premise is implausibly strong. That is to say, it excludes patently causal phenomena from being causes. For example, phenomena like hurricanes, terminal diseases and banking crises have wide-spread effects, but none of these phenomena are numerically identical to their underlying physical phenomena. After all, any hurricane, financial crisis, or terminal disease could have been metaphysically necessitated by a physical phenomenon

⁴I ignore here the possibility of maintaining that only *some* mental phenomena are numerically distinct from any given physical phenomenon. This is merely for ease of formulation.

that contains some greater or fewer number of electrons in its actual physical metaphysical base. Nonetheless, these phenomena and their underlying physical phenomena do not genuinely overdetermine their effects. Instead, these appear to be harmless cases of trivial overdetermination. The upshot is that in familiar cases of non-overdeterministic causation, like hurricanes causing destruction and banking crises causing job losses, there is a sufficient physical cause for the target effect, *and* a cause that is not numerically identical to that sufficient physical cause. These familiar cases thus contradict *Numeric Causal Exclusion* and non-reductionist physicalists can credibly maintain that exclusion arguments against their position require an unreasonably strong exclusion premise.

List and Stoljar argue that the dualist can make a similar case. They maintain that *Causal Exclusion* is an implausibly strong premise and conclude that dualists can respond to exclusion arguments in the same way as non-reductionist physicalists. I will argue that this strategy fails. Even though *Causal Exclusion* is indeed implausibly strong, exclusion arguments against dualism can rely on a weaker exclusion principle that does not encounter the same problems. To understand why this is the case, we can take a look at List and Stoljar's counterexample to *Causal Exclusion*.

5.2 Metaphysically distinct co-causes

Let us say that phenomena that non-overdeterministically cause the same effect against the same background conditions 'co-cause' that effect. Some cases demonstrate that metaphysically distinct phenomena *can* co-cause effects.

Consider the following case derived from List and Stoljar (2017, p. 105). A certain university is organized such that the committee delegated to make tenure decisions always consists of the most successful professors. Given this organizational structure, these professors making a negative decision simultaneously makes it the case that the university made a negative decision. If, in such a case, an applicant loses her job due to the university's decision being negative (*UD*), the most successful professors's decision being negative (*PD*) would *also* count as a cause, despite its being metaphysically distinct from

UD. The example thus provides us with metaphysically distinct phenomena that co-cause an effect and thereby disproves *Causal Exclusion*.

Moreover, the example lines up well with the philosophical literature on *realization*. This allows us to embed the counterexample to *Causal Exclusion* in an established theoretical framework and affords us a closer look at what enables two phenomena to co-cause in such cases.

In cases of realization, one can distinguish between the realized phenomenon, its *total* realizer and its *core* realizer (cf. Shoemaker, 2007, p. 21–22).⁵ The realized phenomenon can be any non-fundamental phenomenon, like *UD*. The *total* realizer of the realized phenomenon is typically a large and complex set of phenomena that metaphysically necessitates the realized phenomenon and is therefore *not* metaphysically distinct from it. The total realizer of *UD* for example, will include *PD* as well as relatively permanent phenomena, such as the organizational structure of the university. The *core* realizer is a salient, non-redundant part of this total realizer, such as *PD*. This phenomenon *is* metaphysically distinct from the realized phenomenon, because its occurrence does not on its own suffice for the occurrence of the realized phenomenon nor does the occurrence of the realized phenomenon suffice for the occurrence of the core realizer. That is to say, there are possible worlds where the professors make the same decision, but the university is organized differently and the university makes a different decision, and there are worlds where the university makes the same decision, but the university is organized differently and the professors make the same decision.

Despite this metaphysical distinctness, it appears that core realizers and their realized phenomena can co-cause effects. This conclusion is supported by two heuristic principles that are widely accepted in the mental causation debate. Let *p* be a phenomenon and *P* the proposition that *p* occurs, and similarly for *q* and *Q*. The first principle is that *p* non-overdeterministically causes *q* if and only if *Q* counterfactually depends on *P*.⁶ The second principle is that counterfactual dependence is defined in accordance with Lewis (1973a):

⁵The realization literature tends to focus on *properties* rather than phenomena. Given that we use ‘phenomenon’ merely as a place-holder, it is harmless to stick to our ‘phenomenon’-terminology.

⁶List and Stoljar rely on this heuristic as well (2017, p. 103–104).

Counterfactual Dependence Q counterfactually depends on P *iff*
 $P \Box \rightarrow Q$ and $\neg P \Box \rightarrow \neg Q$

where the semantics of $\Box \rightarrow$ are such that

$P \Box \rightarrow Q$ is true *iff* there is a possible world where P and Q hold which is closer to the actual world than any possible world where P and $\neg Q$ hold (or there are no possible worlds where P holds).

The idea underlying these principles is that, in most normal circumstances, causes are *difference-makers* for their effects, and counterfactual dependence as captured by the above semantics serves as a reliable proxy for difference-making.

We can tentatively rely on these heuristic principles as follows. If these heuristic principles indicate that p causes q , that provides *prima facie* defeasible evidence that p causes q . Conversely, if these heuristic principles indicate that p does not cause q , that provides *prima facie* defeasible evidence that p does not cause q . If we want to maintain that p *does* cause q we thus require an explanation why the case fails the counterfactual test for causation. This explanation can take different forms. We could for example argue that the case of p and q is a type of case where the counterfactual test systematically delivers the wrong results. Alternatively, we could argue that the case of p and q meets some other reliable test for causation. In the absence of such overriding evidence, we take a failure to pass the counterfactual test for causation to indicate that there is no causal relation between the phenomena under consideration.

In order to use this test, we need some account of what makes one possible world closer to actuality than the other. In mental causation debates, it is customary to adopt Lewis's analysis of the closeness of possible worlds, which relies on the number and size of the 'miracles' that separate possible worlds from the actual world.⁷ Miracles are to be understood as violations of nomic

⁷See Bennett (2003, 2008) and Kroedel (2015). Lewis's analysis is not uncontroversial, as it can be difficult to assess whether one miracle creates more spatiotemporal dissimilarity than another. However, the analysis does provide clear results in cases of dualist mental properties and their target effects (see Section 5.3). See also Kroedel (2020) for an explicit defense for the use of this metric in cases of mental causation. List and Stoljar do not

laws, such as gravity locally failing to attract my body to the earth. In order to determine what distance a miracle creates between two worlds, Lewis proposes the following guidelines (1979, p. 472):

- I. It is of the first importance to avoid big, widespread, diverse violations of law.
- II. It is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails.
- III. It is of the third importance to avoid even small, localized, simple violations of law.
- IV. It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly.

According to these guidelines, the closest possible world where a realized property is excised is typically one where its core realizer is absent as well. Suppose that we are looking for the closest possible world where *UD* did not occur. At least some part of *UD*'s total realizer will have to be absent from that world as well. After all, its total realizer metaphysically necessitates *UD*, and metaphysically impossible worlds are standardly ignored when one evaluates counterfactual dependence.⁸ Therefore, this world will lack either the core realizer, i.e. *PD*, or some of the more permanently ongoing phenomena making up the total realizer, like the organizational structure of the university. *Prima facie*, it will be easier to maintain maximum match of particular fact by excising the core realizer, because changing more permanently ongoing phenomena, like organizational structures, is likely to result in more extensive mismatches of particular fact. For instance, changing the organizational structure of the university such that the committees are composed differently, or their decision no longer settles the university decision, is likely to affect *several* university decisions on tenure applications, rather than just this one. All of these different decisions will spread into further differences in

indicate that their strategy requires a different analysis of counterfactuals, and such a deviation would require a separate defense.

⁸See Lewis (1973b, sect. 1.6), Woodward (2008, p. 254–256) and Chapter 10.

matter of fact: new faculty members get hired, lectures are given by different professors, unsuccessful applicants move to other cities, etc. The resulting world will probably be more different from ours than a world where the organizational structure remains identical, but the most successful professors make a different decision in this specific case. Consequently, the closest possible world where the realized property is excised typically is a world where the core realizer is absent as well.

Given that counterfactuals are evaluated by looking at the closest possible world where the antecedent is true, this means that realized phenomena and their core realizers will often enter into the same relations of counterfactual dependence. For example, in the university decision case, an applicant's job loss will be counterfactually dependent on both *UD* and *PD*:

- (i) $UD \Box \rightarrow \text{JOB LOSS}$
- (ii) $\sim UD \Box \rightarrow \sim \text{JOB LOSS}$
- (iii) $PD \Box \rightarrow \text{JOB LOSS}$
- (iv) $\sim PD \Box \rightarrow \sim \text{JOB LOSS}$

Hence, the example not only conforms with our intuitions and the realization literature, it is also supported by the relevant counterfactuals. Unless one is willing to disregard this evidence and maintain that realized properties are causally excluded by their core realizers, the current exclusion argument has no force against the dualist. We can reasonably conclude that *Dualism*, *Physical Closure*, *Non-Overdetermination*, and *Causal Exclusion* cannot provide a sound argument for *No Mental Causation*, because *Causal Exclusion* is false.

5.3 Exclusion worries persist

According to our nomic naturalist dualism, mental phenomena and their underlying physical phenomena stand in a nomic necessitation relation. List and Stoljar maintain that, if the dualist is willing to maintain that these psychophysical necessitation relations are reciprocal, we should expect dualist mental phenomena and their underlying physical phenomena to co-cause their

effects. They take this claim to be supported by the relevant counterfactuals. Concerning cases where two phenomena F and F^* nomically necessitate one another, they say (2017, p. 104):

[...] to the extent that we are prepared to say, of F , that ‘if it were not instantiated, E would not have happened’, we should be prepared to say exactly the same thing of F^* .

However, the considerations which lead us to reject *Causal Exclusion* do not support this claim. In fact, the relevant counterfactuals contradict it. Consider the following dualist example: my pain is nomically necessitated by the underlying physical phenomenon ‘phys’. Phys is in turn nomically necessitated by my pain. Suppose further that phys is a necessary part of a physically sufficient condition for my wincing a moment after phys and my pain are instantiated.

Despite the reciprocal nomic necessitation relation between my pain and phys, these two phenomena exhibit relevantly *different* patterns of counterfactual dependence, because the closest possible world where my pain is absent is *not* a world where phys is absent. After all, it takes but a small localized miracle in a psychophysical law to excise my pain and hold all physical facts, including the occurrence of phys, fixed. The resulting possible world will still contain phys *and* my wincing, as it still contains a sufficient physical cause for my wincing. Compare that possible world with the closest possible world where both my pain and phys are absent. We can assume that both phenomena can be excised with one small, localized miracle preceding the occurrence of phys, because my pain would not have occurred in the absence of the underlying physical phenomenon. Just like the possible world lacking pain, this world only requires one small, localized *miracle*. Even so, the resulting world is further removed from actuality than the world that just lacks my pain, because it contains strictly more mismatch in *particular fact*. In particular, the resulting world lacks *both* my pain *and* phys, rather than just lacking my pain. Furthermore, the absence of phys will *spread* throughout this possible world. For example, given that the occurrence of phys is a necessary part of the sufficient condition for my wincing to occur, I will not wince in the resulting world, which makes for a further mismatch of particular fact. The closest possible world lacking both phys and pain will thus also lack my wincing, but

it is *not* the closest possible world where pain is lacking, as there is a closer possible world that lacks pain but contains both *phys* and my wincing.

By contrast, the closest possible world where *phys* is absent *will* lack my wincing, because *phys* is a necessary part of the physically sufficient condition for my wincing.⁹ Consequently, my wincing counterfactually depends on *phys*, but not on my pain:

- (v) PAIN $\Box \rightarrow$ WINCE
- (vi) \neg PAIN $\Box \rightarrow$ WINCE
- (vii) PHYS $\Box \rightarrow$ WINCE
- (viii) \neg PHYS $\Box \rightarrow$ \neg WINCE

Assuming that counterfactual dependence is a reliable test for non-overdeterministic causation, these counterfactuals support the exclusionist conclusion that my pain and *phys* do *not* co-cause my wincing. We are thus in need of further support for the conclusion that my pain *does* cause me to wince.

5.4 Weak Exclusion

We can summarize our findings as follows. The relation between realized properties and their core realizers is *tighter* than nomic necessitation in that it puts stronger restrictions on those nearby possible worlds where the first relatum is instantiated and the second is not. In the case of nomic necessitation, these worlds are but a small localized miracle away. In the case of realization, the miracle excising the realized property will make for some further mismatch of particular fact in the total realizer, which increases the departure from actuality. The counterfactuals indicate that it is exactly this further tightness that allows the relata of realization to co-cause effects. Given that our current dualist model does not posit such a tighter relation between conscious properties and physical properties, the upshot appears to be that dualist conscious properties and physical properties cannot co-cause effects.¹⁰

⁹For reasons just addressed, that possible world plausibly lacks my pain as well. However, this is not essential to my argument.

¹⁰Bennett (2008) argues for a similar conclusion by relying on a counterfactual test for genuine overdetermination. Keaton and Polger (2014) use cases of realization to demon-

We should thus expect there to be an exclusion principle that exploits this difference in tightness. The relevant principle would target dualist mental phenomena, whilst allowing for standard cases of co-causing by core realizers and their realized property. Here is one strategy for devising such a principle.

First, we can define tightness as follows:

Tightness For any two phenomena A and B , A is tightly related to B if and only if there is some set of background conditions c that includes neither A nor anything that metaphysically necessitates A , such that B and c together metaphysically necessitate A .

We can stipulate that background conditions cannot include any nomic laws, but only particular matters of fact about which phenomena occur when and where.¹¹ With this definition in mind, we can say that realized phenomena are tightly related to their core realizers. After all, the core realizer is a salient, *non-redundant* part of the total realizer, and the total realizer metaphysically necessitates the realized phenomena. This is just another way of saying that those parts of the total realizer that are not the core realizer do not on their own metaphysically necessitate the realized property, but do so when taken together with the core realizer. Consequently, realized phenomena, like the university decision, stand in a tight relation to their core realizers, like the professor's decision.

I take *Tightness* to carve out an important and familiar feature of some philosophically significant relations, such as realization, partial grounding, and part-whole relations. These relations are such that their holding entails that one relatum metaphysically necessitates the other, given some background conditions. It is that feature that *Tightness* is supposed to carve out. If we allow for the set of background conditions to be empty (and I do not see why we should not), the following relations will turn out to be tight in this sense as well: metaphysical necessitation, grounding, the determinate-

strate that realized properties are often still excluded by their realizers according to Bennett's proposal, which might therefore still be considered to impose too strong requirements on co-causes. My proposal imposes weaker restrictions on co-causes and the resulting exclusion argument thus relies on weaker assumptions.

¹¹As in Chapter 2, I assume that there is a relatively straightforward way to distinguish between laws and particular matters of fact. This might make for complications if one assumes a broadly Humean picture of laws.

determinable relation, supervenience, etc.¹²

I also take *Tightness* to capture the difference between realization and nomic necessitation that was borne out in the previous section. After all, if phenomenon *B* is tightly related to phenomenon *A*, and *B* occurs in the actual world in virtue of *A* occurring in the actual world, then all worlds that do not contain *B* but do contain *A* are more than a small localized miracle removed from actuality, as these worlds require some further mismatch in the particular facts that make up the background conditions against which *A* metaphysically necessitates *B*.¹³ By contrast, in at least some cases where *A* nomically necessitates *B*, and *B* occurs in the actual world in virtue of *A* holding in the actual world, there is a possible world containing *B* but not *A* that is only a small localized miracle removed from actuality. That is what we learned from the case of pain and phys. As we have seen in the previous section, this difference between tight relations and nomic necessitation affects the truth values of the counterfactuals that are relevant when assessing purported cases of co-causation. To mark this difference between nomic necessitation and tight relations, we will say that one phenomenon is *merely* nomically necessitated by the other if and only if it is nomically necessitated by that phenomenon, but there are no tight relations between these phenomena.

With these considerations in mind, we can reformulate the exclusion principle such that it allows for co-causes that are tightly related, but not for co-causes that stand in a mere nomic necessitation relation. We can do so as follows:

Weak Exclusion For any three phenomena *A*, *B* and *C*: if *A* occurs at *t* and is a sufficient cause for *B*'s occurrence at *t* + *x*, no phenomenon *C* occurring at *t* that is not tightly related to *A* and is not tightly related to any of *A*'s parts is a cause of *B*, unless it is a case of genuine overdetermination.

Weak Exclusion is supported by the relevant counterfactuals and is consis-

¹²See also Bennett (2017, p. 60), who imposes a 'necessitation-given-the-circumstances' requirement on all 'building' relations. My definition of tightness is loosely based on that requirement.

¹³This might be true trivially. For example, if *B* is metaphysically necessitated by *A*, there are no such worlds. Consequently, all of those worlds are trivially more than a small localized miracle away.

tent with realization cases. It thus remains unaffected by List and Stoljar's arguments against *Causal Exclusion*.

Plausibly, *Weak Exclusion* threatens nomic naturalist dualism. This is because such a dualism is plausibly committed to there being no tight relations between mental phenomena and physical phenomena. To see this, consider the candidates for background conditions against which a physical phenomenon could metaphysically necessitate a mental phenomenon within such an ontology. These background conditions cannot be physical, for that would mean that mental phenomena are metaphysically necessitated by a set of physical phenomena and would thus contradict the metaphysical distinctness of the mental and the physical. On the plausible assumption that metaphysical necessitation is transitive, these background conditions cannot be metaphysically necessitated by physical phenomena either, for the same reasons. The only other matters of fact left in the dualist's ontology to play the role of background conditions are mental facts. But how would these metaphysically necessitate other mental facts when taken together with physical phenomena but not in the absence of these physical phenomena? It is hard to see how physical phenomena can be allowed to play such a non-redundant part in the metaphysical necessitation of dualist mental phenomena. Consequently, proponents of dualism do not posit any of the relations that we listed as tight between mental and physical phenomena, and quite often they explicitly deny that such relations hold between them. Similarly, standard treatments of dualist ontologies in the mental causation debate, like Bennett (2008); Bourget (2019) and Kroedel (2015, 2020), assume that individual dualist mental phenomena like pain can be excised from their actual worlds with merely a small localized miracle. For now, we have followed suit, and List and Stoljar have not provided us with reasons to deviate from these standard treatments.

Note that even if the dualist can credibly maintain that there are tight relations between mental and physical phenomena, there would remain a significant difference between the case of mental phenomena and cases like the university decision. As argued above, the background conditions against which physical phenomena would metaphysically necessitate mental phenomena on such a picture cannot be physical or metaphysically necessitated by

physical phenomena. By contrast, the university decision plausibly *is* metaphysically necessitated by physical phenomena, as its total realizer is metaphysically necessitated by physical phenomena, and metaphysical necessitation is plausibly transitive.¹⁴ Consequently, the dualist would still be at peril of exclusion arguments that exploit this difference even if she were allowed to posit tight relations between the mental and the physical. In Chapter 7, I briefly discuss how such an exclusion argument follows from addressing some issues surrounding causal sufficiency. However, given that the focus in this chapter is on List and Stoljar's proposed response to exclusion arguments, and they make no suggestion of deviating from the standard view that dualists cannot posit tight relations between the mental and the physical, I set this issue aside for now.

In the absence of evidence to the contrary, we can tentatively state that nomic naturalist dualism is committed to the following claim:

Nomic Dualism Mental phenomena are merely nomically necessitated by physical phenomena.

Consequently, the brand of dualism we set out to defend is threatened by the following exclusion argument:

Nomic Dualism Mental phenomena are merely nomically necessitated by physical phenomena.

Causal Closure Every physical phenomenon has a sufficient physical cause at any given time t (if it has a cause at all at t).

Weak Exclusion For any three phenomena A , B and C : if A occurs at t and is a sufficient cause for B 's occurrence at $t + x$, no phenomenon C occurring at t that is not tightly related to A and is not tightly related to any of A 's parts is a cause of B , unless it is a case of genuine overdetermination.

Non-Overdetermination There is no systematic genuine overdetermination of physical effects with mental causes.

¹⁴I am assuming here that university decisions are not metaphysically dependent on phenomenally conscious phenomena.

No Mental Causation Mental phenomena do not systematically cause physical phenomena in the actual world.

We are therefore still in need of arguments against one the above premises.

In the next chapter, we look at a proposal recently defended by Kroedel (2015, 2020), which denies both *Nomic Dualism* and *Weak Exclusion*. The above discussion provided evidence that the nomic necessitation relation is too weak to allow for co-causing. Perhaps the dualist can adjust her position on the modal status psychophysical laws such that mental phenomena and their nomic bases can co-cause effects, *without* giving up on her central commitment that the mental and the physical are metaphysically distinct. Kroedel proposes such a strategy.

Chapter 6

Supernomological Dualism

Kroedel claims that “[...] dualism can explain mental causation and solve the exclusion problem” (2015, p. 357).¹ He maintains that the dualist can do so by making a slight adjustment in her ontology of the mind. According to Kroedel, the dualist should declare psychophysical laws to be modally stronger than other nomic laws, but still weaker than metaphysical laws. By doing so, the dualist can allow mental phenomena to be causes according to a counterfactual account of causation, whilst maintaining that mental phenomena are metaphysically distinct from physical phenomena.

In this chapter, I argue that Kroedel’s proposal indeed delivers dualist mental causation, but, as is recognized by Kroedel himself, it requires a substantial *ad hoc* hypothesis. All else equal, this proposal would be improved upon by a model that does not require such an *ad hoc* hypothesis.

6.1 Strengthening the laws

Kroedel (2015, 2020) maintains that counterfactual dependence, defined according to the Lewisian framework we discussed in the previous chapter, provides a sufficient condition for causation within the context of the mental

¹Note that this is a stronger claim than List and Stoljar’s, who only maintained that the dualist can respond to exclusion arguments in the same way as non-reductionist physicalists.

causation debate.² If A counterfactually depends on B , then A causes B . Starting from this posit, the task for the dualist is to argue that behavioural phenomena counterfactually depend on mental phenomena according to her ontology.

In the previous chapter, we discussed the problem for dualism on a counterfactual account of causation. If dualism is true, my wincing is not counterfactually dependent on my pain, as the counterfactual ($\neg \text{PAIN} \square \rightarrow \neg \text{WINCE}$) comes out as false. This is because there is a possible world that is physically identical to ours, but, due to a localized lapse in the nomic psychophysical law connecting phys and pain, my pain does not occur. Call that world W_{NoPain} . In W_{NoPain} , I would still wince, given that the physical law connecting phys (+ background conditions) and my wince is still in place. In order for my pain to cause my wincing, there should be a possible world that contains neither pain nor my wincing and is closer to ours than W_{NoPain} , but it is easy to see that this cannot be the case. The mere fact that my wince would be absent from this world already makes it more distant than W_{NoPain} in virtue of the rule on spatiotemporal similarity; and then we have not even considered the miracle it would take to make my wince disappear.

The problem clearly lies with the weakness of the modal law connecting my pain and phys. If one could claim that cutting loose mental phenomena requires a larger miracle than taking out both the mental phenomenon and its purported effect, the problem would be solved. As Kroedel puts it (2015, p. 361):

[...] assuming dualism, the critical condition for establishing that behavioural events counterfactually depend on, and hence are caused by, mental events is that worlds where the actual psychophysical laws are violated are always less similar overall to our world than worlds without such violations, irrespective of violations of ordinary laws of nature.

Unfortunately, it is part and parcel of *Nomic Naturalist Dualism* that the laws connecting mental phenomena and their underlying physical states are of an equal modal strength as the fundamental laws of physics, which govern

²While he presents this as a working assumption in Kroedel (2015), the first two chapters in Kroedel (2020) provide an explicit defense of this claim.

the diachronic development of our universe. So, as the position stands, it cannot meet this critical condition.

Kroedel proposes to avoid this problem by adopting a position he calls ‘supernomological dualism’. But for one small adjustment, supernomological dualism is the same theory of mind as nomic naturalist dualism. Supernomological dualism bestows a special modal status upon psychophysical laws. These laws should be considered as modally stronger than other nomic laws, like the law of gravity, whilst still modally weaker than metaphysical laws, like the law that all squares are rectangles. This would mean that there is a world in which the physical phenomenon underlying my pain does not give rise to my pain, but that world it is further removed from our world than a world where a particle accelerates across the speed of light threshold. Kroedel notes that this proposal does not interfere with the metaphysical distinctness of the mental and the physical (Kroedel, 2020, sect. 2.5):

[...] nothing forces dualists to accept that psychophysical laws are modally on a par with ordinary laws of nature, such as the laws of physics. They are within their rights to claim that psychophysical laws could not have failed so easily as the other laws. They can claim, in other words, that worlds where the psychophysical laws are violated are further from actuality than any worlds where only the ordinary laws are violated.

Kroedel’s proposed adjustment has an important impact on the relevant counterfactuals. If the necessitation relation between mental phenomena and their nomic bases is modally stronger than regular nomic relations, the geography of possible worlds is reorganized: worlds where psychophysical laws miraculously lapse are more distant than worlds where other nomic laws lapse. So a counterfactual claim like: ‘if my pain had not occurred, I would not have winced’ comes out as true. After all, the possible world where both *phys* and *wince* are absent due to a lapse in the physical laws is now closer to the actual world than a world where *phys* is present, but *pain* is not. Consequently, the following counterfactuals hold:

- (i) PAIN $\Box \rightarrow$ WINCE
- (ii) \sim PAIN $\Box \rightarrow$ \sim WINCE

Given Kroedel's proposed assumption that counterfactual dependence suffices for causation, this means that my pain causes my wincing. These observations plausibly generalize to most purported cases of mental causation. Consequently, supernomological dualism in conjunction with a counterfactual account of causation secures *Mental Causation*. The dualist thus has a model of mental causation at her disposal.

This model has one drawback. Although it delivers the right results, it lacks independent motivation. The proposed adjustment is motivated only by its ability to solve the causal exclusion problem for dualists. Kroedel is aware of the *ad hoc* nature of his solution and makes two remarks in response (2015, p. 372). First, he argues that proposing a special modal relation between the mental and the physical is in line with the dualist intuition that the mental is special. Secondly, he suggests that it might be worthwhile to posit this special modal status without any further motivation in order to save mental causation. However, neither of these remarks fully dissolve the worry.

Kroedel is correct in stating that dualism is motivated by the intuition that the mental is special. However, the driving intuition is that consciousness is special because it is metaphysically distinct from the physical. It is true that this intuition is compatible with there being a special category of supernomological laws relating the mental to the physical, but it does not *motivate* positing such a special category. One could equally posit nomic laws, as we did, or even laws that are slightly weaker than nomic laws, but still stronger than social laws (cf. Lavazza and Robinson, 2014, p. 3–4). Claiming that the specialness of consciousness warrants positing a stronger modal link between the mental and the physical than a mere nomic link at least requires further arguments or a further development of what is meant by 'special'.

When it comes to safeguarding mental causation, one could indeed argue that some sacrifice is warranted. Even so, requiring this *ad hoc* posit in order to safeguard mental causation forces the dualist in an awkward dialectical position. Competing theories of the mind without *ad hoc* elements will have an immediate advantage. All else equal, we should prefer a model of mental causation without *ad hoc* elements.

It is worth noting however, that Kroedel's proposal does improve over other *ad hoc* solutions. In particular, supernomological dualism is more *par-*

simonious than a blank denial of *Non-Overdetermination*. As we discussed, one problem with blank denials of *Non-Overdetermination* is that they appear to posit more causal relations than are actually required to explain phenomena. Kroedel avoids this problem by couching his proposal in a counterfactual account of causation. According to Lewisian counterfactualism, facts about causation are metaphysically necessitated by the distribution of particular matters of fact across different possible worlds. Consequently, one does not *add* anything to one's ontology by claiming that one phenomenon causes another. One merely makes a claim about the distances between certain possible worlds. It just happens to be the case that the supernomological dualist's claim is an *ad hoc* one. This is not a decisive defect, but, all else equal, one would prefer a solution that does not require an *ad hoc* posit. Before turning to other solutions, we take a brief look at how supernomological dualism answers the exclusion worry and the common cause worry.

6.2 Against *Weak Exclusion*

According to Kroedel's proposal, the dualist can respond to the exclusion argument that we are currently considering by rejecting *Weak Exclusion*. We can characterize his ontology as follows:

Supernomological Dualism Mental phenomena are merely supernomologically necessitated by physical phenomena.

Here, one phenomenon is taken to *merely* supernomologically necessitate the other if and only if the former supernomologically necessitates the other but there is no tight relation between the two phenomena. It is easy to see how this ontology provides a response to the the exclusion argument.

Taken together with the counterfactual sufficiency criterion for causation and *Physical Completeness* this ontology contradicts the exclusion principle we are currently considering:

Weak Exclusion For any three phenomena *A*, *B* and *C*: if *A* occurs at *t* and is a sufficient cause for *B*'s occurrence at *t + x*, no phenomenon *C* occurring at *t* that is not tightly related to *A* and is not tightly related to any of *A*'s parts is a

cause of B , unless it is a case of genuine overdetermination.

We can see that it contradicts this principle by focusing again on the case of phys, pain and wince. According to *Physical Completeness* my wincing has a sufficient physical cause. According to counterfactuals (i) and (ii) my pain is a cause of my wincing. And finally, according to *Supernomological Dualism* there is no tight relation between my pain and the sufficient physical cause of my wincing. If these three claims are true, then *Weak Exclusion* is false, and the dualist has a response to the exclusion argument.

6.3 The common cause worry

Kroedel does not explicitly address the common cause worry. In this section, I argue that the common cause worry does affect his proposal. Supernomological dualism provides a difference between standard common cause scenarios and purported cases of dualist mental causation, but it does not provide a principled distinction. That is to say, the provided distinction does not explain the significance of the difference between causation and correlation due to a common cause. Overall, a solution that *does* provide a more illuminating distinction between standard common cause scenarios and purported cases of dualist mental causation would be preferable over the distinction provided by supernomological dualism.

Kroedel's proposal explicitly builds on Lewis's semantics of counterfactuals. Part of the motivation for Lewis's specific semantics of counterfactuals and his rules to measure the distance between possible worlds was to avoid that correlations between two phenomena that merely hold in virtue of their having a cause in common come out as causal. Kroedel proposes to adjust the rules for measuring distances between possible worlds by introducing a new category of laws: supernomological laws. Due to this reorganization some effects of a common cause come out as standing in a genuine causal relation in virtue of their having a cause in common. In particular, the correlations between mental phenomena and their target effects are considered as causal in virtue of their having a cause in common.

Consider the case of pain, phys and wince again. On Kroedel's view, phys

causes wince, because the following counterfactuals hold:³

(iii) $\text{PHYS } \Box \rightarrow \text{WINCE}$

(iv) $\neg \text{PHYS } \Box \rightarrow \neg \text{WINCE}$

In the closest world where phys is absent, I would not wince.

Within this framework, phys also causes pain. That is to say, the following counterfactuals are true as well.

(v) $\text{PHYS } \Box \rightarrow \text{PAIN}$

(vi) $\neg \text{PHYS } \Box \rightarrow \neg \text{PAIN}$

In the closest world where phys is absent, I would not be in pain.⁴

From (iii)–(vi), it follows that pain and wince are effects of a common cause. In order to respond to the common cause worry, we require some principled difference between the case of pain, phys and wince on the one hand, and standard cases of confounders on the other. There is only one obvious difference between the two. Standard cases of confounders, like smoking for tar-stained teeth and cancer, do not involve *supernomological* laws. It may seem unclear why this should matter to the causal status of the correlation between tar-stained teeth and cancer, but the supernomological dualist has her answer ready: it matters because it affects the truth value of the relevant counterfactuals. However, it is unclear why this difference should be of significance. It is often important to know which phenomena are causally related and which are merely correlated due to a common cause. For example, it is important to know that smoking causes cancer, but tar-stained teeth do not. If the only difference between standard common cause scenarios and mental causation scenarios is this difference between *supernomological* necessitation and nomic necessitation, it remains unexplained why this difference is of any importance to us. There is still room for an improved reply to the common cause worry for dualist mental causation.

³Kroedel explicitly recognizes this consequence of his view (2015, p. 363).

⁴Kroedel's framework comes with a further quirk: pain causes phys. Given that phys also causes pain, this means that phys and pain form a causal loop. On the assumption that causation is transitive, this means that they are both self-causing as well. This seems peculiar to say the least. However, I do not press this point here. Causal loops are a controversial topic (e.g. Ismael, 2003), as is causal transitivity (e.g. Björnsson, 2007; Hall, 2000; McDonnell, 2018).

6.4 Evaluating Kroedel's proposal

Supernomological dualism provides an improvement over a blank denial of *Non-Overdetermination* and List and Stoljar's proposal. It provides a consistent model of dualist mental causation. The model requires an *ad hoc* postulate, but is not unparsimonious. It also provides a distinction between mental causation and standard confounders, but this distinction is unilluminating. Perhaps this reply would do in the absence of more promising alternatives, but I will argue that we can do better.

Chapter 7

On Causal Sufficiency

As we have discussed in Chapter 5, a variety of the exclusion argument is often levied against non-reductionist physicalists as well. When responding, these non-reductionists frequently remark that exclusion arguments rely on the notion of causal sufficiency and claim that this notion is problematic. If these objections are indeed adequate, the dualist might also be saved from exclusion worries. After all, two central premises in the exclusion argument we are currently considering make reference to sufficient causes: *Weak Exclusion* and *Causal Closure*.

In this chapter, I will argue that these concerns about the role of causal sufficiency in exclusion arguments are relatively superficial. First, I provide a summary of the objections to causal sufficiency. Then I demonstrate how one can reformulate the exclusion argument by relying on *Physical Completeness* and adjusting the exclusion premise accordingly. The resulting exclusion argument avoids the kind of objections that plague standard formulations in terms of causal sufficiency. I conclude that exclusion arguments do not crucially rely on causal sufficiency and objections to this notion therefore pose only a superficial challenge to exclusionists. The upshot is that dualists are still in need of a convincing response to exclusion arguments.

7.1 Causal exclusion and sufficiency

Typical causal exclusion arguments contain two premises that make reference to ‘sufficient causes’: the closure premise and the exclusion premise. Several philosophers have remarked that the notion of a sufficient cause is problematic and maintain that the reliance on this notion undermines causal exclusion arguments. For example, Menzies claims that:

The fundamental error of this [causal exclusion] principle is that it mistakes causal sufficiency for causation (Menzies, 2013, p. 71).

Raatikainen makes a similar claim:

[...] both these assumptions [i.e. *Causal Closure* and *Weak Exclusion*] involve confusing causes with sufficient conditions. There are causes, which are difference-makers; and there are sufficient conditions, which are wholly different issues and not causes of any sort; there are no such things as sufficient causes. Hence, I do not think that these two assumptions are so much false (or true) as mongrels based on a conceptual confusion which fail to make clear sense (Raatikainen, 2010, p. 360).

Many others make mention of the problems they find in the notion of causal sufficiency and its role in the exclusion argument. Examples can be found in Crane and Aradottir (2013); Hitchcock (2012); Koons and Bealer (2010); List and Menzies (2009); Pernu (2013, 2016); Raatikainen (2013, 2018); Woodward (2008) and Zhong (2019).

Such objections to the reliance on causal sufficiency tend to be part of more integrated responses to exclusion arguments as they are raised against non-reductionist physicalism.¹ Note however, that if these criticisms are adequate, and causal exclusion arguments crucially rely on the problematic notion of causal sufficiency, the dualist is safe from causal exclusion arguments as well.² It is therefore worth taking a closer look at this criticism on causal sufficiency.

¹In fact, these responses are quite similar to the responses I rehearse in Chapters 5 and 10).

²This point seems to have gone by unremarked by all the of the authors listed above, except for Koons and Bealer (2010).

7.2 Against sufficient causes

A sufficient cause of an effect can intuitively be understood to be a cause of the effect that on its own suffices for the occurrence of that effect. Upon closer scrutiny, it appears that any phenomenon that is sufficient for the occurrence of an effect will have to be both enormous and maximally specific.³ This makes any phenomenon that is sufficient for an effect an unlikely candidate for being a cause of that effect.

It is easy to see why effects typically cannot have sufficient causes that we are familiar with. For example, there are many familiar causes of the window shattering after Yue threw a rock at it: Yue's throw, the rock's mass, the window's brittleness, etc. None of these familiar causes independently suffices for the window to shatter, as its shattering requires *all* of these familiar causes to occur.⁴ For a cause simultaneous with Yue's throw to be sufficient for the window shattering, its occurrence will have to fix the occurrence of *all* these familiar causes.

In fact, the occurrence of an effect requires more than just the occurrence of all its familiar causes. If these familiar causes are to give rise to the effect, there cannot be any *interference* with the causal process leading up to the effect. Consequently, the occurrence of a sufficient cause of an effect has to make it impossible that any such interference should occur. For example, the phenomenon simultaneous with Yue's throw that is sufficient for the window shattering has to make it impossible that there is interference by a meteor knocking the rock off course. That is a drastic requirement. Physics teaches us that an interference could travel at the speed of light. Consequently, such an interference could be on its way from an enormous distance. In order for the occurrence of a phenomenon to exclude that possibility, it has to span a sufficient amount of space, such that its occurrence fixes what *is* occurring at such enormous distances. Consequently, philosophers have concluded that only phenomena that span at least the entire cross-section of the backwards light cone of an effect can really be sufficient for the occurrence of that effect

³Alternatively, a sufficient cause could be understood to be sufficient for the effect given an agreed upon set of background conditions. This alternative reading will plausibly avoid the size problem, but not the specificity problem (see Chapter 14). As the aim of this section is to rehearse the non-reductionist's criticism on the notion, I focus on the reading that is most susceptible to it.

⁴Crane and Arnadottir (2013, Section 4.3) raise this point.

(e.g. Field, 2003; Loewer, 2007b).⁵ In layman's terms, sufficient causes must be enormous.

Once we start looking at what physics teaches us, it appears that sufficient causes not only need to be enormous, but also maximally specific. Any non-maximally specific phenomenon, like Yue's throw, can be physically realized such that it has a *thermodynamically abnormal future*. For example, the fundamental particles making up the rock might be arranged such that it suddenly emits a particle at an immense acceleration that is orthogonal to the rock's anticipated trajectory, causing the rock to make a sudden turn and therefore miss the window. Most probably, none of the actual rock throws, or enormous physical phenomena involving rock throws, will ever involve a mid-air turn due to the immensely accelerated ejection of a fundamental particle — that is why such a physical realization is thermodynamically *abnormal*. However, such trajectories are not physically impossible.⁶ In order for a phenomenon to be sufficient for an effect, its occurrence has to exclude the possibility that any of the phenomena in the relevant cross-section of the backwards light cone of that effect has a thermodynamically abnormal future that interferes with the occurrence of the effect. Not only must the rock not be poised to change its trajectory, none of the phenomena in the cross-section of the light cone can be such that they result in the ejection of a particle that will knock the rock off course. The safest way to ascertain that none of these phenomena are realized in such an abnormal way, is to fix how they actually *are* realized to a maximal degree of specificity. The underlying physics and theory are no doubt mind-boggling, but the takeaway is quite simple: in order for any phenomenon to be sufficient for any effect, it not only has to be enormous, but also maximally specific.⁷

Such enormous and maximally specific phenomena would make for peculiar causes to say the least. The causes we name in everyday conversations

⁵The cross-section of a backwards light cone of a phenomenon contains all space-time points at a certain time in the phenomenon's past from which a flash of light could have reached that phenomenon. If we assume that it took one second for the rock to reach the window in our example, the relevant cross-section is a sphere with a radius of approximately 300.000 km.

⁶See Albert (2015, Ch. 1) and Field (2003, p. 439) for accessible explanations.

⁷I assume here that disjunctions of phenomena are not phenomena. Otherwise, one could construct less specific sufficient phenomena from the disjunction of maximally specific phenomena that are individually sufficient for the effect.

as well as scientific explanations, like infections and hurricanes, are significantly more local and less specific. Critics of causal sufficiency argue that it is doubtful that such enormous maximally specific phenomena even qualify to be causes according to contemporary philosophical accounts of causation, and add that accounts of causation that would treat sufficient phenomena as causes are contentious or outdated.⁸

This critical attitude is not unmotivated. Many philosophers take causes to be *difference-makers* of their effects and the absence of difference-makers ought to correlate strongly with the absence of the target effect. For example, Yue *not* throwing the rock correlates strongly with the window *not* shattering at the time it did. In contrast, no phenomenon that is enormous and specific enough to be sufficient for the window shattering can be a difference-maker for its shattering. Due to its specificity and size, any change in the relevant cross-section of the backwards light cone of the effect entails that the sufficient phenomenon does not occur. However, there are many changes in the cross-section of the backwards light cone of familiar effects that are irrelevant to these effects. For example, the physical realization of Yue's shoes or the Belgian prime minister's hairdo could have varied substantially without affecting the shattering of the window. This means that for any phenomenon that is sufficient for the window shattering, there are many scenarios in which that phenomenon did not take place, but the window nonetheless shatters at the time it did in the actual world. It therefore appears that sufficient phenomena do not correlate strongly enough with their effects to be difference-makers. If causes are difference-makers, it is unclear why any sufficient phenomenon should be considered a cause — let alone why it should outcompete mental phenomena that often do serve as difference-makers, like intentions and desires.⁹

If the notion of a sufficient cause does not match with our everyday, scientific or philosophical understanding of causes, it is indeed suspicious that it features in the standard formulations of two of the central premises in the

⁸See Crane and Arnadottir (2013, p. 258), Koons and Bealer (2010, p. xix), List and Menzies (2009, p. 489), Woodward (2008, p. 251) and Raatikainen (2010) as quoted above, as well as Raatikainen (2013, p. 24) and (2018, p. 34). See also Hitchcock (2012, p. 53), who argues that accounts that treat sufficient phenomena as causes contradict other assumptions in exclusion arguments.

⁹This is the upshot of Hitchcock (2012, p. 53–55), List and Menzies (2009), Menzies (2013), and Zhong (2019).

exclusion argument: *Causal Closure* and *Weak Exclusion*. However, these worries can be answered in a straightforward fashion. One can formulate valid exclusion arguments that do not rely on the notion of causal sufficiency and do not encounter such worries.

7.3 Exclusion without causal sufficiency

Getting rid of causal sufficiency is not as hard as one might expect. The familiar notion of physical necessitation can do the work exclusionists cut out for causal sufficiency, without inheriting its problematic consequences. Recall that we defined physical necessitation as follows:

Physical Necessitation For any two phenomena A and B , A physically necessitates B if and only if all physically possible worlds that contain A also contain B .

A ‘physically possible world’ is any possible world in which the same fundamental laws of physics hold as in our world. If the exclusionist chooses to rely on physical necessitation rather than causal sufficiency, she can make it unambiguous what is meant by saying that A is ‘sufficient on its own’ for B in the context of the exclusion argument, *without* making any claims about A ’s eligibility for being a *cause* of B : all that is required for the occurrence of B is the occurrence of A .

Moreover, relying on physical necessitation allows us to replace *Causal Closure* with *Physical Completeness*:

Physical Completeness For any actual physical phenomenon P and any time t , there is a purely physical phenomenon that occurs at t and physically necessitates the occurrence of P .

Given that we have posited *Physical Completeness* at the outset of this project, we now have a stand-in for the closure premise that is, in the current context, non-negotiable. Note also that most of the authors listed above as objecting to *Causal Closure* do accept *Physical Completeness*.¹⁰

¹⁰Koons and Bealer might form an exception, but they do maintain that a failure

Now that the closure premise is reformulated in terms of physical necessitation, the exclusion premise must be adjusted accordingly. The obvious reformulation looks as follows:

Physical Exclusion For any three phenomena A , B and C : if A occurs at t and physically necessitates B 's occurrence at $t + x$, no phenomenon C occurring at t that is not tightly related to A and is not tightly related to any of A 's parts is a cause of B , unless it is a case of genuine overdetermination.

This principle imposes a necessary condition on any non-overdetermining cause occurring at t of an effect, namely being tightly related to the effect's physically sufficient phenomena occurring at t . We can summarize the intuition driving *Physical Exclusion* as follows: if the occurrence of a phenomenon is physically necessitated, any non-overdetermining cause of that phenomenon must be tightly related to physical history that necessitates its occurrence. Or, to put it in a slightly different way: the occurrence of any phenomenon that is not tightly related to a phenomenon A that physically necessitates B , must be *redundant* for the occurrence of B . After all, given the occurrence of A , the occurrence of B was already settled. So if this second separate phenomenon were to cause B , the occurrence of B would be overdetermined by A and the second phenomenon.¹¹ Or so the *Physical Exclusion* principle states.

Before turning to the resulting exclusion argument, it is worth setting aside some potential worries about *Physical Exclusion*. First, one might worry that this adjusted principle does not avoid all of the concerns about causal sufficiency. It is often remarked that any effect has a potentially infinite number of sufficient causes at any time preceding the effect. For example, if being male is causally sufficient for not getting pregnant, so is being male and taking birth control, being male and feeling dizzy, etc.¹² In

to distinguish between *Causal Closure* and *Physical Completeness* is the source of the widespread contention that dualists (and perhaps even non-reductionist physicalists) cannot allow for mental causation (2010, p. xix).

¹¹Note that the possible case of overdetermination would have to be a quite peculiar one, as the overdetermining cause would be metaphysically distinct from any physical phenomenon in the backwards light cone of the target effect. It would thus have to be a case of divine or supernatural intervention of the redundant kind. Even so, I leave the non-overdetermination *caveat* in, as such cases are not obviously metaphysically impossible.

¹²List and Menzies (2009, p. 489) and Menzies (2013, p. 72) both make this remark and

the context of the exclusion argument, one might worry that there are too many physically sufficient phenomena to choose from for any effect. Which of these is to exclude phenomena from being causes?

To some extent, these remarks apply to physically sufficient phenomena as well. There appears to be a potentially infinite number of such phenomena for any familiar effect. Start with the maximally specific physical phenomenon spanning the entire cross-section of the backwards light cone of the window shattering and add any redundant phenomenon: physical goings-on outside the backwards light cone, ectoplasmic goings-on inside that light cone, etc. The composite of a physically sufficient phenomenon for the window shattering and an irrelevant phenomenon will also be physically sufficient for the window shattering.¹³

However, *Physical Exclusion* contains clear instructions as to which physically sufficient phenomena to select: all of them. It states that a cause of an effect cannot be distinct from *any* of its strictly nomically sufficient phenomena. Consequently, physical goings-on outside the backwards light cone and ectoplasmic goings-on inside that light cone are excluded, because they are metaphysically distinct from at least one physically sufficient phenomenon: the maximally specific physical phenomenon spanning the relevant cross-section of the backwards light cone.

Second, one might worry that the change to *Physical Exclusion* creates a new problem. In particular, one might object that the resulting overdetermination it posits is not a *causal* overdetermination. After all, we granted the critics of *Causal Exclusion* that if A is physically sufficient for B , A is probably not a cause of B . Consequently, we cannot conclude that, if B has a cause that is not tightly related to A , B has one cause too many. Or, to borrow a phrase from Papineau (2002, Section 1.2), we cannot conclude that B is caused *twice over*. Standard exclusion arguments derive their thrust from the idea that such a systematic *causal* overdetermination is problematic. Our reformulated argument does not pose the same threat for dualists, as the resulting overdetermination is not causal. Or so the objection would go. I think there are two remarks to be made in response to this objection.

First, the resulting overdetermination is objectionable for the same rea-

ascribe it to Salmon (1971).

¹³After all, $\Box(p \rightarrow q)$, entails $\Box(p \& r \rightarrow q)$ for any p , q and r .

son that causal overdetermination is taken to be objectionable. The thesis that the effects of mental phenomena are causally overdetermined is taken to be problematic because it would mean that there is a *redundant* cause for each effect of a mental phenomenon. It is wildly implausible that these effects would systematically have causes that are redundant for their occurrence (cf. Papineau, 2002, Section 1.5). It matters little whether these causes are redundant because of the presence of another cause, or because of the presence of a physically sufficient physical phenomenon that does not stand in a tight relation to the purported cause. Either way, it would be a burden on the dualist if her theory states that all mental causes are redundant for the occurrence of their effects.

Second, if B is overdetermined by a physically sufficient physical phenomenon A and a phenomenon that is not tightly related to A , the causal overdetermination of B follows given a plausible further assumption. After all, it is plausible that A contains some phenomenon a that causes B against the background of all the rest of A — call that $A-$. By hypothesis, the purportedly redundant cause is not tightly related to either a or $A-$. If that is the case, then B is caused twice over: once by a and once by the purportedly redundant cause. After all, both that cause and a cause B against the background of $A-$: B would have occurred even if the metaphysically distinct cause were absent.

The move from *Causal Closure* and *Causal Exclusion* to *Physical Completeness* and *Physical Exclusion* does not introduce new problems, and it avoids the kinds of objections that are raised against the notion of causal sufficiency at work in *Causal Closure* and *Causal Exclusion*. So if *Physical Completeness* and *Physical Exclusion* can generate a valid exclusion argument, we can conclude that these objections can be answered in a straightforward fashion and therefore do not pose a serious threat to exclusion arguments. In the next section, I provide such a valid exclusion argument and make some notes about the reliance on *Physical Exclusion* and cases of realization.

7.4 Exclusion again

We can now formulate an exclusion argument without relying on causal sufficiency. The resulting exclusion argument runs as follows:

Nomic Dualism Mental phenomena are merely nomically necessitated by physical phenomena.

Physical Completeness For any actual physical phenomenon P and any time t , there is a purely physical phenomenon that occurs at t and physically necessitates the occurrence of P .

Physical Exclusion For any three phenomena A , B and C : if A occurs at t and physically necessitates B 's occurrence at $t + x$, no phenomenon C occurring at t that is not tightly related to A and is not tightly related to any of A 's parts is a cause of B , unless it is a case of genuine overdetermination.

Non-Overdetermination There is no systematic genuine overdetermination of physical effects with mental causes.

No Mental Causation Mental phenomena cannot systematically cause physical phenomena.

The argument is valid. *Nomic Dualism* states that mental phenomena are merely nomically necessitated by physical phenomena. *Physical Completeness* and *Physical Exclusion* together entail that causes of physical effects must stand in a tight relation to the co-occurrent physical phenomena that physically necessitate the target effect. Finally, *Non-Overdetermination* entails that purported effects of mental causes are not systematically overdetermined. It follows that mental phenomena systematically cause physical phenomena.

Consequently, objections to the notion of causal sufficiency cannot vindicate dualist mental causation. They pose only a superficial threat to exclusion arguments.

Before continuing, I would like to make a final note on replacing causal sufficiency with physical necessitation. Recall that, in Chapter 5, we followed

standard treatments of dualist ontologies in assuming that there are no *tight* relations between dualist mental phenomena and physical phenomena. Replacing causal sufficiency with physical sufficiency holds some promise of addressing List and Stoljar's objections to *Causal Exclusion* without having to rely on that claim. To see this, consider the following exclusion principle:

Strong Physical Exclusion For any three phenomena A , B and C : if A occurs at t and physically necessitates B 's occurrence at $t + x$, no phenomenon C occurring at t that is metaphysically distinct from A is a cause of B , unless it is a case of genuine overdetermination.

This principle relies on metaphysical distinctness instead of absence of tight relations, and on physical necessitation instead of causal sufficiency. *Prima facie*, it avoids the style of counterexamples that List and Stoljar used against *Causal Exclusion*, which read as follows:

Causal Exclusion For any three phenomena A , B and C : if A occurs at t and is a sufficient cause for B 's occurrence at $t + x$, no phenomenon C occurring at t that is metaphysically distinct from A and is metaphysically distinct from all of A 's parts is a cause of B , unless it is a case of genuine overdetermination.

Recall the counterexample we considered against this principle: a certain university is organized such that the committee delegated to make tenure decisions always consists of the most successful professors. Given this organizational structure, these professors making a negative decision simultaneously makes it the case that the university made a negative decision. If, in such a case, an applicant loses her job due to the university's decision being negative (UD), the most successful professors' decision being negative (PD) would *also* count as a cause, despite its being metaphysically distinct from UD . The example thus provides us with metaphysically distinct phenomena that co-cause an effect and thereby disproves *Causal Exclusion*.

However, it is less obvious that it also disproves *Strong Physical Exclusion*. Even though UD is clearly metaphysically distinct from PD , it is less clear that it is also metaphysically distinct from the physical phenomenon that oc-

curs simultaneously with *UD* and physically necessitates the job loss. After all, that physical phenomenon will be enormous. Consequently, it will metaphysically necessitate an immense number of relevant phenomena. For example, this physical phenomenon will metaphysically necessitate which words are contained in the university protocols in which exact order. Perhaps *PD* in conjunction with the contents of the university protocols *does* metaphysically necessitate *UD*. If this were the case, the university decision example does not contradict *Strong Physical Exclusion*. On the assumption that these observations generalize, one can use *Strong Physical Exclusion* to argue against dualists even if they are allowed to posit tight relations between mental phenomena and physical phenomena.

There are two potential complications. First, some philosophers argue that realized phenomena are metaphysically dependent on phenomena that precede the realized phenomena (see for example, Shoemaker (2007, p. 22) and Bennett (2017, Section 4.3)). Recall that we too have assumed that *UD*'s total realizer contained *ongoing* phenomena, like the organizational structure. This would mean that some realized phenomena are not metaphysically necessitated by co-occurrent enormous physical phenomena, but have diachronic phenomena in their total realizer as well. We can adjust *Strong Physical Exclusion* accordingly by focusing on the entire backwards light cone of the effect, rather than the physically sufficient time-slice of that light cone at *t*. Second, it might be the case that some phenomena are realized by phenomena that are situated outside the backwards light cone of one of their effects. Consequently, they would not be metaphysically necessitated by the physically sufficient phenomenon for the target effects. I am unable to conjure up such examples here, but am not confident that they are impossible either.¹⁴ Defending *Strong Physical Exclusion* against realization cases might require some extra work. Even so, the principle provides a promising back-up plan for the exclusionist in the event that nomic naturalist dualists finds a way to posit tight relations between mental and physical phenomena.

Overall, the above considerations do indicate that it is not always in-

¹⁴Clark and Wildman (2018) could be read as providing some examples, but they focus on *grounding* rather than metaphysical necessitation or realization, and explicitly rely on content externalism. As stated in Chapter 2 we are setting aside issues about content externalism here. I also think the difference between grounding and metaphysical necessitation will make a difference, but cannot argue for it here.

nocent to set aside worries about causal sufficiency. It appears that, all else equal, cases of metaphysically distinct but tightly related co-causes are less compelling against exclusion principles that do not ignore the tricky issues surrounding causal sufficiency. So even though objections to the role of causal sufficiency in exclusion arguments might pose only a superficial threat to these arguments, ignoring the issues surrounding causal sufficiency altogether might still lead us astray. I will set these issues aside here and focus on *Physical Exclusion*, which sidesteps both objections to causal sufficiency as well as objections in terms of realization or other tight relations. On the widely accepted assumption that there can be no tight relations between mental phenomena and physical phenomena in a nomic naturalist dualist ontology, *Physical Exclusion* provides a valid exclusion argument against nomic naturalist dualism.

Chapter 8

Lowe's Models of Mental Causation

The past three chapters discussed dualist strategies to safeguard mental causation by directly addressing one of the premises in the exclusion argument. The objections against the sufficiency premise and the exclusion premise posed only superficial problems for exclusion arguments against dualism. There are straightforward fixes that avoid the objections and still serve the purpose of such exclusion arguments. By contrast, Kroedel's proposal to adjust the dualism premise does not suffer from the same defect and provides the right result, but only does so at the cost of an *ad hoc* postulate. An alternative approach is to start with a model of dualist mental causation and work from there. If the model is credible it will provide its own motivation to reject one of the premises. My own approach will be more of this second kind. In this chapter, we take a look at three models proposed by Lowe.

Throughout his career, Lowe proposed several models of dualist mental causation that were to conform with a scientifically informed world view. I distinguish between three models and discuss them chronologically. First, Lowe proposed to consider dualist mental causes as 'enabling causes' (1992; 1996). Second, he proposed to consider dualist mental causes as explanatory causal mediators (1999). Third, he proposed to consider dualist mental causes as simultaneously occurring causal mediators (2000; 2008). Relative to our

current project, all of these proposals share the same deficiency. If they are to secure dualist mental causation, they either require an *ad hoc* rejection of *Non-Overdetermination*, and thereby provide no benefits over a bare overdetermination model, or they require a rejection of *Physical Completeness*, and thereby fall outside the scope of the current project. Before discussing these proposals however, it is worth elaborating on Lowe's ontology of the mind.

Even though Lowe explicitly brands his view as dualist, it is not entirely clear how well Lowe's brand of dualism fits with the dualism we have been discussing so far. There are at least two salient differences. First, Lowe considers persons to be the substance that is distinct from the physical (e.g. 1996, Ch. 2), whereas we are assuming that only phenomenally conscious phenomena are distinct from the physical. Second, Lowe is sometimes dismissive of necessitation claims about the mental and the physical (1996, p. 44–66), whereas we have assumed that the mental is nomically necessitated by the physical.¹ Based on these dissimilarities one might be inclined to set aside Lowe's proposals as irrelevant to our current endeavours.

Nevertheless, the problems facing our ontology of the mind and Lowe's are structurally the same: if one chooses to respect the completeness of the physical as espoused by naturalists, how does one allow for non-physical phenomena to cause physical phenomena? Consequently, Lowe's proposed answers to this challenge should interest us as well. If there is a way for non-physical persons that are not nomically necessitated by the physical to cause physical phenomena, there might be a very similar way for non-physical phenomena which are nomically necessitated by the physical to cause physical phenomena. Indeed, if Lowe's proposals for dualist mental causation were viable, our current investigation could reach a satisfactory result quite easily. Conversely, if — as I will argue — Lowe's proposals cannot secure dualist mental causation within the boundaries of our project, it will be instructive to see why.

¹In later work however, he proposed that dualist mental phenomena are causally necessitated by the causal phenomena underlying them (2000; 2008). At the very least, this conforms to the spirit of our own nomic necessitation characterization.

8.1 Enabling causes

Lowé dedicates the third chapter of his *Subjects of Experience* (1996) to the topic of mental causation. After ventilating some worries about the viability of physical completeness principles, he goes on to argue that dualist mental causation might be possible even if *Physical Completeness* is true. His argumentation builds on the distinction between ‘initiating’ causes and ‘enabling’ causes. Lowé accepts that *Physical Completeness* requires all *initiating* causes of our behaviour to be physical. He goes on to argue that this is compatible with there being *enabling* causes for our actions that are metaphysically distinct from these initiating causes. Moreover, he maintains that mental phenomena are well-equipped to play the role of such enabling causes for our actions. It would therefore follow that dualist mental causation is compatible with *Physical Completeness*.

Lowé proposes to accept that all our actions are physically necessitated by chains of purely physical phenomena. Hence agreeing with the opponent of dualist mental causation that “[...] every physical event has wholly physical antecedent causes which are necessary and sufficient for its occurrence” (1992, p. 69). He then remarks that these chains of physical phenomena are likely to be interwoven in vastly complex networks of causal interactions; in large part because the neural activation patterns that precede our actions exhibit an intractable complexity. If one observes any intentional action and tracks its physical causes backwards, one will be lost on an ever branching path of physical phenomena. The chaos of these physical phenomena converging on this particular action would be as mysterious as ever narrowing circles closing in on a specific point in a pond, from which a rock is suddenly propelled into one’s hand (cf. 1996, p. 68). That is to say, even though the action is entirely physically necessitated, it would still strike anyone observing only its physical causes as an unexplained oddity.

The causal role of the mental phenomenon is to explain why these individual physical phenomena converged onto this specific action. However, the mental phenomenon cannot play this role by initiating any of the physical phenomena leading up to the action, because all of these phenomena have physically sufficient physical causes. Instead, the mental phenomenon ‘enables’ these chains of physical phenomena to converge on the specific action

they cause. Thereby *explaining* why the action, which in the absence of the mental phenomena would appear to be caused by the physical phenomena as an unexplained oddity, takes place.

In order to clarify this enabling causal role and to contrast it with the initiating causal role of the physical phenomena, Lowe compares the causal role of the mental phenomena to the causal role of a spider's web and the spider's movements when moving across the web (1996, p. 82–83):

[...] the web is what we might call an 'enabling' or 'facilitating' cause, rather than an 'initiating' cause of the spider's movements — it enables and constrains these movements to take place in certain directions rather than others. Now, so too might states of consciousness both facilitate and impose constraints upon patterns of neural phenomena.

The initiating causes of the spider's movements lie within the spider's body, but these movements are also governed by the web on which it moves. Due to its enabling and constraining force, the web should be considered an 'enabling' cause of the spider's movements. Similarly, our actions are initiated by the intricate interaction of neural chains of phenomena, but these phenomena are coordinated towards converging upon this specific action, rather than upon any other action or perhaps upon no action at all, by the mental cause of the action (*ibid.* p. 84).

Further clarification of this proposal is provided by Gibb (2015a). She reads Lowe as claiming that the physical phenomena are sufficient *event-causes* of the subsequent action, but that the mental phenomenon causes the *fact* that these event causes converge on this specific action.² Lowe should therefore be interpreted as distinguishing between two kinds of causation which can come apart: event causation and fact causation. In the case of intentional action, the physical phenomena cause the action event and the mental phenomena cause the fact that these physical phenomena cause the action event.

However, neither this analogy nor the further distinction between fact causation and event causation adequately explains how dualist mental phe-

²And indeed, Lowe explicitly admits that his proposal commits him to the possibility of fact causation (1996, p. 67 fn. 19).

nomena can be enabling causes for our actions. If we set aside the possibility of genuine overdetermination, being an enabling cause of an effect plausibly requires being a difference-maker of this effect. The analogy between the spider web and dualist mental phenomena in a physically complete world is imperfect, in that the spider web *is* metaphysically necessitated by the totality of physical phenomena. This allows the spider web to make a difference to the spider's movement; if the web had not been there, its movements would have been different. If the spider web did not make a difference to the movements of the spider in this minimal sense, we would certainly not consider it an enabling cause of these movements, unless it were a case of overdetermination.³

Correspondingly, those who have distinguished between initiating causes and causal roles that are closer to Lowe's notion of enabling causes restrict their attention to phenomena that *are* difference-makers for the target effects. Consider for example, the following example by Dretske (1993, p. 122–123):⁴

A terrorist plants a bomb in the general's car. The bomb sits there for days until the general gets in his car and turns the key to start the engine. The bomb is detonated (triggered by turning the key in the ignition) and the general is killed. Who killed him? The terrorist, of course. How? By planting a bomb in his car. Although the general's own action (turning on the engine) was the triggering cause, the terrorist's own action is the structuring cause, and it will be his (the terrorist's) action, something he did a week ago that will certainly be singled out, in both legal and moral inquiries, as the cause of the explosion that resulted in death.

The initiating cause of the general's death converged on the general's death rather than on the general's being in time for work because of the enabling — or in Dretske's words 'structural' — cause of the general's death; i.e. the

³Lowe does warn the reader of the limits of this analogy, but the only disanalogy he mentions is that the spider's web is static whilst the mind is dynamic and it is unclear how this has any bearing on their respective causal roles.

⁴Lowe acknowledges that there is some affinity between his own enabling causes and Dretske's structural causes, but also remarks that there are differences. He does not elaborate on what these differences are.

bomb being wired up to the ignition by the terrorist. It should be clear that the bomb-planting is a difference-maker for the general's death in this example. Further, it should be clear that the bomb-planting only plays this enabling causal role by dint of being a difference-maker to the general's death.

Similarly, we should not consider mental phenomena enabling causes of the convergence of physical causal chains on our actions unless they were difference-makers of this convergence or it were a case of overdetermination. As we have seen in Chapter 5, it is at the very least unclear *how* dualist mental phenomena can be difference-makers if *Physical Completeness* is true. In the absence of a proper account of *how* mental phenomena can be difference-makers for physical effects (or, alternatively, of how they can be enabling causes without being difference-makers) Lowe's proposal does not address the real question for dualist mental causation. *Prima facie*, a convincing account of dualist mental phenomena as enabling causes still requires a denial of either *Physical Completeness* or *Non-Overdetermination*.

It matters little here whether the enabling cause is to cause an event or a fact.⁵ For example, it makes no difference whether one considers the terrorist's action as the cause of the general's death or as the cause of the fact that turning the ignition would detonate the bomb. In either case, the purported enabling cause has to be a difference-maker for the target effect, unless it is a case of genuine overdetermination. More generally, even though the debates on how and whether facts can be causes are intricate and ongoing, fact causation will have to exhibit relevant similarities with familiar causal relations if it is to be counted as causation at all. Consequently, the plausibility of most philosophical research on (mental) causation is typically taken to be independent of the outcome of debates on the nature of its relata.⁶ It is thus unclear how restricting the causal role of mental phenomena to causing facts rather than phenomena will be of help for the dualist.

Lowe claimed that one can respect the physical completeness of the physical domain whilst securing dualist mental causation by denying that mental

⁵See also, Engelhardt (2017, Section 2) for a similar objection. Engelhardt goes on to defend a variation of Lowe's position. In particular he defends the position that dualist mental substances can have causally effective properties in as far that these properties are reducible to physical properties.

⁶As is evidenced by the typical disclaimer about causal relata in philosophical papers on causation.

phenomena can be *initiating* causes of our actions and granting them the role of *enabling* causes of our actions instead. In the absence of a proper account of how dualist mental phenomena can do so without being difference-makers for our actions, this claim is not sufficiently supported. Despite the spider web analogy and the proposal to restrict the causal role of mental phenomena to causing facts, it is no more clear how the mental can play this enabling causal role than how it could play an initiating causal role. Without a proper account of how dualist mental phenomena can be enabling causes in the face of *Physical Completeness* and *Non-Overdetermination*, the original problem for the dualist remains: there appears to be no room for mental causation.

8.2 Diachronic mediation

In later work, Lowe proposed to reconcile the completeness of the physical with dualist mental causation by claiming that mental phenomena are causal mediators that have sufficient physical causes of their own. There are two versions of this proposal. According to the first, mental phenomena diachronically mediate between physical phenomena. Lowe maintains that, by doing so, mental phenomena can explain why our actions are non-coincidental rather than coincidental. According to the second, they mediate synchronically between physical phenomena and they are contributory causes on a par with the physical causes. In the next section, we address his proposal in terms of synchronic mediation. But first, we address his proposal in terms of diachronic mediation.

Lowe (1999) extends on his intuition that mental phenomena have a specific causal role in bringing about our actions that allows them to explain those actions in ways that the physical causal chains necessitating those actions cannot explain them. In particular, he maintains that mental phenomena could explain why an action is non-coincidental rather than coincidental. On Lowe's mental model, they could explain this by virtue of causally mediating between different physical chains of phenomena that lead up to the action.

Even if an action is fully necessitated by the physical chains of phenomena converging on this action, we could still consider this action's occurrence as

coincidental, according to Lowe, if the physical chains of phenomena are entirely separate before converging on this action. Two physical chains of phenomena are to be considered separate if they have no common causes in their history. An event is thus taken to be *coincidental* if the causal chains leading up to it have no causes in common. Conversely, an event is taken to be *non-coincidental* if the causal chains leading up to it have a cause in common.

Lowe admits that this conception of coincidence and non-coincidence will require further tinkering for general use, but maintains that it suffices to sketch a model of mental causation according to which mental phenomena explain why our actions are not coincidental. He illustrates this model by considering the scenarios depicted in Figure 8.1 In both worlds the action

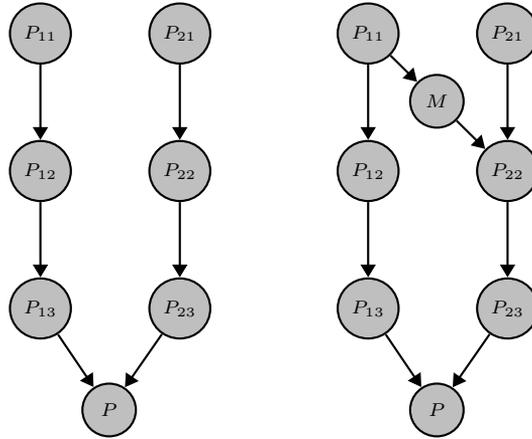


Figure 8.1: World 1 on the left, World 2 on the right

' P ' is necessitated by the physical phenomena P_{13} and P_{23} . In World 1, but not in World 2, the occurrence of P is coincidental, because in World 1, but not in World 2, the two causal chains leading up to P are separate. The presence of mental cause M , which causally mediates between P_{11} and P_{22} in World 2, thereby makes for a significant causal difference between World 1 and World 2. If mental phenomena actually operate causally according to the above model, dualist mental causation is compatible with *Physical Completeness* according to Lowe. He acknowledges that it is an empirical question whether mental causation actually works this way, but maintains

that we ought not to conclude that dualist mental causation is impossible given *Physical Completeness* because it *could* operate according to this model.

Whatever the empirical adequacy of the model turns out to be, it cannot dissolve the dualist's mental causation worries on its own. Consider the occurrence of P_{22} in World 2. Either M is causally redundant for P_{22} 's occurrence, or P_{22} is not physically necessitated by the physical phenomena occurring at the time of M 's occurrence. In other words, the proposed model of mental causation either proposes that mental causation is always a case of overdetermination, or it proposes that *Physical Completeness* is false. By contrasting World 2 with World 1, Lowe appears to opt for overdetermination. However, the contrast might be deceptive, since Lowe mentions that the causal laws are different in these two worlds (1999, p. 199). Perhaps he means by this that P_{21} is causally sufficient for P_{22} in World 1 but not in World 2. Either way, the proposed model provides no significant advancements for dualist mental causation.

One could stop here to argue that the relevant difference was supposed to be an explanatory one and that, whatever the difficulties with closure and overdetermination turn out to be, it *is* the case that the occurrence of P is a coincidence in World 1 but not in World 2 according to Lowe's conception of coincidence. Hence, there *is* a relevant difference between World 1 and World 2, and therefore M *does* play an important causal role.⁷

Such an argument fails to resuscitate Lowe's proposal. It is unclear how mental phenomena could explain our actions without being difference-makers of those actions. Or at least, that is the case if we set aside known complications such as overdetermination. The notion of explanation is itself contentious, but most accounts of explanation require that explanatory phenomena are difference-makers of the phenomena to be explained.⁸ Of course there might be exceptions involving overdetermination, but this still puts us back at square one: either dualist mental phenomena are difference-makers and *Physical Completeness* is false, or all instances of dualist mental causation are instances of overdetermination.

⁷Perhaps a similar argument could have been made about Lowe's previous proposal as well. Enabling causes were supposed to explain the occurrence of their effects. *Mutatis mutandis*, the same two remarks apply to the case of dualist enabling causes as explainers.

⁸e.g. Woodward (2003, p. 157–161 and 175–181).

Further, it is worth noting that Lowe himself is dismissive, if not downright contemptuous, of mental causation models which threaten to restrict the causal role of the mental to an explanatory one. Such models amount to admitting that the mental has no real causal role to play according to Lowe, as he states: "This is, effectively, to take a *non-realist* view of the causal and ontological states of the mental" (1996, p. 75).⁹

Overall, this proposal does not provide a fully developed model of dualist mental causation. The original problem for dualist mental causation reoccurs on this model as well. If mental phenomena are diachronic explanatory causal mediators between the physical causal chains which necessitate our actions, either the physical domain is not complete or our actions are overdetermined.

8.3 Synchronic mediation

Lowe provides a third model of dualist mental causation in terms of *synchronic mediation*. When providing this model, Lowe explicitly denies *Physical Completeness* and replaces it with a different completeness principle. This denial elicits two notes. First, one might take this late explicit denial to cast some light on the previous models as well. As we have seen, these appear to require a denial *Physical Completeness* or *Non-Overdetermination* in the absence of further explanation. Note however, that Lowe was explicit about granting his opponents the presence of sufficient physical causes for behavioural effects when putting forward the previous two models. True, he did not specify whether or not he took these causes to be sufficient in the way typically understood by proponents of exclusion arguments,¹⁰ but it would have been odd to say the least if he intended these causes to be sufficient in a way that obviously differs from what his opponents have in mind.¹¹

Second, given its denial of *Physical Completeness*, the third model lies outside the scope outside of the current investigation. After all, we have set aside varieties of dualism that deny *Physical Completeness* at the outset of

⁹He directs this criticism at Heil (1992); Mele (1992); Yablo (1992) and Dennett (1991b).

¹⁰Of course, as we have discussed in Chapter 7, this notion might be inherently problematic as well, but the further development of these two models does not suggest that this was the weakness in exclusionist thinking he was seeking to exploit.

¹¹Moreover, the two previous models are not compatible with Lowe's proposed brand of completeness either.

this problem. Even so, I will briefly discuss this third model here, if only to set it aside. My reason for doing so is that the third model can be interpreted as an improved version of the second model. It further explains how a model that relies on causal mediation can be made to work whilst at least paying lip service to principles like *Causal Closure* and *Causal Completeness*. Consequently, the synchronic mediation model, while falling outside the scope of our current investigation, provides our discussion of Lowe’s models of dualist mental causation with a natural endpoint.

Lowe’s third model is a variation on his causal mediation proposal. This variation does not rely on the contrast between coincidental and non-coincidental causes. Instead, he proposes the following model of mental causation represented in Figure 8.2. As in his previous model, the mental phenomenon

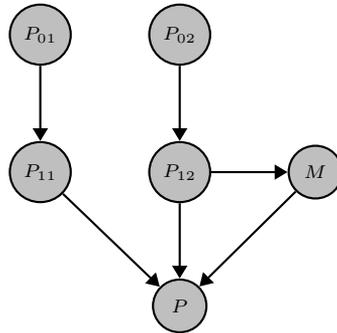


Figure 8.2: Lowe’s synchronic mediation case

M is a causal mediator between two physical phenomena. However, rather than causally mediating between two otherwise disconnected causal chains, M mediates between its own underlying physical phenomenon (P_{12}) and the purported physical effect (P). According to the above sketch, P would thus have three distinct contributory causes: P_{11} , P_{12} and M .

When presenting this model, Lowe is less ambiguous about the causal significance of M for P . He states that “[...] if M had not occurred, then the conjunction of P_{11} and P_{12} , even if it had occurred, would not have sufficed to cause P ” (2008, p. 72). He thereby confirms that M is not an overdetermining cause in the sense that we are operating under, because its role is by no means redundant.

The cost of positing that M is a difference-maker for P is a breach of *Physical Completeness*. Lowe accepts this and goes on to argue that there is a significant sense in which this proposal respects the closure principles at work in mental causation debates. The physical causes at the time of M 's occurrence t_1 are sufficient for the occurrence of P at t_2 in a non-trivial sense. P_{11} and P_{12} will, together with the fundamental laws of nature, necessitate the occurrence of P . It just happens to be the case that they will do so by necessitating the simultaneous occurrence of M , which in turn will — together with P_{11} and P_{12} — cause P . Even if the necessitation from the physical conditions to the physical effect requires the psychophysical laws as well as the physical laws, it is still the case that P has sufficient physical conditions at t in the sense that these conditions *nomically* necessitate P .

Nevertheless, the synchronic mediation model clearly violates *Physical Completeness*, as well as the spirit of the causal closure principles that are standardly propounded in mental causation debates. These principles posit that for every physical effect and any time t , there are physical conditions obtaining at t that in conjunction with the laws of physics *alone* necessitate the given effect. This can clearly be seen in the argumentations provided in favour of *Physical Completeness*:

All physical effects have complete physical causes ('complete' in the sense that those causes on their own suffice by physical law to fix the chances of those effects). (Papineau, 1993, p. 22)

If a physicist encounters a physical event for which there is no ready physical explanation, or physical cause, she would consider that as indicating a need for further research; perhaps there are as-yet undiscovered physical forces. At no point would she consider the possibility that some non-physical force outside the space-time world was the cause of this unexplained physical occurrence. (Kim, 2010, p. 113)

Lowe anticipates this criticism by arguing that the causal contributions of the mental would be invisible to scientists who restrict their attention to the physical causes of our actions. Given that one of the physical causes of any action will cause the required mental cause in all cases that can be

empirically studied, the scientist who focuses solely on the physical causes will have no reason to believe that one of the causes is missing from her causal explanation of any action. That scientists never have to look outside the physical domain for causes should thus not move us to endorse *Physical Completeness* rather than Lowe's proposed model of mental causation.

Be that as it may, Lowe's synchronic mediation proposal only provides us with dualist mental causation by denying *Physical Completeness*. As I have flagged in Part I of this dissertation, we will not be concerned with the debates on *Physical Completeness*. Lowe's third model of dualist mental causation therefore lies beyond the scope of this dissertation.¹²

8.4 Evaluating Lowe's models

The central defect of Lowe's models is that they do not address a central question for models of dualist mental causation. *How* can dualist mental phenomena (non-overdeterministically) cause physical effects, if these physical effects are already physically necessitated by purely physical phenomena? Given that these effects are *necessitated* it does not seem like mental phenomena can make any difference to their occurrence. This question is not answered by merely reconsidering mental phenomena as *enabling* or *mediating* causes. After all, mediating causes and enabling causes typically make a difference to the occurrence of their effects. Without a convincing answer to this central question, Lowe's proposals provide little advancement for the nomic naturalist dualist.

Lowe's proposals share this defect with the objections raised against *Causal Exclusion* and *Causal Closure* in Chapters 5 and 7. These objections do not explain how dualist mental phenomena can be causes and can therefore be countered with relative ease. In order to satisfactorily address the dualist's mental causation problem, we require an explanation of how non-physical phenomena can cause physical phenomena in worlds where *Physical Completeness* holds. Kroedel provides such an explanation: if dualists posit that psychophysical laws are modally stronger than regular nomic laws but

¹²See Robb (forthcoming) for a sympathetic yet critical discussion of Lowe's third model of dualist mental causation.

weaker than metaphysical laws, then the relevant counterfactuals support the conclusion that dualist mental phenomena cause our behaviour (Chapter 6). However, Kroedel's proposal requires an *ad hoc* posit about the relative distances between possible worlds. The dualist would be better off if she could allow for mental causation without relying on an *ad hoc* posit.

This concludes my discussion of the contemporary solutions to the mental causation problem as it arises for the nomic naturalist dualist. We now turn to Part III of the dissertation, in which I present and defend my preferred model of dualist mental causation. This model starts from the interventionist account of causation, so that is where Part III starts as well.

Part III

Dualist Mental Causation

Chapter 9

Minimal Interventionism

As mentioned, my proposed model for dualist mental causation will build upon an *interventionist* account of causation. In this chapter I present and discuss a minimal version of this account based on its two central definitions: the definition of causation (M) and the definition of an intervention variable (IV). This *minimal interventionism* will serve as a starting point for the final part of this dissertation, in which I discuss some problems from philosophy of causation and develop a model of dualist mental causation. My aim is to argue that solutions to these problems about causation motivate an interventionist model of causation that allows for dualist mental causation in worlds where *Physical Completeness* is true.

The overall strategy for arguing my point looks as follows. First, I present minimal interventionism (this chapter). I then demonstrate how this model allows for non-reductionist physicalist mental causation, but *not* for dualist mental causation in Chapter 10. In particular, so-called ‘holding fixed’-requirements on physical nomic bases stand in the way of dualist mental causation according to the minimal interventionist model. In Chapter 11, I argue that minimal interventionism allows for too much causation. In cases where higher-level phenomena are metaphysically necessitated by their underlying lower-level phenomena, minimal interventionism entails that all effects of the lower-level phenomena are effects of the higher-level phenomena as well. Minimal interventionism therefore results in *spurious* higher-level causation. I propose to solve this problem by adding a *robustness* requirement on causa-

tion to the interventionist account of causation. In Chapter 12, I argue that, somewhat surprisingly, adding the robustness requirement opens the door to an interventionist model of causation that *does* allow for dualist mental causation. It does so because it renders the ‘holding fixed’-requirements on physical nomic bases, which stood in the way of dualist mental causation, redundant. I formulate such an account of causation, which I dub ‘insensitive interventionism’, and show how it allows for dualist mental causation and how it provides principled answers to both the exclusion worry and the common cause worry. In the remaining three chapters, I argue that objections to insensitive interventionism can be countered by relying on recent developments in philosophy of causation.

Before doing so, one disclaimer is required. The purpose of the minimal interventionism presented in this chapter is to illuminate the obstacles for an interventionist model of dualist mental causation. However, this account is *not* supposed to represent a fully developed account of causation that is actively defended in the literature. (M) and (IV) are central definitions in Woodward’s (2003) interventionist account of causation, but they are only meant provide *minimal* adequacy conditions for causal claims and Woodward acknowledges that further factors come into play when assessing such claims (cf. Woodward, 2008, 2018). I too will be using these definitions as a starting point and, as we shall see, the adjustments I will propose align with suggestions made by Woodward and other interventionists.

9.1 Interventionist causation

Interventionist accounts of causation analyze causation in terms of what would happen under certain ‘interventions’ on phenomena. The first detailed philosophical proposal of this kind is offered in Woodward’s (2003) *Making Things Happen*, but the essence of such accounts goes back to improvements in causal modelling (e.g. Pearl, 2000). The central idea is simple and intuitive: what it means for phenomenon *A* to be a cause of phenomenon *B* is for it to be possible to affect *B*’s occurrence or the probability of *B*’s occurrence by intervening on *A*’s occurrence. For example, what it means for smoking behaviour to be a cause of lung cancer is for it to be possible to affect (the

probability of) lung cancer occurring by intervening on the occurrence of smoking behaviour. Aside from its intuitive appeal, interventionism gained further credibility through its broad applicability in scientific disciplines. The interventionist framework is used to elucidate the role of causation and causal claims in psychology (Woodward, 2007), psychiatry (Kendler and Campbell, 2009), biology (Woodward, 2010), neurobiology (Woodward, 2017), decision theory (Stern, 2017, 2019), thermodynamics (Zwier, 2017), organic chemistry (Statham, 2017), information theory (Andersen, 2017) and statistical mechanics (Leeds, 2010).

The increasing popularity of interventionism holds some promise for non-reductionism about the mental. Several authors argue that non-reductionist physicalist mental phenomena can be causes according to the interventionist definition of causation.¹ However, these interventionist accounts of non-reductionist mental causation are not without their opponents.² In the next chapter we will assess whether or not there is a plausible interventionist model of non-reductionist physicalist mental causation and whether or not this model could accommodate dualist mental causation as well. In this chapter, we take a closer look at the interventionist definitions of causation and intervention.

The central definition of causation in the interventionist literature reads as follows (Woodward, 2003, p. 59):³

- (M) A necessary and sufficient condition for X to be a type-level direct cause of Y with respect to a variable set V is that there be a possible intervention on X that will change Y or the probability distribution of Y when one holds fixed at some value all other variables Z_i in V .

Informally, (M) states that what it means for one variable X to cause another variable Y relative to a variable set V is for there to be a change in the value

¹Some notable examples are Campbell (2008, 2010); Eronen (2012, 2017); Eronen and Brooks (2014); List and Menzies (2009); Menzies (2013); Polger et al. (2018); Raatikainen (2010, 2013, 2018); Shapiro (2010); Shapiro and Sober (2007) and Woodward (2008, 2015).

²Most notably, Baumgartner (2009, 2010, 2013, 2018).

³Woodward's entire (M) definition is somewhat lengthier as it also aims to cover cases of indirect causation, which we will set aside here. This abbreviated version will do for our discussion and is also the starting point of discussions on interventionism in the mental causation literature (e.g. Baumgartner, 2009, 2010; Woodward, 2008, 2015).

of Y in some scenario where one ‘intervenes on’ the value of X without changing the value of other variables in V . We will turn to the definition of ‘intervention’ soon, but for now it suffices to understand that an intervention on the value of a variable is a *manipulation* of the value of that variable that meets some further requirements. For example, changing the number of smokers in a population is a manipulation of the number of smokers in that population. Whether or not this *manipulation* is an *intervention* will then depend on some of its further features.⁴

We can elucidate (M) with an example. According to this definition, the number of smokers in a population is a cause of the occurrence of lung cancer in that population if and only if there are changes in the occurrence of lung cancer in scenarios where one causes a change in the number of smokers and holds fixed the values of any other variable in the variable set we are currently considering. Such other variables could be: air pollution, genetic predisposition, eating habits, etc. All of these other variables need to be ‘held fixed’ at a certain value when intervening on X in order for subsequent changes in Y to be sufficient for it being the case that X causes Y , as is represented in Figure 9.1. This is a reasonable requirement, as we would not

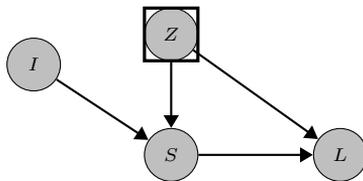


Figure 9.1: The intervention (I) changes the value of smoking behaviour (S) whilst other variables in the system (Z) are held fixed at a certain value

want to conclude that X causes Y , just because bringing about a change in X whilst a change in another variable takes place is followed by a change in Y . To see this, consider another example. We raise the price of apples without controlling for changes in the smoking behaviour of the population

⁴From here on, I will use ‘manipulation’ for a change in value that either does not meet the criteria for being an intervention, or of which it is uncertain that it meets these criteria. The term ‘intervention’ will be reserved for manipulations that meet the criteria listed below.

or the air quality in the area. If we subsequently observe an increase in lung cancer, we should not conclude that the price of apples causes lung cancer. For all we know, variations in the smoking behaviour or air quality actually caused the increase in lung cancer. The ‘holding fixed’-requirement ensures that the observed changes in Y cannot be due to changes in variables other than X . As we shall see, this requirement will also play an important role in our discussions on mental causation and the formal characterization of an intervention itself. Before turning to these issues, some further remarks on (M) are in order.

First, a note on the causal relata. As I mentioned in the introduction, I will pay no special attention to the assumptions about causal relata made in definitions of causation. I chose to use ‘phenomena’ as the term to denote causal relata, but emphasized that this term functions merely as a placeholder. Interventionists make a similar move. They formulate their definitions in terms of ‘variables’ and sometimes ‘values’ and treat these terms more or less like placeholders. For example, in the introduction to his book on causation Woodward makes the following disclaimer (Woodward, 2003, p. 22):⁵

[Interventionism] is most naturally formulated in terms of variables — quantities or magnitudes that can take more than one value. Causal relationships, of course, have to do with patterns of dependence that hold in the world, rather than with relationships between numbers or other abstracta, but in the interest of avoiding cumbersome circumlocutions, I will often speak of causal relationships as obtaining between variables or their values, trusting that it is obvious enough how to sort out what is meant.

In the interest of continuity with standard interventionist definitions, I will also adopt the convention of sometimes talking as if variables are causal relata. As is the case with my talk of ‘phenomena’, I assume that translations in terms of the reader’s preferred causal relata are straightforward.

Second, notice that causation is defined in terms of an ‘intervention’. This

⁵At this stage, Woodward still uses the term ‘manipulability theory’ rather than ‘interventionism’, which became the more popular term later on. I adjusted the quote to fit with the rest of this text.

not only sounds like a causally laden term, but is explicitly acknowledged by Woodward as denoting a causal notion (e.g. 2003, p. 98). The interventionist definition of causation is therefore not a reductive one. It does not attempt to analyze causation in non-causal terms, like transfers of energy, physical necessitation or distances between possible worlds. Instead, it aims to analyze causal facts, like whether or not X is a cause of Y , in terms of other causal facts, such as facts about what happens if one causally affects X whilst stopping other variables from causing changes in Y . If changes brought about by an intervention on X still correlate with changes in the value of Y under such circumstances, then X is a cause of Y . Consequently, the interventionist does not analyze the causal relation between X and Y in terms of a causal relation *between* X and Y , but in terms of the causal relations *surrounding* X and Y (cf. Woodward, 2003, p. 104–107). The interventionist concludes from these considerations that the circularity in (M) is not problematic. After all, (M) still elucidates the notion of causation to such an extent that we can use it to evaluate causal claims without assuming or positing anything about the truth of the causal claim involved.

Third, the definition concerns *type-level* causes. The account is thus developed to evaluate type causal claims such as ‘smoking causes lung cancer’ and ‘pain causes wincing’, rather than token causal claims such as ‘Wendy’s smoking caused her lung cancer’ or ‘my pain caused my wincing’. However, an account of token-causation naturally follows from this definition of type-causation (cf. Woodward, 2003, section 2.7). Following (M), we can say that a type-causal claim, like ‘smoking causes lung cancer’, is true if there is an intervention on the value of smoking behaviour that is followed in a change of value of lung cancer risk with all other variables in the variable set held fixed at *some* value. A token-causal claim, like ‘Wendy’s smoking caused her lung cancer’ is true if there is a possible intervention on the *actual value* of the variable representing her smoking behaviour that will result in a value change in the variable representing the risk of her incurring lung cancer, whilst all other variables in the variable set are held fixed at their *actual value*.

It follows that token causation and type causation are intimately related. The truth of a type causal claim, requires (and thus entails) the possibility of a corresponding token causal claim. In order for smoking to cause cancer,

there must be some metaphysically possible scenario in which the smoking behaviour variable taking a certain value causes the lung cancer risk variable to take a certain value. Entailments in the other direction are harder to pin down. Type-level causal claims are typically taken to be *generic* claims, meaning something like ‘normal tokens of type X typically cause tokens of type Y ’. We will set aside the intricacies surrounding generic claims and the meaning of ‘typically’ and ‘normal’ here.⁶ As before, our examples will focus on token causation, and we will assume that the relevant kind of generalization is available. Even if this should turn out to be false and we establish only that tokens of dualist mental phenomena can cause tokens of physical phenomena, we will have attained a significant result.

Fourth, this account of token causation will still face the usual challenge of accounting for cases of genuine overdetermination, (late) pre-emption and back-up causes (cf. Björnsson, 2007; Schaffer, 2001b; Yablo, 2002). However such cases all concern causation in circumstances that are in a significant sense abnormal, and it is unlikely that cases of (purported) mental causation are abnormal in that sense (cf. Section 3.1). Consequently, we set these kinds of counterexamples aside here and in the remainder of the dissertation.⁷

Finally, (M) concerns *direct* causes, rather than *indirect* causes. Indirect causes of an effect bring about the affect *via* another cause, whereas direct causes do not. In its current form, (M) can only model *direct* causes, because it requires that *all* variables in V , except for X and Y , are held fixed when one intervenes on X . If V were to contain causal intermediaries between X and Y , this means that even those causal intermediaries should be held fixed. By holding fixed these intermediaries, one will *de facto* ensure that there is no change in Y even if X is an indirect cause of Y . To see this, consider the following case. We consider a variable set containing the following variables: buying cigarettes (B), smoking behaviour (S) and lung cancer risk (L). Plausibly B causes S and S causes L , as is represented in Figure 9.2. However, if we were to intervene on the value of B whilst holding fixed the value of S , there is unlikely to be a change in the value of L . After all, buying cigarettes

⁶See Lewis (1973a, p. 558) and Carroll (1991) for some challenges for spelling out the right kind of typicality in accounts of type-causation

⁷Moreover, the treatments of such abnormal cases proposed in the above cited sources are compatible with the model of causation I will defend here.

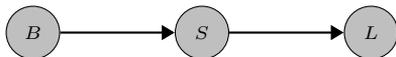


Figure 9.2: Buying cigarettes (B) causes smoking behaviour (S) which causes lung cancer (L)

only leads to higher cancer risk by causing certain smoking behaviour. Consequently, changes to a population's cigarette buying behaviour will not result in changes in lung cancer risk in that population if it does not affect the smoking behaviour. A representation of the intervention that complies with (M) is provided in Figure 9.3. This means that (M) does not capture cases

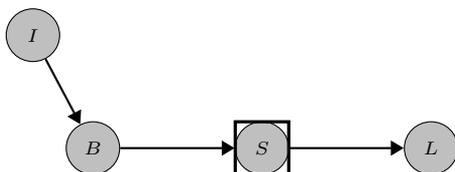


Figure 9.3: The intervention (I) changes the value of buying cigarettes (B) whilst other variables in the system, namely smoking behaviour (S) are held fixed at a certain value. No change in target variable L is expected.

of *indirect* causation.

We will not be concerned with the difference between direct and indirect causation in what follows. There are two reasons for setting this difference aside. First, one can easily adjust (M) to handle cases of indirect causation as well. One can do so by excepting variables that lie on the causal pathway from the investigated variable to the target variable from the holding fixed requirement in (M) (Woodward, 2003, p. 59). As we shall see, the definition of an intervention does something very similar. Second, note that being a direct cause is defined relative to a variable set. It will thus depend on the selection of variables whether a variable X is a direct cause of a variable Y . A direct cause relative to one variable set can fail to be a direct cause in a set that includes an extra variable that lies on the causal pathway from this variable to the effect variable. For example, relative to a variable set that only contains the variables for cigarette buying, the price of apples and lung

cancer, it is likely that buying cigarettes *is* a direct cause of lung cancer. It will simplify matters somewhat if we focus on the relevant direct causal claims in what follows by omitting the purported causal intermediaries from our variable sets. We can safely assume that the corresponding claims of indirect causation can be captured given some simple adjustments in (M).

With these remarks taken care of, we can turn to the definition of ‘intervention’ and ‘intervention variable’.

9.2 Intervention variables

The change in variable I counts as an intervention on the variable X relative to a target effect variable Y if and only if this change in I causes the value taken by X and I is an *intervention variable* for X with respect to Y . The central definition of an ‘intervention variable’ reads as follows (Woodward, 2003, p. 98):

I is an intervention variable for X with respect to Y if and only if I meets the following conditions:

(IV)

I1 I causes X .

I2 I acts as a switch for all other variables that cause X . That is, certain values of I are such that when I attains those values, X ceases to depend on the values of other variables that cause X and instead depends only on the value taken by I .

I3 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 .

- I4** I is (statistically) independent of any variable Z that causes Y and that is on a directed path that does not go through X .

I1 states that an intervention should affect the investigated variable (X) by causing it. It thereby further emphasizes that interventionism is not a reductive definition of causation.

The remaining three conditions are designed to avoid scenarios where manipulations of X do correlate with changes in the target variable Y , but the change in Y is caused by a change in a distinct variable Z , rather than the change in X . In such cases, we will say that the correlation between changes in X and changes in Y is *confounded* by Z . For example, there is (plausibly) a positive correlation between tar-stained teeth and lung cancer risk in the global population, but this correlation is *confounded* by the smoking behaviour in the global population. This means that the smoking behaviour is doing the actual causing, whereas the tar-stained teeth are just a symptom of this actual causing being done.⁸ The presence of tar-stained teeth in cases with high lung cancer risk is a mere symptom of the presence of a plausible cause of high lung cancer risk, namely relatively frequent smoking behaviour. Similarly, the absence of tar-stained teeth in cases with low lung cancer risk is a mere symptom of the absence of that plausible cause. Consequently, the correlation between tar-stained teeth and lung cancer risk is a mere symptom of the causal relation between a confounder and the target effect.

In such cases of confounded correlations, there could be manipulations of the symptom that correlate with changes in the target effect. This is because one can manipulate the symptom by manipulating the confounder. For example, one can manipulate the number of tar-stained teeth in a population by manipulating the number of smokers. Such a manipulation will plausibly correlate with changes in the lung cancer risk in that population. If interventions on X relative to Y are to be indicative of there being a causal relation between X and Y , then interventions should be defined such that the correlation between changes in X and changes in Y under an intervention on Z cannot be confounded by a third variable. There are several ways in which a

⁸We have discussed such scenarios of confounded correlations as ‘common cause scenarios’ in Section 3.2.

manipulation of X might result in a change in Y due to such a confounder. I2, I3 and I4 each serve to avoid different kinds of scenarios in which this happens.

I2 states that the value of X should, in some scenarios, only depend on the value of I . This means that, for some values of I , other variables that normally affect X should be able to take whatever value in their domain without this affecting the value of X . I2 is often summarized by stating that interventions have to be *arrow-breaking*; they disconnect the value of the investigated variable from the values of other variables, as is represented in Figure 9.4. The rationale for I2 is the following. If some other variable Z can

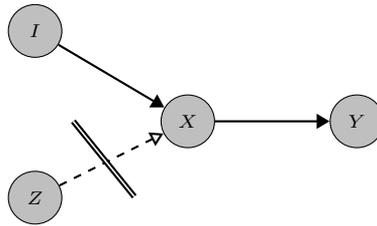


Figure 9.4: Intervention I renders the value of X causally independent of other variables Z

affect the value of X even when we are manipulating X , we cannot be certain that the changes in Y we observe are due to the manipulation of X rather than variations in this further variable Z which also affects X . Consequently, it is possible, for all we know, that the changes in Z cause *both* the changes in X and the changes in Y , thereby confounding the correlation between X and Y .

Consider the following example. We manipulate the number of tar-stained teeth in a population, but this number is still dependent on the smoking behaviour in the population as well. If we decrease the number of tar-stained teeth by our manipulation, and the population, for independent reasons, stopped smoking, we will observe a correlation between the decrease in tar-stained teeth and lung cancer risk. However, we should not conclude from this that tar-stained teeth cause lung cancer. After all, the decrease in lung cancer was not due to our manipulation on tar-stained teeth, but due to the decrease in smoking behaviour. Even though manipulations that fail I2 are

not by themselves *confounding*, because they do not directly affect the value of a confounder, one cannot be certain that there are no other confounders at work if the manipulation fails I2.

I3 states that I ought not to affect the value of Y independently from its influence through X . This is to avoid manipulations that are in themselves *confounding*; i.e. manipulations that affect another cause of the target variable when manipulating the investigated variable. Such would be the case, for example, if we manipulate the number of tar-stained teeth (T) in a population to test its effects on lung cancer risk (L) by manipulating the smoking behaviour (S) in that population, as is represented in Figure 9.5. Even if we

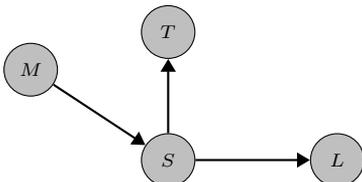


Figure 9.5: M on T violates I3

were to observe changes in L , we would not want to conclude that T causes L . After all, S does the actual causing, and T is just a symptom of this causing being done.

I4 embodies the interventionist dictum that interventions ought to be *exogenous* to the causal system under investigation. In order for manipulations to be indicative of causal relations in a system, the manipulation must come from outside the system.

This requirement works in two directions. First, the values of I ought not to affect the values of any variable that is causally relevant to Y , except for X or any variable on the causal path from X to Y . This is for the same reasons that prompted I3; manipulations that fail this requirement are themselves *confounders*. Second, the values of I ought not be dependent on other variables in the system. Suppose we were to manipulate the tar-stained teeth of only the heavy smokers with a family history of lung cancer in a certain population. We respect I3 by bleaching their teeth, rather than changing their smoking behaviour. After such an experiment we are likely to observe an inverse correlation between tar-stained teeth and the occurrence

of lung cancer: the research subjects with reduced tar stains on their teeth ended up having a higher risk of getting lung cancer. Of course, it would be wrong to conclude that tar-stained teeth are a cause of lung cancer, because our manipulation is causally dependent on two confounders: heavy smoking and a genetic predisposition to develop lung cancer. All in all, I4 avoids manipulations that causally affect or are causally affected by confounders of the target effect and the investigated variable.

(IV) imposes stringent requirements on interventions. In order for a manipulation to qualify as an intervention, one must control for possible confounding by *all* causes of the target variable when manipulating the investigated variable. This is because, unlike (M), (IV) is *not* relativized to some preselected variable set. Instead, it is relativized to a target effect Y ; it defines an intervention variable for X *with respect to* Y (cf. Woodward, 2003, p. 103). Consequently, the requirements comprised in (IV) apply to *all* cause variables of Y that do not lie on the causal path from I through X to Y , rather than just to those variables one chooses to include in the variable set under consideration. I should be (statistically) independent of all those variables (I4), should not affect Y via any of those variables (I3), and should be able to make the value of X independent of all those variables (I2). By adopting (IV), the interventionist thus extends the ‘holding fixed’-requirement in (M) to variables that are not included in the variable set. In particular, (IV) imposes such a ‘holding fixed’-requirement on any potential cause of the target variable that is not on the causal trajectory between the investigated variable and the target variable.

Given the purpose of I2, I3 and I4, this should come as no surprise. These requirements are supposed to stop us from treating confounded manipulations as interventions, and a manipulation can still be confounded by a variable that is not included in the variable set of choice. For example, suppose that we decide to investigate whether the colour of one’s teeth causes lung cancer relative to a variable set that only includes those two variables and a variable representing the price of apples (P). A plausible causal graph of the scenario is represented in Figure 9.6. However, if the ‘holding fixed’-conditions comprised in (IV) were relativized to our selected variable set, we would conclude that T is a cause of L . After all, there are manipulations of T that correlate

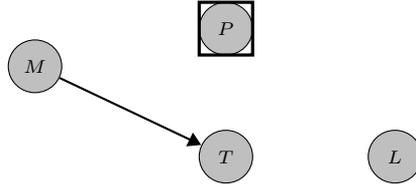


Figure 9.6: M manipulates T whilst only holding P fixed

with changes in L and do not affect the values of any other variable in the value set except for T and L . Namely those manipulations that change the value of T by changing the smoking behaviour. This is obviously the wrong result. We should still conclude that there is no causation between these two variables, because the correlation between tar-stained teeth and lung cancer is confounded by a variable that is not included in our variables set — i.e. smoking behaviour. If we want interventions to be indicative of causation, we should hold fixed *all* potential confounders, not just those that are in our variable set of choice. Consequently, (IV) is not relativized to variable set, but rather to the investigated variable and the target variable. I2, I3 and I4 *de facto* impose a ‘holding fixed’-requirement on any potential cause of the target variable that is not on the causal path between the investigated variable and the target variable. I will refer to these requirements collectively as the ‘holding fixed’-requirements from now on.

As we shall see, these ‘holding fixed’-requirements will play a central role in the debates on interventionism and mental causation. Before turning to these debates, I would like to draw attention to the stringency of these requirements and what it means for the relation between interventions and actual manipulations.

9.3 Interventions and actual manipulations

It is quite plausible that the ‘holding fixed’-requirements in (IV) make it practically impossible for an actual manipulation to qualify as an intervention. To see this, suppose our toy investigation was in fact executed: a group of researchers investigates whether, within a certain population, the occurrence

of tar-stained teeth (T) causes lung cancer risk (L). Aware of the obvious potential confounder S , i.e. the smoking behaviour in the population, the researchers diligently hold fixed its value when manipulating T , as represented in Figure 9.7 Even so, it is practically impossible to control for all potential

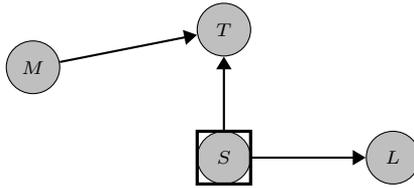


Figure 9.7: M manipulates T whilst holding S fixed

confounders. Here are four (only moderately far fetched) scenarios.

First, imagine that the mere realization that they are part of a medical experiment causes some of the research subjects to suffer from extreme stress. They reason that big pharma cannot be trusted and they are therefore certain to be treated with toxic substances. This stress decreases the efficacy of their immune system, which eventually raises the probability of these subjects incurring lung cancer. The manipulation violates I3 and I4 and therefore fails to be an intervention. Second, imagine that some of the research subjects feel

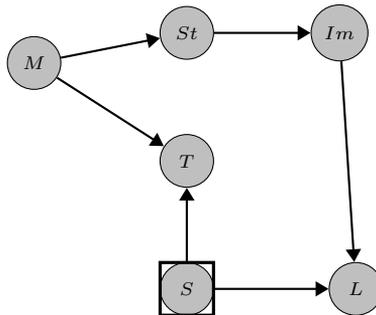


Figure 9.8: M manipulates both T and St , thereby affecting L

comforted by the regular medical check-ups that come with the investigation. They reason that, if they have a health issue, the investigators will pick up on it immediately and they subsequently stop worrying about their health altogether. The resulting decrease in stress causes an increase of their immune

system's efficacy, which eventually lowers the probability of these subjects incurring lung cancer. The manipulation violates I3 and I4 and therefore fails to be an intervention. The causal diagram for both these cases is provided in Figure 9.8

Third, imagine that the research is state funded. The application for funding was approved because the committee was aware of a steady increase in lung cancer risk among the global population. As it turns out, the increased lung cancer risk is mainly due to air pollution (AP). The air pollution caused an increase in lung cancer risk, which caused the approval (A), which caused the investigation (I), which caused the manipulation (M). Consequently, the manipulation fails I4, because it is not statistically independent of one of the causes of its investigated variable L .

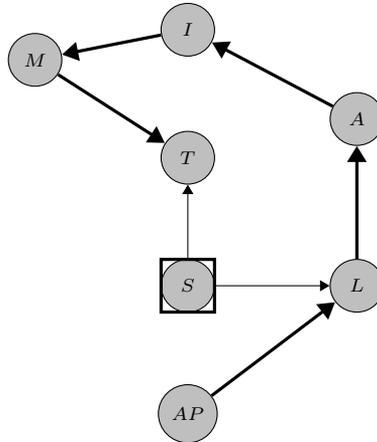


Figure 9.9: The manipulation M is indirectly caused by AP , which is also a cause of the target variable L

Finally, imagine a different experiment by the same research group. In order to eliminate stress as a confounder, they investigate the effects of stress (St) on the efficacy of the immune system (Is). Plausibly, they will have difficulty finding an actual manipulation that ‘acts as a switch’ or is ‘arrow-breaking’ in the way required by I2. Personal disposition and history have an influence on susceptibility to stress that is hard to eliminate. Consequently, it is hard to imagine a manipulation of stress level that ‘breaks’ these dependence relations. It is even harder to imagine that such a manipulation would

actually be executable and allowed by any ethical board.

These examples indicate that it is hard to make actual manipulations meet I2, I3 and I4. Some general considerations support the same conclusion. (IV) demands that interventions are *surgical* in that they do not affect any potential confounder, and *exogenous* in that they are not affected by any potential confounder. As we have seen, (IV) thereby extends the ‘holding fixed’-requirement in (M) to all variables that causally influence Y . In practice, it is impossible to keep track of and control for *all* potential confounders.⁹ Moreover, it is quite plausible that the phenomena in the actual world are interconnected in such a variety of complex ways that any actual manipulation is in fact *hamfisted*, rather than surgical, in that it affects multiple variables at once. Similarly, because the world is such a complex web of causal relations, actual manipulations might inevitably have causes in common with the target variable, rather than being exogenous to the investigated system. For these reasons, it might very well be practically impossible to intervene on any variable. At the very least, it appears that actual manipulations are unlikely candidates for interventions.¹⁰

These remarks are important because they counteract a potential confusion. Interventionism does *not* entail that A causes B if and only if *agents like us* can manipulate B by manipulating A . There are accounts in the causation literature come closer to claiming something like this, such as the agency account proposed by Menzies and Price (1993), but Woodward explicitly distances the interventionist account of causation from such proposals (e.g. Woodward, 2009). The notion of an intervention at work in interventionist theories of causation is a highly idealized one and should not be taken to pick out our everyday interactions with our environment. As we shall see in Section 14.3, there is some disagreement about *how* idealized these interventions should be. For now, we can think of them as “[...] events in which the ‘hand of God’ comes down and alters the value of X directly” (Franklin-Hall, 2016, p. 556–557).

⁹Recall that, according to contemporary physics, such confounders could be located at immense distances from the investigated cause and effect. See Section 7.2.

¹⁰*Pace* Woodward, who states that “[...] there will be realistic cases in which manipulations carried out by human beings will qualify as interventions [...]” (2003, p. 103). I find it hard to imagine such cases. Either way, it should be obvious that it is hard to make actual manipulations meet these requirements.

9.4 Summary

We can summarize our first pass at an interventionist account of causation as follows. Interventionism analyzes causation between two variables X and Y relative to a variable set V in terms of correlations between changes in X and changes in Y in scenarios where the causal facts surrounding X and Y meet the requirements provided in (M) and (IV). In practice, this means that X causes Y if and only if causing changes in X , whilst holding all possible causes of Y that do not lie on the causal path from our manipulation of X through X to Y , correlates with some changes in Y . Much more remains to be said about (M) and (IV). In the chapters to come, I present some problems for minimal interventionism and discuss some possible solutions. However, this rudimentary characterization of interventionism does suffice to explain how interventionism interacts with the issues of non-reductionist mental causation and causal exclusion. In the next chapter, I discuss the standard interventionist strategy to safeguard mental causation for non-reductionist physicalism, and why this strategy does not work for dualism.

Chapter 10

Interventionism and Non-Reductionism

In this chapter I investigate how minimal interventionism interacts with the problem of non-reductionist mental causation by drawing from recent discussions on non-reductionist physicalism and interventionist mental causation. It can seem that the so-called ‘holding fixed’-requirements in (M) and (IV) stand in the way of both non-reductionist physicalist mental causation and dualist mental causation. Upon closer scrutiny, it will appear that these requirements stand in the way of non-reductionist physicalist requirements only if they are applied more stringently than was intended. By contrast, they *do* stand in the way of dualist mental causation when applied as intended.

10.1 Interventionist exclusion

We often manipulate mental phenomena in order to bring about changes in the physical world. For example, we can make someone wear a coat by telling them it is chilly outside and we can reduce a person’s risk of depressive behaviour by reducing the number of instances in which she feels humiliated.¹ It can therefore seem that mental phenomena easily qualify as causes according to interventionism. However, it has been argued that intervention-

¹I borrow this example from Kendler and Campbell (2009).

ism, rather than redeeming non-reductionist mental causation, generates a powerful exclusion argument of its own, because manipulations on mental phenomena cannot meet the qualifications for being an intervention relative to behavioural effects (Baumgartner, 2009, 2010, 2013; Gebharder, 2017). In this section I briefly summarize this line of argument. In the remainder of this chapter I formulate the standard interventionist reply to this argument.

A cursory look at (M) provides an indication of the challenge for non-reductionists (Woodward, 2003, p. 59):

- (M) A necessary and sufficient condition for X to be a type-level direct cause of Y with respect to a variable set V is that there be a possible intervention on X that will change Y or the probability distribution of Y when one holds fixed at some value all other variables Z_i in V .

The definition states that X can be a cause of Y only if some intervention on X which holds the values of all other relevant variables fixed result in changes in Y . However, manipulations of mental phenomena can result in behavioural changes only under the supposition that these manipulations also change some physical phenomena. That is the upshot of *Physical Completeness*: there can be no change in physical phenomena without a change in their physical history. Plausibly, our actual manipulations are followed by behavioural changes because they also affect the physical phenomena underlying the mental phenomena. For example, manipulating one's sense of humiliation will affect their behaviour because it also affects the physical phenomena underlying the sense of humiliation. In general, it seems implausible that one can influence behaviour by manipulating mental phenomena whilst keeping all other physical phenomena, among which the physical phenomena underlying mental phenomena, fixed. This 'holding fixed'-requirement in (M) stands in the way of interventionist mental causation for non-reductionists of both the physicalist and the dualist variety.

Some have remarked that (M) only provides requirements for causation relative to a set of variables V . Non-reductionists could therefore claim that mental phenomena are causes of our behaviour relative to variable sets that do not include their physical bases.² However, circumventing 'holding fixed'-

²See Eronen (2012, Section 3), Eronen and Brooks (2014) and Polger et al. (2018) for

requirements is not that easy. As we have seen in Section 9.2, (IV) is *not* relativized to a preselected variable set and *de facto* imposes a ‘holding fixed’-requirement on any potential cause of the target variable. Even if one decides not to represent the nomic or metaphysical base of the mental phenomena in the variable set, (IV) still requires that one would hold that base fixed when intervening on the mental phenomenon. Moreover, recall that this de-relativized ‘holding fixed’-requirement plays an important role in the interventionist account: it prevents us from concluding that tar-stained teeth cause cancer relative to some creatively constructed variable sets.

So even if (M) on its own might not pose a problem for non-reductionists, (IV) *does* seem to pose a problem. Manipulations of mental phenomena that correlate with behavioural effects appear to violate I3 and I4, which require that an intervention does not causally affect any cause of the target effect other than the investigated variable, and that the intervention is statistically independent from any such other cause. Such manipulations appear to violate these requirements by causally affecting the physical phenomena that underlie the purported mental causes, as those physical bases plausibly affect our behaviour. Therefore, it appears that manipulations of mental phenomena cannot result in behavioural change without violating the ‘holding fixed’-requirements that are built into (IV).

We can summarize the interventionist exclusion argument as follows. According to (M), non-reductionist mental phenomena are causes of our behaviour if and only if there are interventions on mental phenomena that result in changes in our behaviour. Even though we frequently manipulate mental phenomena in order to affect behaviour, (IV) seems to indicate that no manipulation of mental phenomena can meet the requirements on interventions relative to the targeted behavioural changes. If there can be no such manipulations, there are no interventions on mental phenomena that result in behavioural changes and interventionism results in epiphenomenalism for the non-reductionist. Or so the opponents of interventionist models of non-reductionist mental causation argue.

a further development of such an approach to interventionism and exclusion. I will not engage with their proposal directly in the text, as it would deviate the discussion from my own proposal. Where appropriate, I will address their comments on views like mine in footnotes.

10.2 Drainage

Many philosophers do not believe the situation to be so dire. They maintain that there are good reasons to *exempt* some phenomena from the holding fixed requirements and that, once the relevant exemptions are in place, the non-reductionist physicalist is safe from the interventionist exclusion argument. In this section, I discuss Woodward's version of this response and show that it is of no help to the dualist.³

The 'holding fixed'-requirements in (M) and (IV) are designed to avoid picking out confounded correlations as causal. By applying these requirements to the phenomena that metaphysically necessitate the values of the manipulated phenomena, one treats these metaphysical bases as potential confounders of the correlation between the manipulated phenomena and the target effects. Woodward (2015) argues that it is a mistake to treat metaphysical bases as potential confounders in such cases. He traces this mistake to a misunderstanding about the scope of (M) and (IV). These definitions were developed to map the *causal* relations between variables. By representing metaphysical bases in interventionist models, one represents non-causal dependence relations like metaphysical necessitation in the models. Consequently, the heuristic use of the 'holding fixed'-requirements breaks down. He states (2015, p. 336):

One can't simply assume that because it is appropriate to control for ordinary confounders in cases in which no non-causal dependency relations are present, it must also be appropriate to control for factors like [metaphysical] bases which do represent non-causal dependency relations.⁴

Woodward contends that the 'holding fixed'-requirements should not apply to metaphysical bases of the investigated phenomena. He provides two arguments for this contention. First, he maintains that treating metaphysical bases as potential confounders by subjecting them to 'holding fixed'-requirements has absurd consequences: it eliminates the possibility of higher-level causation. Second, he argues that the 'holding fixed'-requirements can-

³See Weslake (2011) and Yang (2013) for similar replies.

⁴Woodward speaks of 'supervenience bases' instead of 'metaphysical bases', the differences between these two do not matter much for our purposes.

not be meaningfully applied to phenomena that metaphysically necessitate the purported causes to begin with. I will briefly discuss these two arguments here and explain why they do not apply to *nomic* bases. The upshot is that Woodward's arguments can safeguard non-reductionist physicalist mental causation, but are of no help for dualists.

First, Woodward remarks that applying the 'holding fixed'-requirements to metaphysical bases has absurd consequences (2015, p. 336–338). In particular, it eliminates the possibility of *higher-level* causation, i.e. causation by phenomena that are metaphysically necessitated by, but are not identical to, the physical phenomena underlying them. Phenomena like hurricanes, terminal diseases and banking crises have many effects, but none of these phenomena are identical to their underlying physical phenomena. After all, any hurricane, financial crisis, or terminal disease could have been metaphysically necessitated by a physical phenomenon that contains some greater or fewer number of electrons than its actual physical metaphysical base. Because these phenomena can be realized by different physical phenomena, none of them are identical with any one such physical realization.⁵ If we apply the 'holding fixed'-requirements to metaphysical bases, we have to conclude that there is no such thing as higher-level causation, and there is causation only at the fundamental physical level. That is a patently absurd conclusion.

This objection to applying the 'holding fixed'-requirement to metaphysical bases parallels a popular line of argument in the literature on causal exclusion arguments. Recall that, according to some versions of the causal exclusion argument, physical causes causally exclude all phenomena that are numerically distinct from these physical causes (cf. Chapter 5). Block (2003) argued that if all metaphysically necessitated phenomena are causally excluded by their metaphysical bases, and (almost) everything is metaphysically necessitated by the physical, (almost) all causation 'drains away' to the fundamental physical level and there can be no psychological, biological, economical, meteorological, or any other non-fundamental causation.⁶ Given that there manifestly *is* such *higher-level causation*, something must be wrong with exclusion arguments.⁷ The structure of the underlying argument

⁵This is a familiar point, found in Pereboom (2002) and Paul (2006, 2007) among others.

⁶A similar thought can already be found in Yablo (1992). See also Anthony (2015).

⁷Moreover, some doubt that there is any causation at the fundamental physical level

against exclusion arguments looks as follows:

1. If metaphysically necessitated phenomena are causally excluded by their metaphysical bases, there is causation only at the fundamental physical level.
 2. It is not the case that there is causation only at the fundamental physical level.
-
3. Therefore, metaphysically necessitated phenomena are not causally excluded by their metaphysical bases.

This *causal drainage* argument is easily translated to an argument about the ‘holding fixed’-requirement:

1. If the ‘holding fixed’-requirements apply to metaphysical bases, there is causation only at the fundamental physical level.
 2. It is not the case that there is causation only at the fundamental physical level.
-
3. Therefore, the ‘holding fixed’-requirements do *not* apply to metaphysical bases

Unless one is willing to deny 2, the ‘holding fixed’-requirement needs to be adjusted. Denying 2 results in universal causal drainage and thereby flies in the face of common sense and scientific practice. This provides a strong incentive to restrict the scope of the ‘holding fixed’-requirements. Moreover, denying 2 is particularly unattractive for those who maintain that mental causation is a problem specific to non-reductionism about the mind, be it of the physicalist or the dualist kind. After all, if there is no causation at any level except for the fundamental physical level, it can hardly be held against the non-reductionist that she cannot allow for causally effective mental phenomena. Non-reductionist mental phenomena would share their epiphenomenalist fate with banking crises, hurricanes and infections. Most of our causal

at all, see Russell (1912) and Field (2003). See Frisch (2014) for a counter-point. Consequently, causal drainage might entail that there is no causation at all. I engage with the question of causation and fundamental physics briefly in Chapter 14.

judgments would be wrong anyway, and it would therefore be unclear why *Mental Causation* should be maintained.

Even if one is willing to embrace causal drainage, there is another reason not to apply ‘holding fixed’-requirements to metaphysical bases. Applying the ‘holding fixed’-requirements to metaphysical bases requires us to consider metaphysically impossible scenarios. Given that phenomenon *A* is metaphysically necessitated by phenomenon *B*, it is metaphysically impossible to change *A* without changing *B*. Consequently, any scenario in which one intervenes on *A* whilst holding *B* fixed is metaphysically impossible. Woodward (2015, p. 335) argues that we cannot derive any useful causal information from such metaphysically impossible scenarios because we cannot coherently conceive such scenarios.⁸ For example, we cannot coherently conceive of a world that is physically identical to ours, but hurricane Katrina or the 2008 banking crisis never occurred. Consequently, we cannot derive any causal information from considering such a scenario. Or so the reasoning goes. The upshot is that applying the ‘holding fixed’-requirements to metaphysical bases not only results in causal drainage, it would require us to consider scenarios that are inconceivable.

Woodward takes these two considerations to apply to any dependency relation that is stronger than mere nomic or causal dependency (e.g. 2015, p. 308). The relations that he has in mind appear to be of the kind we have called ‘tight’ earlier in this text. Recall that tightness was defined as follows:

Tightness For any two phenomena *A* and *B*, *A* is tightly related to *B* if and only if there is some set of background conditions *c* that includes neither *A* nor anything that metaphysically necessitates *A*, such that *B* and *c* together metaphysically necessitate *A*.

I remark on two such tight relations which fall short of metaphysical necessitation.

First, Woodward discusses *definitional* relations between the variables:

⁸The non-reductionist physicalist replies to causal exclusion worries proposed in Bennett (2003, 2008) and Shapiro (2010) also explicitly rely on the metaphysical impossibility of such scenarios. Note however, that the conceivability of metaphysically impossible scenarios is a strained issue. Lewis (1973b, Section 1.6) and Williamson (2016) appear to share Woodward’s contention. Nolan (2017) and Berto and Jago (2019) disagree.

Total Cholesterol is defined as the sum of *High Density Cholesterol* and *Low Density Cholesterol* (Woodward, 2015, p. 327). The values of none of these variables metaphysically necessitates the value of any of the other *individually*, but the values of any two of these variables taken together *do* metaphysically necessitate the value of the third. For example, in all worlds where my *Low Density Cholesterol* is 100 mg/dL and my *Total Cholesterol* is 200 mg/dL, my *High Density Cholesterol* is 100 mg/dL. Suppose now that the risk of heart disease is causally dependent on all three of these variables. This should at least be a possible scenario. However, if we apply the holding fixed conditions to definitionally related variables, there is no possible intervention on the total cholesterol level that results in changes in heart disease risk. It is metaphysically impossible to change the value of the total cholesterol level without changing the value of the high density cholesterol level or the low density cholesterol level. It would thus be impossible for all three of these to cause heart disease risks. On the assumption that many definitionally related variables cause the same effects, it follows that applying the ‘holding fixed’-requirements to such variables would result in considerable drainage and will require us to derive causal information from metaphysically impossible scenarios.

Second, a similar case can be made for the realization relation we discussed in Chapter 5. We distinguished between the core realizer, the total realizer and the realized phenomenon. The realized phenomenon is metaphysically necessitated by the total realizer, but not by the core realizer, which is a salient part of the total realizer. Even so, it seems natural to exempt core realizers as well when intervening on realized phenomena. Given that we are allowed to bring about changes in the total realizer when intervening on the realized property, we should be allowed to bring about changes in the *parts* of the total realizers as well. It would be unprincipled to exempt the total realizer from the ‘holding fixed’-requirements, whilst imposing these requirements on its parts.⁹ Consequently, I will follow Woodward and take the considerations about metaphysical necessitation and ‘holding fixed’-requirements to apply to other tight relations as well. For clarity, I will attempt to restrict

⁹One guesses imposing such a requirement on a whole without imposing it on its parts would result in inconsistencies, but I take it that the rationale is intuitive enough without further argument.

my examples of tight relations to cases of metaphysical necessitation in what follows.

10.3 (M*) and (IV*)

Based on the above drainage objection and considerations about conceivability, Woodward argues that variables that stand in tight relations like metaphysical necessitation to the investigated variable and target variable should not be treated as ordinary confounders. They should be *exempted* from the ‘holding fixed’-requirements. That is to say, variables whose values cannot be fixed independently from the values of the investigated variable or the target variable because they are related by a tight relation ought *not* to be held fixed when one manipulates the investigated variable. In this section I discuss the relevant adjustments to the interventionist definitions in light of these findings, and explain why they are of no help to the dualist.

In light of the findings in the above section, we can reformulate the interventionist definition of causation as follows:¹⁰

(M*) A necessary and sufficient condition for X to be a type-level direct cause of Y with respect to a variable set V is that there be a possible intervention on X that will change Y or the probability distribution of Y when one holds fixed at some value all other variables Z_i in V *except for those that stand in a tight relation to X or Y .*¹¹

This definition requires some unpacking. In particular, we require an account of what it means for two variables to stand in a tight relation to one another. We can start by looking at a standard example of a tight relation: metaphysical necessitation. Being auburn metaphysically necessitates being red. Consequently, we can say that a variable A that represents a piece of cloth being auburn (a_1) or not (a_0) metaphysically necessitates a variable R

¹⁰Cf. Baumgartner (2010) and Woodward (2015).

¹¹Strictly speaking, it is not required to adjust (M) to allow for higher-level causation, because one can choose to omit bases from the relevant variable set. However, (M*) allows us to model causal relations in variable sets that contain non-causal and non-nomic dependency relations. This adjustment is therefore helpful for modelling higher-level causation, but not necessary.

that represents that cloth being red (r_1) or not (r_0). Note that this does *not* mean that all values of A metaphysically necessitate some value of R . After all, not being auburn does not metaphysically necessitate not being red; scarlet pieces of cloth are red as well. In the context of the drainage problem and the ‘holding fixed’-requirement we should require that only *some* values of a variable metaphysically necessitate *some* of the values of another variable in order to exempt the former from ‘holding fixed’-requirements when intervening on the latter.¹² This is because, when intervening, we should be allowed to make those changes that make a change in the investigated variable metaphysically possible. Consequently, even variables that can take *some* values that metaphysically necessitate a certain value of the investigated variable deserve exemption from the ‘holding fixed’-requirement. From here on, I will call such variables ‘metaphysical base variables’, and variables whose values are metaphysically necessitated by another variable under consideration ‘metaphysically necessitated variables’. Metaphysically necessitated variables and their respective metaphysical base variables are thus tightly related variables in virtue of the metaphysical necessitation relation between some of their variables.

We can extend on this example to include other cases of tightly related variables as well. If some values of variable A stand in a tight relation to some values of variable B , then variables A and B are tightly related, where a tight relation is understood as defined in the previous section. As we have seen, examples of tight relations in addition to metaphysical necessitation include realization, definitional relations, and (partial) grounding. For simplicity, we will focus on cases of metaphysical necessitation in what follows, but I take my arguments to generalize to other cases of tight relations as well.

With these remarks taken care of, we can turn to the adjusted definition of an intervention. The change in a variable I counts as an intervention on the variable X relative to a target effect variable Y if and only if this change in I causes the value taken by X and I is an *intervention variable* for X with respect to Y . We can define ‘intervention variable’ as follows:

¹²I follow an explicit proposal found in Woodward (2015, p. 315) and Woodward (2018, p. 15), although he prefers the terms ‘logical’ and ‘conceptual’ over ‘metaphysical’. Based on the examples he uses, I take it there is nothing more than a terminological difference here.

I is an intervention variable for X with respect to Y if and only if I meets the following conditions:

(IV*)

I1 I causes X .

I2 I acts as a switch for all other variables that cause X . That is, certain values of I are such that when I attains those values, X ceases to depend on the values of other variables that cause X and instead depends only on the value taken by I .

I3* 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 either *tightly related to X or Y* , or 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 .

I4* I is (statistically) independent of any variable Z that (a) *is not tightly related to X or Y* , (b) causes Y and (c) is on a directed path that does not go through X .

In order to capture these adjustments in the causal diagrams, we will represent tight dependence relations like metaphysical necessitation with double lined arrows and reserve single lined arrows for causal relations. A simple representation of a higher-level causation relation is provided in Figure 10.1.¹³

¹³Eronen and Brooks (2014) argue that (M*) and (IV*) and their corresponding diagrams are in fact incapable of representing non-causal dependence relations *and* incapable of distinguishing between the causal roles of higher-level phenomena and their bases. I will not engage with their first objection in this text, but as I understand it, interventionist models were never supposed to capture non-causal dependence relations and it is worth noting that their own proposal bans non-causal dependency relations from interventionist models altogether. My adjustments to minimal interventionism will go some way towards addressing their second criticism. See Section 15.3.

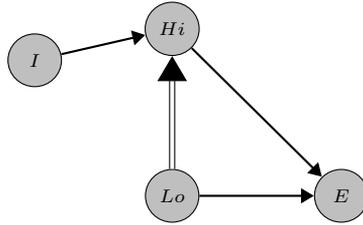


Figure 10.1: Hi is metaphysically necessitated by Lo ; both cause E

According to the resulting account of causation, there is *causal inheritance* instead of universal causal drainage. That is to say, higher-level phenomena *inherit* causal powers from the lower-level phenomena that metaphysically necessitate them, rather than having their causal powers draining away towards their metaphysical bases. After all, any intervention on a metaphysically necessitated phenomenon will also be an intervention on its metaphysical base (cf. Woodward, 2015, p. 331). Whatever effects such an intervention on the physical base might have on a target variable, will, according to (M*) and (IV*), be an effect of the metaphysically necessitated phenomenon as well. In Chapter 11, I will argue that this resulting picture is too generous in ascribing causal power to higher-level phenomena. For now however, it is worth pointing out how this minimal account allows for non-reductionist physicalist mental causation, but not for dualist mental causation.

Non-reductionist physicalists can benefit from the resulting model of causation. After all, they maintain that mental phenomena are *metaphysically* necessitated by physical phenomena, as is represented in Figure 10.2. Mental causation is just a variety of higher-level causation according to their view. Consequently mental phenomena can inherit causal powers from their physical metaphysical bases, because all interventions on mental phenomena are also interventions on their underlying physical phenomena. By adopting the more perspicuous definitions (M*) and (IV*) the non-reductionist physicalist can convincingly respond to mental causation worries.

However, these clarifications are of no help to the nomic naturalist dualist. As we have seen in Chapter 5, she typically maintains that there are no tight relations between mental phenomena and their underlying physical

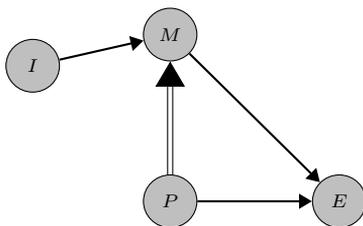


Figure 10.2: Mental phenomenon M is metaphysically necessitated by P ; both cause E

phenomena. The former are merely *nomically* necessitated by the latter, and nomic bases are not exempted from the ‘holding fixed’-requirements. It thus appears that minimal interventionism, which consists of only (M*) and (IV*), remains inhospitable to dualist mental causation.¹⁴

Moreover, Woodward’s arguments for exempting metaphysical bases from holding fixed requirements do not carry over to nomic bases. Requiring that nomic bases are held fixed when one intervenes on nomically necessitated phenomena does *not* result in universal causal drainage.¹⁵ Furthermore, changes in nomically necessitated phenomena without changes in their nomic bases *are* metaphysically possible and informative according to the dualist.¹⁶ We have not yet seen any reason to maintain that scenarios in which one intervenes on a nomically necessitated phenomenon whilst holding its nomic base fixed are uninformative. Applying the ‘holding fixed’-requirements to nomic bases does not lead to absurd consequences, nor is it absurd to begin with. So not only is the current model inhospitable to dualist mental causation, it appears that the measures that make it inhospitable are well-motivated.

It is worth making one further remark on the ‘holding fixed’-requirements. Having taken a closer look at how these requirements interact with the issue of non-reductionist mental causation, we can see how minimal interventionism still pays lip service to heavyweight accounts of causation (cf. Chapter 4). In particular, minimal interventionism pays lip service to the idea that causes

¹⁴Woodward (2008, p. 253–254) also suggests that his adjustments will be of no help for the ‘minority position’ that takes the relation between mental properties and their underlying physical properties to be causal, i.e. dualism.

¹⁵Or at least, that is the case if one accepts the *Physicalism about the Non-Mental* claim that is characteristic of nomic naturalist dualism.

¹⁶That is the whole point of conceivability arguments for dualism (cf. Chalmers (1996)).

are physically sufficient for their effects. Even though (M*) and (IV*) do not require that all causes are sufficient for their effect, they do require that all causes are tightly related to the strictly physically sufficient phenomena for the target effect. After all, being tightly related to that phenomenon is the only way to exempt it from the ‘holding fixed’-requirements, and if one changes a purported cause variable whilst holding fixed the physically sufficient phenomenon for the target effect, there will be no change in the target effect. Consequently, causes must be tightly related to the physically sufficient phenomena of their effects according to (M*) and (IV*).

This requirement supports the exclusion arguments against dualism. That is to say, *Physical Exclusion* is true according to minimal interventionism. Remember the crucial premise in the exclusion argument as formulated earlier:

Physical Exclusion For any three phenomena A , B and C : if A occurs at t and physically necessitates B ’s occurrence at $t + x$, no phenomenon C occurring at t that is not tightly related to A and is not tightly related to any of A ’s parts is a cause of B , unless it is a case of genuine overdetermination.

This premise states that non-overdetermining causes must be tightly related to a physically sufficient phenomenon for the target effect. As we established in Chapter 3 and 5 allowing for dualist mental causation will probably require denying principles like *Physical Exclusion*. It should thus come as no surprise that minimal interventionism does not allow for dualist mental causation.

10.4 Summary

The purpose of this chapter was to discuss how minimal interventionism interacts with the issue of non-reductionist mental causation. We can summarize the situation as follows. Unless one exempts the physical phenomena underlying mental phenomena from the ‘holding fixed’-requirements in (M) and (IV), minimal interventionism appears inhospitable to mental causation. The current literature contains convincing arguments for exempting these underlying physical phenomena *if* they stand in a tight relation to the purported mental causes. These arguments motivate some adjustments to (M)

and (IV). However, these arguments do *not* apply to *nomic* bases. Given that the dualist maintains that mental phenomena are merely *nominally* necessitated by their underlying physical phenomena, minimal interventionism provides no model for dualist mental causation.

It is up to the dualist to argue that nomic bases deserve an exemption from the ‘holding fixed’-requirements as well. I believe that the most convincing way to do so is to focus on the original motivation for the ‘holding fixed’-requirements. These requirements are designed to control for potential confounders. Cases of higher-level causation have indicated that certain restrictions on the ‘holding fixed’-requirements are in order. For example, applying the ‘holding fixed’-requirements to metaphysical bases can wrongfully treat them as confounders of the correlations between the phenomena that are metaphysically necessitated by these bases and the target effect. The dualist should argue that the same holds for nomic bases: applying the holding fixed conditions can wrongfully treat them as confounders of the correlation between the nominally necessitated phenomenon and the target effect. It would follow that nomic bases ought to be exempted from ‘holding fixed’-requirements as well.

The rest of this dissertation will develop an argument for extending the exemption of ‘holding fixed’-requirements to nomic bases. This will require a serious amount of patience and effort. As we have seen, nomic bases certainly seem like prime examples of confounders (cf. Section 3.2), interventionists emphasize that the relation between the exempted variables and the purported cause variables should be non-nomic and non-causal (e.g. Woodward, 2015, p. 308), and the two motivations for believing that bases cannot be confounders of the phenomena they metaphysically necessitate do not carry over to nomic bases.

Dualists will thus have to present other arguments to motivate exempting nomic bases as well. I will argue that there are considerations from general philosophy of causation that support such an exemption. In particular, I will argue that some of the problems that philosophical accounts of causation generally, and interventionism specifically, meet prompt solutions that support the exemption of nomic bases. These problems, arranged in the order we will discuss them are: the problem of spurious higher-level causation,

the dilemma of negative causation, and the problem of making causation compatible with physics. Once the solutions to these problems are in place, applying the holding fixed requirement to nomic bases will appear redundant and undermotivated. The upshot is that a brand of interventionism that is resistant to these three problems should exempt nomic bases from ‘holding fixed’-requirements. At the very least, it will lack the resources to motivate applying these requirements to nomic bases. In the next chapter, I address the problem of spurious higher-level causation.

Chapter 11

Spurious Higher-Level Causation

In the previous chapter, we saw that the interventionist replaced her original definitions (M) and (IV) with (M*) and (IV*) to avoid confusion when modelling higher-level causation. On the resulting account, metaphysically necessitated phenomena can *inherit* the causal powers of their metaphysical bases, rather than being causally excluded by them. In this section, I will argue that minimal interventionism allows for *too much* causal inheritance and thereby spuriously ascribes causal powers to higher-level phenomena. This is because (M*) and (IV*) exclude the possibility that metaphysical bases are confounders for the correlation between their effects and the phenomena that are metaphysically necessitated by these bases. However, there is strong evidence that metaphysical bases sometimes *do* confound such correlations. Consequently, metaphysical bases sometimes threaten the causal status of the phenomena they metaphysically necessitate, and minimal interventionism is not equipped to deal with this possibility. I propose a solution to this problem, briefly compare it with proposals by Woodward (2008, 2015) and List and Menzies (2009), and draw some conclusions for our overall project.

11.1 The problem

According to minimal interventionism, a phenomenon A can inherit the causal powers of another phenomenon B relative to a target effect E by meeting two criteria. First, phenomenon A needs to stand in tight relation to phenomenon B . Second, there has to be an intervention on A that, by affecting the occurrence of B , results in a change in the target effect. Structurally, the scenarios should correspond to the representation in Figure 11.1. It is easy to find phenomena that meet these criteria for a given effect, but do

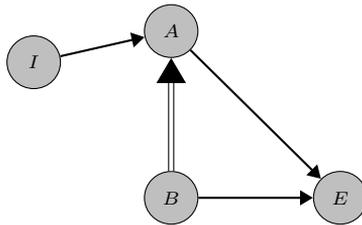


Figure 11.1: A is metaphysically necessitated by B ; both cause E

not seem to be causes of that effect. Consider the following three examples.

First, consider Yue. Yue died because he ingested half a gram of cyanide and that is exactly the lethal dose for a human of his constitution. His ingesting half a gram of cyanide metaphysically necessitates his ingesting a milligram of cyanide. Moreover, any intervention that stopped him from ingesting a milligram of cyanide would have saved his life. Nevertheless, we would not want to say that Yue died because he ingested a milligram of cyanide. After all, if he had merely ingested one milligram of cyanide, he would have survived.¹

According to minimal interventionism, Yue's ingesting a milligram of cyanide is a cause of his death. To see this, consider the following variable set.² The variable mg represent Yue's ingesting a milligram of cyanide, the variable $0,5g$ represent his ingesting half a gram of cyanide, and the vari-

¹Cf. List and Menzies (2009, p. 483) and Walter (2010).

²In this example and the two that follow it, the variables are selected to bear out the spurious higher-level causation problem. One should not assume however, that selecting the variables differently will make the problem go away. After all, the conditions comprised in (IV*) apply equally to variables outside of the variable set under consideration (cf. Sections 9.2 and 10.1).

able D represent Yue dying (roughly) at the time he did. For simplicity's sake, we can say that all of these variables only have an 'on' value and an 'off' value. The I variable represents an intervention as usual. The relevant causal graph is provided in Figure 11.2. Relative to this variable set, there

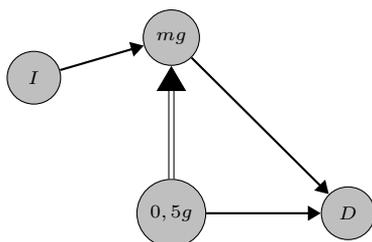


Figure 11.2: ingesting mg is metaphysically necessitated by ingesting $0, 5g$; both cause the death D according to minimal interventionism

is an intervention on mg that meets the requirements in (IV*) and results in a change in the value of D . In particular, suppose that all variables are at their 'on' value. In such a scenario, *all* interventions on mg will affect D . Consequently, minimal interventionism spuriously treats mg as a cause of D .

Second, consider Sarah. Sarah is a pigeon trained to peck exclusively at auburn objects. She is presented with an auburn pebble, which causes her to peck. The pebble being auburn metaphysically necessitates its being red. Moreover, any intervention on the pebble's being red would have stopped Sarah from pecking at it, because all non-red pebbles are non-auburn as well. Nevertheless, we would not want to say that Sarah pecked because the pebble was red. After all, she would not have pecked if the pebble had been red but not auburn.³

According to minimal interventionism, the redness of the pebble is a cause of Sarah's pecking. To see this, consider the following variable set. The variable R represents the pebble being red, the variable A represents the pebble being auburn, and the variable P represents Sarah pecking. For simplicity's sake, we can say that all of these variables only have an 'on' value and an 'off' value. The I variable represents an intervention as usual. The relevant causal graph is provided in Figure 11.3. Relative to this variable set, there

³Cf. List and Menzies (2009, p. 494), Yablo (1992, fn. 23) and Zhong (2019, p. 4).

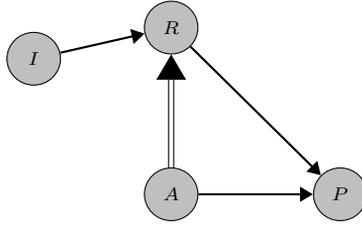


Figure 11.3: Redness (R) is metaphysically necessitated by Auburn (A); both cause the pecking (P) according to minimal interventionism

is an intervention on R that meets the requirements in (IV*) and results in a change in the value of P . In particular, suppose that all variables are at their ‘on’ value. In such a scenario, *all* interventions on R will affect P . Consequently, minimal interventionism spuriously treats R as a cause of P .

Third, consider the copper wire connecting my keyboard to my computer. This wire is opaque. The exact physical realization of the wire metaphysically necessitates the wire’s opacity. Moreover, there is an intervention on the opacity of the wire that would have stopped this ‘p’ from appearing on this screen, namely one that changes the opacity of the wire connecting my keyboard to my computer by replacing it with a transparent plastic wire. Nevertheless, we would not want to say that the wire’s being opaque caused this ‘p’ to appear on my screen. After all, if the wire was opaque but did not conduct electricity, no ‘p’ would have appeared on my screen.⁴

According to minimal interventionism, the opacity of the wire causes the occurrence of the ‘p’. To see this, consider the following variable set. The variable O represents the wire being opaque, the variable EP represents the wire having its exact physical realization, and the variable p represents the occurrence of a ‘p’ on my screen. The I variable represents an intervention as usual. For simplicity’s sake, we can say that all of these variables only have an ‘off’ value and an ‘on’ value. The relevant causal graph is provided in Figure 11.4. Relative to this variable set, there is an intervention on O that meets the requirements in (IV*) and results in a change in the value of p . In particular, suppose that all variables are at their ‘on’ value. In such

⁴Cf. Jackson and Pettit (1990a, p. 204).

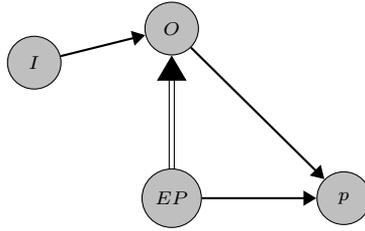


Figure 11.4: The opacity of the wire (O) is metaphysically necessitated by its exact physical realization (EP); both cause the occurrence of ‘ p ’ (p) according to minimal interventionism

a scenario, many interventions on O will affect the probability of p taking its ‘on’ or its ‘off’ value. Consequently, minimal interventionism spuriously treats O as a cause of p .

These examples bear out a general problem for (M^*) and (IV^*) . By allowing *all* phenomena to inherit the causal powers of their metaphysical bases, minimal interventionism allows for too much causal inheritance. Sometimes, phenomena are metaphysically necessitated by causes without inheriting their causal powers. Minimal interventionism fails to distinguish between such spurious cases of higher-level causation and good cases of higher-level causation. In order to provide the right fix for this problem, we need to pin down what distinguishes the spurious cases of higher-level causation from the good cases. In the next section, I propose that the relevant difference lies in the correlation patterns that are involved in these cases. These patterns indicate that the metaphysical bases are confounders in cases of spurious higher-level causation, whereas in good cases of higher-level causation they are not.

11.2 The screening off pattern

In this section, I propose to analyze the spurious higher-level causation problem in terms of correlation patterns. In particular, I argue that the relevant difference between spurious cases of higher-level causation and good cases can be explained in terms of *screening off* relations holding between the variables

involved.⁵ First, I introduce the screening off relation and show how it sheds light on some simple cases of spurious higher-level causation and good higher-level causation. Second, I show how more complex cases can be explained by a more specific variety of this relation. Third, I conclude that the relevant screening off pattern indicates that metaphysical bases can in fact be confounders for correlations between their target effects and the phenomena that are metaphysically necessitated by these bases. In the next section, I propose a solution to the spurious higher-level causation problem based on this analysis and briefly compare it to some other solutions.

A variable is often said to threaten the causal status of another variable relative to a certain effect by *screening off* the correlation between the other variable and the target effect. First, let us stipulate that there can be a correlation between changes in C and changes in B only if there can be changes in B and C . The screening off relation can now be defined as follows:

Screening Off For any three non-identical variables A , B and C , A *screens off* the correlation between B and C if and only if, for every value x of A , if A is held fixed at x , then there is no correlation between changes in the value of C and changes in the value of B .⁶

Put less formally, this means that, if A screens off the correlation between B and C , then the correlation between changes in B and changes in C disappears if we restrict our population to cases where A has one fixed value. For example, if one controls for smoking behaviour by restricting the population to cases with identical (or very similar) smoking behaviour, the correlation between changes in the colour of one's teeth and variations in lung cancer risk disappears. In this situation, it is natural to conclude that smoking behaviour confounds the correlation between tar-stained teeth and the risk of

⁵Woodward has recently proposed an account in some ways similar to the one developed here. See Woodward (2018). I postpone until a later occasion a comparison between the two accounts. Relative to the current project, the most relevant difference between our two accounts is that mine focuses more on the parallel between confounding by a common cause and confounding by a metaphysical base, whereas Woodward's focus is on selecting the right level of abstraction for causal explanations.

⁶In the interest of continuity with (M*) and (IV*), I stick to the interventionists's preferred terminology of 'variables' and 'values' when defining these terms as it is more natural to formulate claims about correlations in terms of variables and their values.

incurring lung cancer. That is to say, smoking behaviour does the actual causing, whereas the correlation between tar-stained teeth and lung cancer is spurious. Perhaps we can analyze cases of spurious higher-level causation in the same way.

In many cases of spurious higher-level causation, the relevant variables exhibit the same screening off pattern. In particular, the correlation between the metaphysically necessitated phenomenon and the target effect is often screened off by the metaphysical base in such cases. For example, recall Sarah, the pigeon trained to peck exclusively at *auburn* objects. She is presented with an auburn pebble, and subsequently pecks at it. The causal graph is provided in Figure 11.5. Even though the pebble being red is meta-

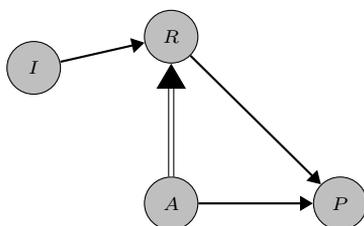


Figure 11.5: Redness (R) is metaphysically necessitated by (A); both cause the pecking (P) according to minimal interventionism

physically necessitated by its being auburn, it seems wrong to say that the pebble being red caused her to peck at it. In this case, the pebble being auburn screens off the correlation between redness and Sarah's pecking: if the being-auburn variable is held fixed at one particular value, changes in the redness no longer correlate with changes in the pecking. This is easy to demonstrate. Suppose that the being-auburn variable is held fixed at 'on': the pebble is auburn. In this scenario, changes in the redness do not correlate with changes in the pecking because there can be no changes in the redness. After all, the pebble being auburn metaphysically necessitates its being red. Alternatively, suppose that the being-auburn variable is held fixed at 'off': the pebble is not auburn. In this scenario, changes in the redness do not correlate with changes in the pecking either, because Sarah pecks at red objects only if they are auburn. Consequently, being auburn screens off the correlation between being red and Sarah's pecking, just like smoking behaviour

screens off the correlation between the colour of one's teeth and one's risk of getting lung cancer. In both cases, the screened off correlation is spurious, and the actual causing is done by the variable that does the screening off — the 'screen variable' for short.

Compare this to a simple good case of higher-level causation. Suppose Sophie is a pigeon trained to peck at exclusively *red* objects.⁷ She is presented with an auburn pebble, and subsequently pecks at it. It seems right to say that the pebble being red caused Sophie to peck. In Sophie's case, the correlation between the metaphysically necessitated phenomenon and the target phenomenon is *not* screened off by the metaphysical base. True, if the being-auburn variable is held fixed at the 'on' value, changes in the redness do not correlate with changes in the pecking because there can be no changes in the redness. However, if the being-auburn variable takes value 'off', there is still a strong correlation between changes in the pebble's being red and Sophie's pecking. Therefore, it is *not* the case that correlations between changes in the value of the metaphysically necessitated variable disappear for *any* fixed value of the metaphysical base variable. The correlation between changes in the metaphysically necessitated variable and changes in the target variable is to some extent independent of the value of the metaphysical base variable. To speak in terms of phenomena, we can say that, in good cases of higher-level causation, the occurrence of the metaphysically necessitated phenomenon affects the occurrence of the target phenomenon independently of the occurrence of the metaphysical base phenomenon. Whereas in cases of spurious higher-level causation, it does not.⁸

Based on these two simple cases, one might be inclined to conclude that the difference between good cases of higher-level causation and spurious cases boils down to the difference between correlations that are screened off and correlations that are not. This would provide us with a convenient similarity between spurious higher-level causation cases and standard cases of confounding by a common cause. After all, in both cases the spurious correlation is

⁷This example has become commonplace in philosophy of causation and is originally due to Yablo (1992, p. 257).

⁸It is worth noting that some authors say that metaphysical bases always screen off their metaphysically necessitated phenomena from target effects (e.g. Jackson and Pettit, 1990b, p. 111). However, they appear to operate on another notion of 'screening off' that is closer to 'pre-emption'.

screened off by a third variable that appears to do the actual causing. Consequently, we would have a general explanation at our disposal to dismiss both kinds of spurious correlations as non-causal: the correlation is spurious because it is screened off by a third variable.

However, a brief look at some more complex cases demonstrates that the situation is not that simple. There are spurious cases where the relevant correlation is *not* screened off by the metaphysical bases, and there are good cases where the relevant correlation *is* screened off by the metaphysical base. I will discuss these cases in that order, and show how they can be handled by relying on the screening off relation as well.

First, consider the case of the opaque wire again. I press the ‘p’-key on my keyboard and a ‘p’ occurs on the screen (p). The wire connecting my keyboard to my screen is opaque (O) and the opacity of that wire is metaphysically necessitated by its exact physical realization (EP). For simplicity’s sake, we can say that all of these variables only have an ‘off’ and ‘on’ value. The relevant causal graph is provided in Figure 11.6. Now, let us

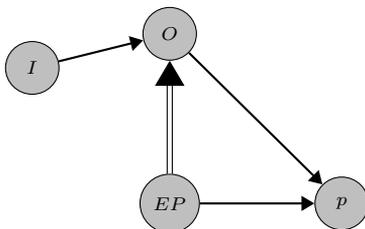


Figure 11.6: The opacity of the wire (O) is metaphysically necessitated by its exact physical realization (EP); both cause the occurrence of ‘p’ (p) according to minimal interventionism

stipulate that there is a strong correlation between the opacity of materials and their conductivity.⁹ If this is indeed the case, it is not obvious that the exact physical realization of the wire screens off the correlation between the opacity of the wire and the appearance of the ‘p’. If we hold EP fixed at its ‘off’ value, the correlation between changes in the opacity and changes in the occurrence of ‘p’ does not disappear.

Explaining such cases requires a slight adjustment to our analysis. Af-

⁹Even if this stipulation is false, I take there to be similar cases.

ter all, the metaphysical base of the wire's opacity does not screen off the correlation between the opacity and the appearance of 'p'. However, the metaphysical base *does* metaphysically necessitate the values of a variable that *does* screen off this correlation: the conductivity of the wire. After all, in scenarios where the wire conducts electricity, changes in its opacity do not correlate with changes in the occurrence of 'p', nor do they correlate in scenarios where the wire does *not* conduct electricity. The conductivity of the wire screens off the correlation between the opacity of the wire and the occurrence of the 'p'. The relevant causal graph is provided in Figure 11.7.

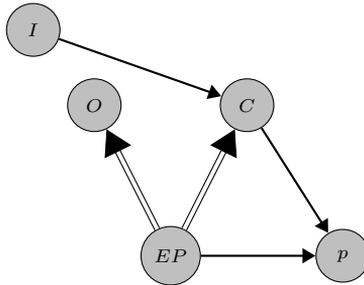


Figure 11.7: The opacity of the wire (O) and the conductivity of the wire (C) are both metaphysically necessitated by its exact physical realization (EP), (C) and (EP) cause the occurrence of 'p' (p), but (O) does not.

We can thus extend our analysis of spurious higher-level causation cases as follows. The correlation between a metaphysically necessitated phenomenon and a target effect is *spurious*, and therefore not indicative of causation if and only if, the metaphysical base screens off the correlation between the metaphysically necessitated phenomenon and the target effect *or* the metaphysical base metaphysically necessitates a screen variable for that correlation.¹⁰ This adjustment allows our analysis to capture more complex cases of spurious higher-level causation. It now remains to extend our analysis to capture outstanding cases of good higher-level causation.

Consider a variation on Sophie's case. There are several pebbles scattered

¹⁰Perhaps this is still too strong. Perhaps the metaphysical base could simply be tightly related to the screen variable without metaphysically necessitating it. I do not see why that should be impossible. However, I cannot come up with an example either. For this reason, and because the relevant fix is straightforward, I leave the definition as is.

around Sophie. All of these pebbles have one of the following four colours: scarlet, auburn, cyan and turquoise. Recall that Sophie exclusively pecks at red objects. Suppose now that she pecks at an auburn object. It seems right to say that she pecks at this pebble because it is red. Nevertheless, there is a metaphysical base variable that screens off the correlation between the redness variable and the pecking variable. To see this, consider the following variable set. R represents redness, and takes two values ‘on’ and ‘off’. S represents the specific shade of the pebble, and takes four values: ‘scarlet’, ‘auburn’, ‘cyan’, and ‘turquoise’. P represents Sophie’s pecking behaviour and only takes two values ‘on’ and ‘off’. The relevant causal graph is represented in Figure 11.8. In this scenario, the values of S metaphysically

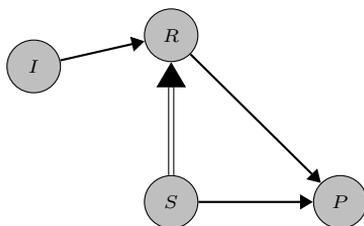


Figure 11.8: Redness (R) is metaphysically necessitated by the specific shade (S); both cause the pecking (P).

necessitate the values of R . Moreover, S also screens off the correlation between R and P . After all, if one holds S fixed at any one value, there can be no change in the value of R . Consequently, there can be no changes in R that correlate with changes in P when one holds S fixed at any particular value. This seems to be a good case of higher-level causation where the metaphysical base *does* screen off the correlation between the metaphysically necessitated phenomenon and the target effect.

Upon closer scrutiny, it appears that such cases can be analyzed in terms of the screening off relation as well. Even though it is the case that S screens off the correlation between R and P , it is also the case that R screens off the correlation between S and P . This is easy to demonstrate. First, consider scenarios in which R takes the ‘on’ value. The values of S are now restricted to ‘scarlet’ and ‘auburn’, and changes between these two values do not correlate with changes in P . Second, consider scenarios in which R takes the ‘off’ value.

The values of S are now restricted to ‘cyan’ and ‘turquoise’, and changes between these two values do not correlate with changes in P . Therefore, R screens off the correlation between S and P . Given that S was also shown to screen off the correlation between R and P , we can say that R and S *symmetrically* screen off one another from P .

Compare this to cases of spurious higher-level correlation. In Sarah’s case, we established earlier that the variable A , which represents the pebble being auburn, screens off the correlation between R and P . It is *not* the case that R in return screens off the correlation between A and P . If we hold R fixed at ‘on’, changes in whether or not the pebble is auburn will still correlate with changes in the value of P . Consequently, it is *not* the case that holding R fixed at any value will stop the correlation between changes in A and changes in P . We can say that the correlation between R and P is *asymmetrically* screened off by A :

Asymmetric Screening Off For any three non-identical variables A , B and C , A *asymmetrically screens off* the correlation between B and C if and only if (i) A screens off the correlation between B and C , and (ii) B does *not* screen off the correlation between A and C .

This, I submit, marks the relevant difference between cases of spurious higher-level causation and good cases of higher-level causation: in spurious cases, the correlation between the metaphysically necessitated phenomenon and the target effect is *asymmetrically screened off* by either (i) its metaphysical base, or (ii) a phenomenon that is metaphysically necessitated by that base. In good cases, the correlation between the metaphysically necessitated phenomenon is *not* asymmetrically screened off by either (i) its metaphysical base, or (ii) a phenomenon that metaphysically necessitated by that base.

Two disclaimers about this analysis are in order. First, the situation is likely to be less black and white than is captured in the definition of *Asymmetric Screening Off*. We might in fact allow for causation by a metaphysically necessitated variable that is screened off by its metaphysical base, but does not screen off changes between *all* of the values of that base. For example, suppose that Aubrey the pigeon is trained to peck exclusively at

red objects, *except* for those that are a very specific shade of auburn.¹¹ There are several pebbles scattered around Aubrey. These pebbles have a variety of different colours, including many shades of red, and there is one pebble that is the very specific shade of auburn that Aubrey does not peck at. If Audrey pecks at one of the red pebbles, it seems right to say that the redness of the pebble caused her to peck. However, the variable that takes all the shades of the pebbles in their full specificity as its values will asymmetrically screen off the correlation between changes in redness and Audrey's pecking behaviour. After all, if the redness variable is held at 'red' there are still changes in the specific shade variable that correlate with changes in the pecking behaviour. In particular, changing the value from any shade of red other than the specific auburn to that specific auburn (or *vice versa*) will correlate with a change in the pecking behaviour. Hence, redness does not screen off the specific shade variable. However, the specific shade variable screens off redness for reasons that are by now familiar: once one fixes the variable at one value, there can be no changes in the redness variable.

In such cases, we can say that redness *mostly* screens off the specific shade variable. For *most* of the values of the metaphysical base variable, it is the case that changing from or to that value does not correlate with changes in Audrey's pecking if one holds fixed the redness variable at any of its values. It will no doubt be hard to pin down exactly how much a metaphysically necessitated variable should screen off its metaphysical base to maintain its causal status. For instance, it is possible that the strictness we impose on the screening off behaviour of metaphysically necessitated variables varies with the context. If, for example, we cannot manipulate at such a fine-grained level as is represented in the metaphysical base variable, we might be more forgiving with regards to the screening off behaviour of the metaphysically necessitated variable. The interaction between context, variable selection and causation deserves further discussion. I postpone this discussion to Chapter 13, where we will discuss cases where omissions are causes, because such negative causation cases better bear out the intricacies of that interaction.

For the sake of simplicity, I ignore grey-area cases like Aubrey's in what follows. I expect that accounting for all of these cases will be very difficult,

¹¹See Woodward (2018, p. 20–22) for similar examples from thermodynamics and fluid flow modelling.

but I also expect that the screening off pattern will play an integral role in accounting for them. It still seems to be the case that the metaphysically necessitated variable counts as a cause in virtue of *mostly* screening off the metaphysical base variable. So even with this first disclaimer in mind, I do think that the asymmetric screening off correlation patterns mark the relevant distinction between spurious cases and good cases of higher-level causation.

The second disclaimer concerns *gerrymandered* variables. As is demonstrated by Franklin-Hall (2016), one can gerrymander a variable that *does* asymmetrically screen off the correlation between the metaphysically necessitated variable and the target effect in most good cases of higher-level causation. For example, one could do so by constructing a variable that takes the disjunction of all the sufficient phenomena for the target effect as its ‘on’ value, and the disjunction of all the sufficient phenomena for the non-occurrence of the effect as its ‘off’ value. The resulting variable will screen off the correlations between most plausible cause variables and the target effect. After all, there has to be *some* change in the value of that gerrymandered variable in order for there to be a change in the occurrence of the target effect. And changes in the value of that gerrymandered variable that correspond to a change in the target effect do not require a change in the plausible cause variable.¹² In order for the analysis in terms of screening off patterns to work, there have to be some restrictions on variables that prevent gerrymandering. I postpone a discussion of gerrymandered variables until Section 15.2, because such variables pose a problem for interventionism in general, and thus deserve a separate treatment.

With these two disclaimers in place, we can turn to the upside of my proposed analysis. The asymmetric screening off pattern still provides a convenient similarity between spurious higher-level causation and standard cases of confounding by a common cause. Consider again the case of smoking, tar-stained teeth and lung cancer. Plausibly, the correlation between tar-stained teeth and the risk of incurring lung cancer is asymmetrically screened off by smoking behaviour: if one holds smoking behaviour fixed, correlations between changes in tar-stained teeth and changes in the risk of incurring lung cancer disappears, if one holds fixed the tar-stainedness of the teeth, cor-

¹²For example, if Sophie is presented with a red pebble, but shot in the neck before she gets a chance to peck, she will not peck.

relations between changes in smoking behaviour and changes in the risk of incurring lung cancer do *not* disappear. Based on these correlation patterns, we conclude that smoking behaviour confounds the correlation between tar-stained teeth and lung cancer. The correlation between the latter two is just a symptom of the causal relation between smoking and lung cancer. I propose a similar analysis for cases of spurious higher-level causation: the metaphysical base confounds the correlation between the metaphysically necessitated phenomenon and the target phenomenon.

We can summarize the problem of spurious higher-level causation as follows. Minimal interventionism controls for confounders in the form of common causes, but not in the form of metaphysical bases. It did so because controlling for metaphysical bases in the same way as one does for confounders in the form of common causes is inherently problematic and results in causal drainage. A closer look indicated that metaphysical bases can in fact threaten the causal status of the phenomena they metaphysically necessitate. They can do so in the same way that common causes can, namely by asymmetrically screening off the correlation between those phenomena and the target effect (or by metaphysically necessitating another phenomenon that asymmetrically screens off the correlation between those phenomena and the target effect). In such cases, it appears that the metaphysical base (or the threatening phenomenon it gives rise to) *does the actual causing*, whereas the metaphysically necessitated phenomenon appears to be a mere symptom of this underlying causal relation. Consequently, the correlation between the metaphysically necessitated phenomenon and the target effect is confounded by the metaphysical base (or another phenomenon that is metaphysically necessitated by that base) in such cases. Minimal interventionism is not adapted to this possibility and consequently requires an adjustment.

Of course, the proposed adjustment cannot be to give metaphysical bases the standard treatment for potential confounders. Subjecting them to the 'holding fixed'-requirements will still result in causal drainage. We thus require another method to control for confounders that are not common causes.

11.3 The robustness solution

In this section I propose my ‘robustness’ solution to the spurious higher-level causation problem for minimal interventionism and briefly compare it to earlier proposals by Woodward (2008, 2015) and List and Menzies (2009). I argue that, while my solution is similar in spirit to these earlier proposals, it is more precise than Woodward’s and avoids some of the criticism that is levied against List and Menzies (2009). I conclude that solutions to the spurious higher-level causation problem will have to track the robustness of the relevant correlations.

As we have seen, minimal interventionism takes correlations between two variables X and Y to be causal if these correlations persist in scenarios that meet the conditions contained in (M*) and (IV*). That is to say, if a correlation persists when one holds fixed at some value all other potential causes of Y that are not tightly related to X or Y , then that correlation is causal. The problem of spurious higher-level causation arises because some correlations that meet these criteria do not strike us as causal. In particular, those correlations that are asymmetrically screened off by the metaphysical base (or some phenomenon metaphysically necessitated by that base), do not strike us as causal.

One can solve the spurious higher-level causation problem by imposing extra requirements on those correlations that we count as causal. In particular, one can solve the problem by requiring that such correlations are ‘robust’ in the following sense:¹³

Robustness If there is a correlation between variable A and variable B , and variable C is a base variable of A , then the correlation between A and B is robust relative to C , if and only if, the correlation between A and B is *not* asymmetrically screened off by (i) C or (ii) any variable metaphysically necessitated by C .

¹³Terms like ‘robust’ and ‘robustness’ are frequently used in the causation literature in ways that differ subtly from mine. For example, Usher (forthcoming) uses them to denote insensitivity to changes in the background conditions — a property which I will call ‘stability’ in Chapter 13. Overall, terms denoting insensitivity, such as ‘stability’ and ‘robustness’, are often used interchangeably. My use of those terms aims to make distinctions of a finer grain.

By ‘base variable’ I mean either a metaphysical base variable or a nomic base variable, i.e. a variable that underlies A and the values of which either metaphysically or nomically necessitate the values of A . As we shall see, robustness relative to nomic bases will play an important role in our overall project. Given that we are concerned with metaphysically necessitated causes here, we can focus on robustness relative to metaphysical bases in this section.

Now we can adjust our definition of causation accordingly. We can do so by simply adding a robustness requirement to (M^*) and keeping (IV^*) as it was. Call the adjusted definition of causation (M^{**}) :

(M^{})** X is a type-level direct cause of Y with respect to a variable set V if and only if

- (i) there is a possible intervention on X that will change Y or the probability distribution of Y when one holds fixed at some value all other variables Z_i in V that do not stand in a tight relation to X or Y .
- (ii) the correlation between X and Y under such interventions is robust relative to X ’s base variables.

(i) avoids treating correlations confounded by common causes as causal. (ii) avoids treating correlations confounded by metaphysical bases of the investigated variable as causal. The resulting account of causation avoids spurious higher-level causation.

Before discussing the consequences of this robustness proposal for our overall project, I briefly compare it to two other proposed solutions to spurious higher-level causation from the literature on higher-level causation. I discuss two proposed treatments of the problem: one by Woodward (2008, 2015) and one by List and Menzies (2009). I argue that my robustness proposal captures the idea underlying both of these treatments and avoids some of the objections raised against the proposal by List and Menzies.

Woodward (2015, p. 320–321) briefly considers the problem of spurious higher-level causation. He introduces a problem case that is similar to the three problem cases we considered above and concludes that there must be a “a relatively stable pattern of non-zero correlation” between (metaphysically necessitated) causes and their effects. He does not provide any further expla-

nation of what he means by a ‘stable pattern’, nor does he adjust his definition of causation to cope with such cases. However, there is an intuitive sense in which the correlation between the metaphysically necessitated phenomenon and the target effect is unstable in spurious higher-level causation cases: it breaks down under natural changes in the metaphysical base. For example, Sarah will no longer peck if the pebble were any other shade of red, whereas Sophie would. In other texts, Woodward uses the term ‘stable’ applied to correlations to mean that these correlations do not break down under natural changes in background conditions (e.g. Woodward, 2006, p. 13–14). However, his specific examples when talking about higher-level causation suggests that what he has in mind is resistance to natural changes in metaphysical bases when using the term in that context.¹⁴

I take robustness to capture this kind of insensitivity to natural changes in the metaphysical base. Suppose that the correlation between a metaphysically necessitated phenomenon A and a target effect E is robust and thus *not* asymmetrically screened off by metaphysical base variable B . This means that, if there are changes in B that are compatible with keeping A fixed, the correlation between A and E will not break down under these changes. Consequently, the fact that certain values of A are followed by certain values of E is not contingent on B having any particular value. Conversely, suppose that the correlation between A and target effect E is not robust and thus *is* asymmetrically screened off by A ’s metaphysical base variable B . This means that the fact of particular values of A resulting in particular values of E is contingent on these values of A being realized by a particular value of B . Once B is fixed at any particular value, there is no correlation between changes in A and changes in E , and if A is fixed at a particular value, the value of E is still dependent on the value of B . In other words, if a correlation is not robust, particular values of the metaphysically necessitated variable A ’s being followed by particular values of the target effect E is sensitive to the value of metaphysical base variable B .

List and Menzies (2009) provide a similar proposal. They argue that higher-level causation must be *realization-insensitive*:¹⁵ the presence of the

¹⁴See also Woodward (2008, p. 242–243), where he emphasizes the importance of resistance to natural changes in metaphysical bases in cases of higher-level causation.

¹⁵It is worth noting that their discussion is restricted to cases of metaphysical necessita-

metaphysically necessitated phenomenon resulting in the target effect should not be dependent on the presence of the former's actual metaphysical base. For example, the redness of a pebble can only be the cause of a pigeon pecking if the pigeon would still have pecked in a variety of scenarios where the redness of the pebble had a different metaphysical base. Its redness does *not* count as a cause if the pigeon would not peck in a variety of scenarios where the redness has a different metaphysical base. Or in the terminology used above: the metaphysical base should not asymmetrically screen off the correlation between the metaphysically necessitated phenomenon and the target effect.

However, List and Menzies's proposal also bears some relevant dissimilarities to mine. First of all, they support their proposal with a counterfactual test for causation, whereas my proposal relies on the interventionist account and correlation patterns.¹⁶ Second, they propose to use their counterfactual test to exclude some metaphysical bases as well as some metaphysically necessitated phenomena. I will not discuss List and Menzies's counterfactual test here,¹⁷ but it is worth briefly elaborating on their proposal to exclude metaphysical bases in favour of the phenomena they metaphysically necessitate — if only to distance my proposal from such a commitment.

List and Menzies claim that metaphysically necessitated causes causally exclude their metaphysical bases.¹⁸ They maintain that, in the good cases, metaphysically necessitated causes are better difference-makers for their effect. In Sophie's case for example, it seems reasonable to say that if a red pebble had not been red, she would not have pecked at it. However, it is less certain that if an auburn pebble had not been auburn, Sophie would not have pecked at it. After all, the pebble could have been crimson rather than auburn in this counterfactual scenario. In such a scenario, Sophie would still have pecked at it. Consequently, redness is a better difference-maker than

tion. It is therefore unclear whether they take 'realization' to include the relation between core realizers and realized properties as discussed in Chapter 5.

¹⁶Nonetheless, they express that they are sympathetic to interventionist accounts of causation (List and Menzies, 2009, p. 481).

¹⁷Some have argued that it requires the problematic assumption that the metaphysically necessitated phenomenon is still present in the closest possible world where the base is absent (e.g. Woodward, 2015, fn. 1). See also Williamson (2000, Ch.3) for reasons to prefer correlation across different scenarios over counterfactual dependence as a guide to causal relevance. I will not pursue this line of criticism here. The evaluation of counterfactuals and their relation to causation are tricky matters that would take us far afield, and we have our hands full with the interventionist account of causation.

¹⁸Zhong (2019) uses a similar line of reasoning to deny the causal closure of the physical.

any of its metaphysical bases in Sophie's case. In terms of the correlation patterns involved, we can say that the metaphysically necessitated variable screens off the correlation between the metaphysical base and the target effect. According to List and Menzies, this means that the metaphysically necessitated variable is a cause, but the metaphysical base is not.

Many have objected that List and Menzies's proposal to causally exclude metaphysical bases clashes with our causal judgments (e.g. McDonnell, 2017; Shapiro and Sober, 2012). Their reasoning goes as follows. First, they note that we readily accept causal claims such as (McDonnell, 2017, p. 1468):

- (1) Socrates's drinking hemlock caused him to die.

Second, they note List and Menzies's proposal cannot accommodate such causal claims. In the case of (1), for example, there is a phenomenon that is metaphysically necessitated by Socrates's drinking hemlock *and* screens off the correlation between his drinking hemlock and his dying (roughly) when he did: Socrates's drinking poison. After all, if Socrates had drunk poison that was not hemlock, he would still have died (roughly) when he did. Even so, we would not want to say that (1) is false just because there is a phenomenon that is metaphysically necessitated by Socrates's drinking hemlock and that screens off the relevant correlation. Objectors conclude that List and Menzies's account of causation does not provide the right results, and that it cannot be the right way to secure higher-level causation or (non-reductionist) mental causation (McDonnell, 2017).

Our robustness proposal does not encounter this kind of problem. After all, *Robustness* does *not* allow metaphysically necessitated variables to exclude their metaphysical bases from being causes of certain effects. Even if a metaphysically necessitated variable would asymmetrically screen off the correlation between its metaphysical base variable and the target effect, this does not exclude the metaphysical base from causing that effect. Or at least, not according to *Robustness*. Consequently, (1) is true according to (M**), which thus avoids a central objection raised against List and Menzies proposed treatment of the spurious higher-level causation problem.

We can conclude that the difference between robust correlations and non-robust correlations marks the relevant distinction between spurious cases of

higher-level causation and good cases of higher-level causation. This is not to say that my robustness proposal is the best solution to the problem. Perhaps there are solutions that are simpler or more general in that they solve several problems about causation all at once. However, these alternative solutions had better track the difference between robust and non-robust correlations if they are to solve the problem of spurious higher-level causation.

11.4 The upshot

We now have an analysis of, and a solution to, the problem of spurious higher-level causation. It is worth taking some time considering the importance of these findings for our overall project.

We found that metaphysical bases can in fact act as confounders for the relevant higher-level correlations by asymmetrically screening off these correlations. It is only when the higher-level correlation is robust that it qualifies as causal. We adjusted our definition of causation accordingly and on the resulting picture metaphysically necessitated phenomena can be causes in virtue of exhibiting robust patterns of correlations with their target effects. The need for a robustness requirement on causation raises a question that is of particular interest for both the interventionist and the dualist. If what matters for higher-level causation is exhibiting a robust pattern of correlation, then why should nomically necessitated phenomena, such as dualist pain, be excluded from being causes by their nomic bases? Put differently: if metaphysically necessitated phenomena can be causes in virtue of exhibiting robust patterns of correlations with their target effects, then why shouldn't nomically necessitated phenomena be allowed to be causes by exhibiting robust patterns of correlation with target effects?

To make the questions more tangible, consider the following example. Suppose that we are considering a purported case of mental causation: my pain is necessitated by phys, and phys causes me to wince. We find that the correlation between pain and wince is indeed robust relative to phys. However, we are not in a position to know whether the necessitation relation is of metaphysical or nomic strength. The situation, as far as we know it, is represented in Figure 11.9. It is hard to see of what use further information about

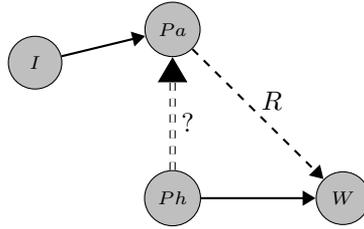


Figure 11.9: Pain (*Pa*) is necessitated by Phys (*Ph*) with undisclosed modal strength; there is a robust correlation (*R*) between *Pa* and my wincing (*W*)

the modal strength of the necessitation relation would be. Given the robustness of the pain-wince correlation, we know that pain is not asymmetrically screened off from wince. That is to say, the pain-wince correlation persists across changes in the pain's base. This pattern of correlation is therefore similar to higher-level causal correlations, such as the correlation between redness and Sophie's pecking, and markedly different from confounded correlations, such as the correlation between tar-stained teeth and lung cancer risk or the correlation between red objects and Sarah's pecking. Given that the interventionist proposes to analyze causation in terms of patterns of correlation, it is hard to see how she can motivate attaching any further importance to the modal strength of the necessitation relation.

This line of reasoning is particularly pressing given the rationale driving interventionism. The interventionist account is not only advertised as providing a metaphysically lightweight account of causation (e.g. Woodward, 2015, p. 312–313); one of the motivations for its development and the proposed adjustments is to isolate those correlations that are exploitable for reliable prediction, manipulation and control (e.g. Woodward, 2007, p. 76–77). In particular, the rationale for the 'holding fixed'-requirements is exactly to isolate such correlations, and adjustments to these requirements are motivated by that goal. Woodward states (2014, p. 710):

In other words, the role of or rationale for such control for off-path variables is to enable us to distinguish cases in which a correlational relation can be reliably exploited for purposes of manipulation and those cases in which it cannot and to prevent us from

being misled, by the presence of confounders, into thinking that a correlational relation can be so used when it cannot. In this sense there is an obvious functional justification for controlling for off-path or potentially confounding variables.

This rationale motivates a robustness requirement on causation. *Ceteris paribus*, robust correlations can be more reliably exploited for the purposes of manipulation than non-robust correlations. In the case of Sophie, for example, we can reliably make her peck by presenting her with red objects. In the case of Sarah, we cannot.¹⁹ Plausibly, these observations generalize to other higher-level correlations. As Woodward notes, this rationale does not motivate applying ‘holding fixed’-requirements on metaphysical bases, as this would exclude many correlations that can be reliably exploited for manipulation from being counted as causal. However, it does not motivate applying the ‘holding fixed’-requirements to nomic bases either. For example, figuring out the modal force of the necessitation relation in scenarios such as the one sketched in Figure 11.9 will not provide us with any extra information about the availability for manipulation or control of the relevant correlation.

Based on these considerations, it appears that allowing for causation by nomically necessitated phenomena is more in keeping with the spirit of interventionism than excluding it.²⁰ This means that standard interventionist models such as (M**) and the models provided by List and Menzies (2009) and Woodward (2008, 2015) go beyond the spirit of interventionism. That is to say, these models impose more requirements on correlations than their just being available for reliable manipulation and control, thereby suggesting that there is more to causation than just patterns of correlations that are thus available for manipulation and control. In particular, standard interventionist models *de facto* demand that causes are tightly related to those phenomena that are physically sufficient for their target effects (cf. Chapters 4 and 10). The result is an account of causation that is lightweight, but not *thoroughly* lightweight: it does not demand that causes produce or physically necessitate their effects, but it does demand that causes are tightly related

¹⁹Unless all the red objects available to us are also scarlet; we should be so lucky.

²⁰Or at least, it accords with the spirit of interventionism in as far as Woodward (2014) reliably represents that spirit.

to the phenomena that physically necessitate those effects.²¹ However, if one focuses on the original motivation for the ‘holding fixed’-requirements, a thoroughly lightweight account is more appropriate. The above quotation suggests that we should treat correlations that are available for manipulation and control *regardless* of how intimate the relation between the purported cause and the relevant physically sufficient phenomena turn out to be. If we were to follow this suggestion, we should *not* impose the ‘holding fixed’-requirements on nomic bases. Instead, one should exempt nomic bases as well as metaphysical bases and impose a robustness requirement on causation. It seems to be worthwhile for the interventionist to investigate whether or not a model of causation that is more in line with the rationale driving her account can credibly be defended.

In the next chapter, I will develop an interventionist account of causation that follows the above line of thought. As opposed to standard interventionist models such as (M**) and the models provided by List and Menzies (2009) and Woodward (2008, 2015), the model I propose does not attach any importance to the modal force of the necessitation relation when evaluating purported cases of higher-level causation. On this account, pain is a cause of wince in the case represented in Figure 11.9 *regardless* of the modal strength of the vertical necessitation relation. It would be a significant step forward for the dualist if such an account of causation can indeed be credibly defended.

²¹On the assumption that these physically sufficient conditions also produce their effects, this means that causes should also be tightly related to phenomena that produce their effects (cf. Ney (2009, 2012) and Chapter 4).

Chapter 12

Interventionism for Dualists

In this chapter I present a first approximation of my interventionist model of dualist mental causation, which I dub *insensitive interventionism*. First, I formulate the central definitions of insensitive interventionism (IM), *Insensitive Token Causation*, and (IIV). Second, I show how insensitive interventionism allows for dualist mental causation in worlds where *Physical Completeness* is true and briefly discuss under which conditions there is dualist mental causation in physically complete worlds according to insensitive interventionism. I then show how insensitive interventionism also provides a principled answer to both the exclusion worry and the common cause worry. I conclude that, in as far as insensitive interventionism provides adequate sufficient conditions for causation, it is plausible that there is mental causation in the actual world even if dualism is true. I then give a brief preview of the following three chapters, in which I provide further support for those sufficiency conditions. In particular, I will argue that they are supported by recent developments in debates on causation by absences and the relation between causation and physics.

12.1 Insensitive interventionism

As we have seen, standard interventionist definitions of causation, such as (M**) do not allow for dualist mental causation in physically complete worlds. This is because they impose ‘holding fixed’-requirements on nomic bases. If we take (M**) as a starting point, we can motivate exempting nomic bases from these requirements and the resulting model of causation straightforwardly allows for dualist mental causation in worlds where *Physical Completeness* holds. Or so I will argue. Let us first look at the account of causation itself, which I will call ‘insensitive interventionism’.

We can start with the definition of type-causation, call it (IM):

- (IM) X is a type-level direct cause of Y with respect to a variable set V if and only if
- (i) there is a possible intervention on X that will change Y or the probability distribution of Y when one holds fixed at some value all other variables Z_i in V that do not stand in a synchronic nomic necessitation relation or a tight relation to X or Y .
 - (ii) the correlation between X and Y under such interventions is robust relative to X ’s base variables.

As with (M), (M*), and (M**), (IM) is a definition of direct type-level causation. However, definitions of both indirect causation and token causation follow naturally from this definition. I will provide a definition of indirect causation, but it is well worth it to work out a definition of token causation based on (IM). This will allow us to evaluate token causal claims, such as the claim that my pain causes my wincing. We can define insensitive interventionist token causation as follows:

Insensitive Token Causation X taking value x is a direct token cause of Y taking value y relative to variable set V if and only if, X meets (IM) relative to Y and V , and there is a possible intervention on x that would be followed by a change in y .

For example, my actual pain is a cause of my actual wincing, if and only if, there is an intervention on my pain that results in a change in my wincing, and the correlation between my pain and my wincing is robust. That is to say, the correlation between my pain and my wincing is not asymmetrically screened off by the base — be it nomic or metaphysical — of my pain.

We can now define the notion of an intervention variable as follows:

I is an intervention variable for *X* with respect to *Y* if and only if *I* meets the following conditions:

(IIV)

I1 *I* causes *X*.

I-I2 *I* acts as a switch for all other variables that cause *X*. That is, certain values of *I* are such that when *I* attains those values, *X* ceases to depend on the values of other variables that cause *X* (*but do not stand in synchronic nomic necessitation relation to X*) and instead depends only on the value taken by *I*.

I-I3 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, for those causes of *Y*, if any, that either *stand in a synchronic nomic necessitation relation or tight relation to X or Y*, or 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*.

I-I4 *I* is (statistically) independent of any variable *Z* that (a) *does not stand in a synchronic nomic necessitation relation or tight relation to X or Y*, (b) causes *Y* and (c) is on a directed path that does not go through *X*.

(IIV) is just an adaptation of (IV) with the exemption clause for tight relations extended to include synchronic nomic necessitation relations as well.

(IM), (IIV) and *Insensitive Token Causation* together provide a first approximation of an account of causation that I will call *insensitive interventionism* from here on. Inensitive interventionism aims to pick out as causal those correlations that are insensitive, in the sense that they persist across a variety of scenarios. In particular, it picks out a correlation between changes in X and changes in Y as causal if and only if that correlation persists in scenarios where one changes the value of X and the causes of Y that do not stand in a tight relation or nomic necessitation relation to X (or any other potential cause of Y) are held fixed at a certain value, as per (i), and in scenarios where there are variations in X 's base variable, as per (ii). Condition (i) serves to avoid treating correlations that are confounded by variables that do not stand in tight relation or nomic necessitation relation to the purported cause variable as causal. Condition (ii) serves to avoid treating correlations that are confounded by variables that do stand in a nomic necessitation relation or tight relation to the purported cause variable as causal. In light of the observations in the previous chapters, I take insensitive interventionism to do what interventionist accounts were supposed to do: pick out as causal those correlations that are available for manipulation and control.

As we shall see in the following chapters, the current formulation of insensitive interventionism will require some adjustments to deal with counterexamples. However, this first approximation will suffice to illustrate how a model of causation along these lines can allow for dualist mental causation. By exempting variables that stand in a mere nomic necessitation relation to the purported cause from the 'holding fixed'-requirements, the model provides a *thoroughly* lightweight account of causation in the sense outlined in Chapter 4: it does not require that causes stand in tight relations to the phenomena that physically necessitate or produce their target effects. As we shall see in the next section, this thorough lightness will make insensitive interventionism amenable to dualist mental causation. It then remains to be argued that insensitive interventionism, or some adjusted variety that is equally lightweight, can in fact provide a plausible account of causation.

12.2 Dualist mental causation

According to insensitive interventionism, it is plausible that dualist mental phenomena cause behavioural phenomena. Consider again the example of my pain and my wincing. My being in pain (Pa) is nomically necessitated by some physical condition (Ph), and I subsequently wince (Wi). Let us start with a simple version of this case and assume that all three variables only have an ‘on’ value and an ‘off’ value. Assume further that they are all turned on in the actual scenario: I am in pain, phys occurs, and I subsequently wince. The relevant question is whether or not my pain is a token cause of my wincing.

According to insensitive interventionism, the answer is likely to be yes. Pa having its ‘on’ value, is likely to come out as a cause of Wi having its ‘on’ value, because the correlation between Pa and Wi meets both condition (i) and condition (ii) imposed by insensitive interventionism. The causal relations between these variables are likely to be as is represented in Figure 12.1. Let us look at condition (i) and condition (ii) in turn.

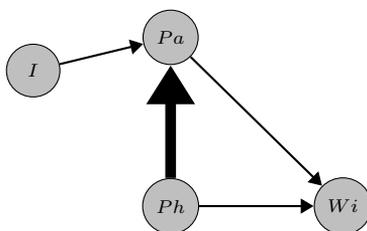


Figure 12.1: The thick full arrow represents the nomic necessitation relation between Pain (Pa) and (Ph); both of which cause my wincing (Wi)

First, it is plausible that there are logically possible manipulations of my pain that meet (IIV) and result in changes in my wincing. In particular, there are manipulations that stop me from wincing by stopping me from being in pain. Such manipulations would also affect the nomic base of my pain, but (IIV) allows for that. *Prima facie*, my pain meets (i) relative to my wincing.

Second, the correlation between my pain and my wincing appears to be robust. That is to say, the correlation between my pain and my wincing is not

asymmetrically screened off by its nomic base. After all, if I hold fixed Ph at its ‘off’ value, there are still changes in Pa that correlate with changes in Wi . For example, changing the value of Pa from ‘off’ to ‘on’ will still correlate with changes Wi , because there are pains that are not nomically necessitated by phys and that are typically followed by my wincing.¹ Consequently, Ph does not screen off the correlation between Pa and Wi and *a fortiori* does not asymmetrically screen off the correlation between Ph and Wi . *Prima facie*, my pain meets (ii) relative to my wincing.

There are two further remarks worth making on (ii) for cases of dualist mental causation. First, the above reasoning is contingent on pain actually being multiply realizable by physical nomic bases and pain’s being followed by wince not being contingent on just any one of those nomic bases. It has to be the case that my being in pain results in my wincing across variations in the physical nomic bases of my pain. I am not in a position to claim that such a correlation pattern actually exists. However, the existence of such patterns is taken to be plausible in mental causation debates, and plays an essential role in establishing mental causation for non-reductionist physicalists (e.g. Yablo, 1992; Campbell, 2008, 2010; Woodward, 2008, 2015). I see no reason why the dualist should not be allowed to assume the same correlation patterns.

Second, the situation becomes less straightforward once we start considering more complex variables. Consider a scenario that is structurally the same as in Figure 12.1, but in which Ph can take the following four values: $\{phys_1, phys_2, phys_3, phys_4\}$. $phys_1$ and $phys_2$ are physical phenomena that nomically necessitate pain and result in wincing. $phys_3$ and $phys_4$ are physical phenomena that nomically necessitate a feeling of bliss and do not result in wincing. In this scenario, Ph plainly screens off the correlation between Pa and Wi . Once one fixes Ph at any of its values, there is no more correlation between changes in Pa and changes in Wi . Pa would maintain its robustness relative to Wi if it were to screen off the correlation between changes in Ph and changes in Wi , but it is not immediately clear that it does. If we consider nomically impossible scenarios, Pa clearly does not screen off the correlation between Ph and Wi . If we hold Pa fixed at any of its values, we can still change freely between the values of Ph by breaking the nomic ne-

¹As the reader will have noticed, the case is structurally the same as Sophie’s case with simple variables in Chapter 11.

cessitation between its values and the value of Ph . The correlation between such changes in Ph and changes in Wi is likely to persist. By contrast, if we ignore nomically impossible worlds, Pa screens off the correlation between Ph and Wi , and therefore maintains its robustness. We can say that the correlation between Pa and Wi is robust across nomically possible worlds, but not across nomically impossible worlds.

This second remark might be taken to raise a worry for dualist mental causation. All the good cases of higher-level causation we discussed in the previous chapter *are* robust across nomically impossible worlds. Why should we take peace with robustness across nomically possible worlds? I think there are two things the dualist should say in response to this worry. First, depending on how the mental-physical correlations turn out to be in the actual world, the restriction to nomically possible worlds might not be required. Second, such a restriction to nomically possible scenarios need not be problematic. It is in fact not uncommon to restrict one's the scope to nomically possible scenarios in accounts of causation, and such a restriction is likely to be less restrictive than one would expect (cf. Albert, 2000, 2015; Dorr, 2016; Loewer, 2007a, 2008, 2012). I discuss both responses in turn.

First, take a look at the variable we have used to formulate this worry. We have not qualified the four values of Ph any further than their representing physical phenomena and their relation to the values of Pa and Wi . Indeed, we reverse-engineered these values such that we could construct a screen variable for the Pa - Wi correlation. As we have briefly discussed in the previous chapter, it is easy to reverse-engineer such a screen variable for almost any correlation. However, there is no saying whether or not there is a non-gerrymandered variable that can play this role. Or at least, it is an empirical question whether or not there is a type of physical phenomenon that can play the role of Ph as spelled out in our example and I am in no position to assess the plausibility of there being any such phenomenon. Even so, it is worth noting that it is a non-trivial commitment to state that there are such non-gerrymandered screen variables for all or even most mental-physical correlations. It might still be the case that no non-gerrymandered screen variables can be found for mental-physical correlations, and there consequently is no need for mental variables to screen off the correlations between such screen

variables and their target effects. It would thus not matter whether they would do this screening off only across nomically possible scenarios rather than across nomically impossible scenarios as well.

Moreover, note that the threat of potential physical screen variables is not unique to dualism. The non-reductionist physicalist too should worry that there are no non-gerrymandered screen variables for mental-physical correlations that are not in turn screened off by the relevant mental variable. Suppose for example that there is a non-gerrymandered Ph variable with four values. $phys_1$ and $phys_2$ metaphysically necessitate pain, whereas $phys_3$ and $phys_4$ metaphysically necessitate a state of bliss. If it turns out that $phys_1$ and $phys_3$ are typically followed by wincing, and $phys_2$ and $phys_4$ are typically not followed by wincing, the pain-wincing correlation is not robust across nomically impossible worlds either. If there is such a variable, the causal status of pain relative to wincing would be in peril for the non-reductionist physicalist as well.² As mentioned, we will address issues surrounding variable selection and gerrymandered variables in the upcoming chapters. It is worth bearing in mind that these issues interact with the issues surrounding robustness.

Second, it is unclear why robustness across nomically possible worlds will not do. Insensitive interventionism aims to pick out the correlations that are available for reliable prediction, manipulation and control as causal. With that purpose in mind, it is justified to ignore nomically impossible worlds when evaluating robustness.³ If the correlation between Pa and Wi is robust across nomically possible worlds, we can predict and control occurrences of my wincing by manipulating my pain. It is *not* the case that the correlation between Pa and Wi will suddenly disappear if I manipulate Pa in a different way than before, as would be the case for the correlation between tar-stained teeth and lung cancer (cf. *infra*). We can safely ignore nomically impossible scenarios when assessing robustness. In fact, restrictions on the scenarios we consider when assessing the insensitivity of a correlation are quite common. We will discuss some such restrictions in the next chapter as well. Here I

²See also Woodward (2008, p. 259–261).

³One could protest that such nomically impossible worlds play an indispensable role in distinguishing correlations due to common causes from correlations that are causal. This is a version of the common cause worry, which I address soon.

would just like to address the restriction to nomically possible scenarios.

Notable accounts of causation and counterfactuals require us to ignore nomically impossible scenarios. According to the thermodynamical accounts of causation and counterfactuals, proposed by Albert (2000, 2015) and Loewer (2007a, 2008, 2012), we should restrict ourselves to nomically possible scenarios when using counterfactual scenarios to evaluate causal or counterfactual dependence. In a similar vein, Dorr (2016) argues that we should not rely on nomically impossible scenarios when evaluating counterfactuals, but should instead consider scenarios where the initial conditions were different, such that they evolved according to the nomic laws in the counterfactual scenario. For example, to evaluate ‘if I were left-handed, I would not use these scissors’, we should not consider a scenario where I turn out to be left handed because the nomic laws are different. Instead, we should consider a scenario in which the initial conditions of the universe are different in such a way that they, following the same nomic laws, evolve into me being left-handed. Or so Albert, Dorr and Loewer argue. I will not elaborate on their arguments here, but one point that is highlighted by these authors is worth illuminating: more is nomically possible than one would expect.

The laws that we are asked to obey in considering counterfactual scenarios allow for scenarios that we intuitively feel to be counternomic. For example, consider the following scenarios provided by Albert (2015, p. 1)

Suppose the world consisted entirely of point masses, moving in perfect accord with the Newtonian law of motion [...]. And consider a rock, traveling at constant velocity, through an otherwise empty infinite space [...]. And note that nothing whatsoever in the Newtonian law of motion, together with the laws of the interparticle forces, together with the stipulation to the effect that those interparticle forces are all the forces there are, is going to stand in the way of that rock’s suddenly ejecting one of its trillions of elementary particulate constituents at enormous speed and careening off in an altogether different direction, or (for that matter) spontaneously disassembling itself into statuettes of the British royal family, or (come to think of it) reciting the Gettysburg Address.

A rock that spontaneously transforms into something that recites the Gettysburg Address certainly strikes us as nomically impossible. However, it appears that it is allowed by at least one candidate for a set of fundamental laws: the Newtonian laws. Albert goes on to argue that such seemingly impossible scenarios are in fact possible according to any fundamental physical account of the world that is currently taken seriously (2015, p. 2–3). At most, we can say that such scenarios are vastly improbable according to such theories. Robustness across nomically possible scenarios alone therefore provides us with a robustness across more scenarios than one would expect. Scenarios containing smokeless fires, suddenly levitating trains, and honest politicians are all nomically possible, even though they break many generalizations we like to call ‘laws’. We should bear in mind that the laws we typically consider, such as economical, biological or social laws, are not *nomie* laws in the sense that we outlined in Chapter 2. Nomic laws are fundamental, in that they are not explained by any further laws or matters of fact,⁴ but rather explain further matters of facts and certain *ceteris paribus* laws. By contrast, economical, biological and social laws are exactly the kind of *ceteris paribus* laws that are (at least partially) explained by these fundamental nomic laws. It turns out that those laws which are general enough to be fundamental, are also so liberal as to allow for scenarios that we typically regard as violating the laws of nature. A restriction to scenarios that are allowed by those laws is therefore not as restrictive as one would think.

Now that we have dealt with these remarks, we can conclude that our current approximation of insensitive interventionism at least allows for dualist mental causation. It does so by exempting nomic bases from the ‘holding fixed’-conditions and instead demanding that causal correlations be robust. Whether or not this means that there actually is dualist mental causation in the actual world will depend on whether or not dualism is true and on how the mental-physical correlations turn out to be. I do not address either of these issues in this dissertation. However, I did remark that, even if these correlations are only robust across nomically possible worlds, this need not interfere with their causal status. It remains to be argued that such thoroughly lightweight accounts of causation can in fact deliver adequate suf-

⁴Unless one is a Humean about laws, in which case they are explained by further matters of fact, but not by other laws.

iciency conditions for causation. Before turning to this task, I briefly discuss how insensitive interventionism provides a response to both the exclusion worry and the common cause worry.

Exclusion worries

According to insensitive interventionism, the *Physical Exclusion*-premise of the exclusion argument is false. Insofar as insensitive interventionism is a well-motivated account of causation, it provides a well-motivated reply to the exclusion argument.

Recall the exclusion argument that threatens the dualism we are currently considering:

Nomic Dualism Mental phenomena are merely nomically necessitated by physical phenomena.

Physical Completeness For any actual physical phenomenon P and any time t , there is a purely physical phenomenon that occurs at t and physically necessitates the occurrence of P .

Physical Exclusion For any three phenomena A , B and C : if A occurs at t and physically necessitates B 's occurrence at $t + x$, no phenomenon C occurring at t that is not tightly related to A and is not tightly related to any of A 's parts is a cause of B , unless it is a case of genuine overdetermination.

Non-Overdetermination There is no systematic genuine overdetermination of physical effects with mental causes.

No Mental Causation Mental phenomena cannot systematically cause physical phenomena.

Mental Causation and *Nomic Dualism* together entail that physical phenomena are systematically caused by phenomena that are merely nomically necessitated by physical phenomena. *Physical Completeness* and *Physical Exclusion* together entail that there are no phenomena that are merely nomically necessitated by physical phenomena and cause physical phenomena, *unless*

it is a case of genuine overdetermination. Finally, *Non-Overdetermination* states that genuine overdetermination is not an option. One of these five propositions has to go.

Denying *Physical Exclusion* appears to be the only option within our project. We decided to treat *Physical Completeness* as non-negotiable at the outset of our project. Our dualist starting position is plausibly committed to *Nomic Dualism* — and even if it turns out not to be, there is an only slightly stronger exclusion principle that will affect our dualist starting position (cf. Section 7.4). The aim of our project is to secure *Mental Causation*, and denying *Non-Overdetermination* seems to require a substantial *ad hoc* postulate. We thus require a motivation to reject *Physical Exclusion* to respond to the exclusion argument.

If the sufficiency conditions comprised in insensitive interventionism are adequate, *Physical Exclusion* is false. As we have seen, dualist mental phenomena can be causes of physical effects in a world where *Physical Completeness* is true according to insensitive interventionism. On the plausible assumption that those dualist mental phenomena are merely nomically necessitated by the physical phenomena that are physically sufficient for those effects, this means that *Physical Exclusion* is false. Some effects *do* have (non-overdetermining) causes that are merely nomically necessitated by their physically sufficient conditions; namely the effects of dualist mental phenomena.

Denying *Physical Exclusion* seems radical. This should come as no surprise, as it is a direct consequence of exempting nomic bases from the ‘holding fixed’-requirements, which also seemed radical. However, as was the case with exempting nomic bases, denying *Physical Exclusion* is in line with the motivations driving interventionism: it allows us to select those correlations that are available for prediction, manipulation and control as causal.

Common cause worries

Prima facie, it is plausible that any dualist mental phenomenon and its purported effect have a cause in common; namely the nomic base of the mental phenomenon. This poses a challenge for accounts of dualist mental causation. Such accounts should allow for the correlations between dualist mental phenomena and their purported effects to be causal, without counting spurious correlations between effects of common causes as causal. I argue that insensitive interventionism addresses this challenge, because it provides a principled distinction between standard cases of correlations that are confounded by a common cause and cases of dualist mental causation.

According to insensitive interventionism, dualist mental phenomena and their purported effects do exhibit a standard common cause structure. The nomic base is plausibly a cause of the dualist mental phenomenon and its purported effect. For example, phys is plausibly a cause of my pain and of my wincing, because there are interventions on phys that result in changes in my pain and my wincing. Recall the relevant causal diagram as represented in Figure 12.2. Inensitive interventionism thus counts some correlations that

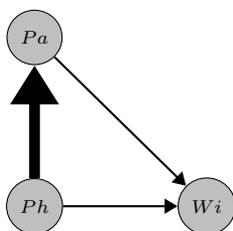


Figure 12.2: The thick full arrow represents the nomic necessitation relation between Pain (Pa) and (Ph); both of which cause my wincing (Wi)

are due to a common cause as causal.

However, insensitive interventionism does not count this correlation as causal *just* because the relata have a common cause, but because the correlation is robust. One has to bear in mind the real problem with counting all correlations due to common causes as causal: it results in counting confounded correlations as causal. Inensitive interventionism is designed to *not* count confounded correlations as causal and does deliver the right results

in standard cases of confounding like correlations between tar-stained teeth and lung cancer risk. Even though insensitive interventionism counts some correlations due to common causes as causal, it does provide a principled distinction between confounded correlations and causal relations.

The common cause worry for dualist mental causation is thereby answered. Even though the nomic bases of dualist mental phenomena plausibly cause both the dualist mental phenomena and their purported effects, this only threatens the causal status of dualist mental phenomena if the nomic base asymmetrically screens off the correlation between the dualist mental phenomena and their purported effects. This is exactly what insensitive interventionism predicts.

We can conclude the following. Insensitive interventionism provides a model of dualist mental causation in a world where *Physical Completeness* and *Nomic Dualism* holds. It provides a response to exclusion worries by motivating the rejection of *Physical Exclusion* and it provides a response to the common cause worry by drawing a principled distinction between cases of dualist mental causation and spurious correlations due to common causes. In as far as the sufficiency conditions in insensitive interventionism are adequate, it is plausible that dualist mental causation is possible in a physically complete world. It now remains to be argued that these sufficiency conditions are in fact adequate.

12.3 Looking forward

Our first approximation of insensitive interventionism provides a thoroughly lightweight account of causation. By exempting nomic bases from the ‘holding fixed’-requirements, it allows phenomena that are merely nomically necessitated by the physically sufficient phenomena for a certain effect to cause that effect. In particular, it allows mental phenomena to cause physical phenomena in a nomic naturalist dualist ontology. In the following three chapters, I defend such a model against some objections. In this section, I provide a brief summary of these chapters.

First, a remark on the scope of my defense is required. Given that I aim to provide a credible model of causation that *allows* for dualist mental cau-

sation, I will focus on defending the *sufficient* conditions provided by models like insensitive interventionism. In particular, I will defend the thorough lightweightness of these sufficient conditions. After all, it is this thorough lightweightness that renders insensitive interventionism hospitable to dualist mental causation. This means that I will allow revisions to our first approximation of insensitive interventionism in the light of the upcoming objections as long as these revisions do not interfere with the thorough lightweightness of its sufficiency conditions.

Broadly speaking, there are two ways to object to the sufficiency conditions provided by our first approximation of insensitive interventionism. First, one can object that there is an essential feature of causation that these conditions fail to capture. Second, one can argue that they systematically overascribe causation. In the following three chapters, I address objections of both kinds.

In Chapter 13, I address the objection that causation must be productive. Insensitive interventionism does not require causes to produce their effects. For example, it allows pain to cause wincing in a nomic naturalist dualist ontology. Given that pain cannot transfer physical energy on wincing, and that is how production is understood in these debates, pain does not produce wincing. Consequently, insensitive interventionism must be mistaken. So the objection goes.

I will respond to this objection by relying on cases of causation by absences. In short, I argue that the prevalence of such *negative causation* provides a powerful reply to the objection from production. However, by allowing for negative causation insensitive interventionism runs the risk of overascribing causation. I critically evaluate some of the solutions to this problem from the literature and conclude that we could adopt any of these without giving up on the thorough lightweightness of insensitive interventionism.

In Chapter 14, I address the objection that causation must be closely related to physical necessitation. More precisely, I address the objection that causes must physically necessitate their effects given background conditions. Our first approximation of insensitive interventionism does not capture this requirement. As we have seen, this model allows for dualist pain to cause wincing, but one's pain will not meet the physical necessitation requirement

relative to wincing according to nomic naturalist dualism. After all, the only background conditions relative to which pain physically necessitates wincing will be the occurrence of physical phenomena that on their own physically necessitate wincing. Thoroughly lightweight models like insensitive interventionism allow for pain to cause wincing, and must therefore be mistaken. So the objection goes.

I will respond to this objection by relying on recent work on the relation between causation and physics. In short, I argue that a closer look at this relation reveals that most familiar causes do not physically necessitate their effects given background conditions. This finding provides a powerful reply to the objection from physical necessitation. However, the discussion on causation and physics bears out a further challenge for our model of causation. There is some risk that it will spuriously ascribe *backwards* causation: it predicts that phenomena cause target effects that lie in their past. I argue that it is at least unclear that this is a problem that is specific to models like insensitive interventionism and point out that many of the solutions in the literature are compatible with a thoroughly lightweight account of causation.

In Chapter 15, I address some remaining objections that require less extensive treatments. Let us now turn to the objection from production and cases of negative causation.

Chapter 13

Negative Causation

Some philosophers maintain that lightweight accounts of causation such as (insensitive) interventionism are false because causation must be productive. Production is understood as a transfer of energy and interventionist accounts do not require causes to transfer energy on their effects. Therefore, such accounts of causation must be mistaken. So the objection goes. This is a commonplace reductionist objection to non-reductionist accounts of mental causation. For example, Kim holds that any notion of causation that does not require production is unable to underwrite our sense of agency in purported cases of mental causation (Kim, 2005, p. 17–18); (Kim, 2007, p. 235–236). If this is correct, our mental phenomena must produce our behaviour in order for our sense of agency to be accurate. Any correct account of mental causation must thus be a heavyweight account of causation, according to Kim.

In this chapter, I focus on a popular kind of counterexample to production requirements on causation: causation by absences or omissions. My discussion of such ‘negative causation’ cases can be divided in two parts. In the first part, I rehearse the arguments made by Schaffer (2004) and Russo (2016) to the extent that accounts that track our causal judgments, and in particular our causal judgments about mental causation, *must* allow for negative causation and therefore cannot impose a production requirement on causes.

In the second part, I discuss the problem of *spurious* negative causation.

Lightweight models like insensitive interventionism run the risk of allowing too many absences to be causes of a given effect. This poses a considerable challenge for our model of dualist mental causation, as we do not want our model to be too permissive. I critically survey four solutions to this problem and I argue that the correct solution relies on a particular kind of sensitivity of correlation patterns that I will dub ‘stability’. I conclude the chapter by discussing the relevance of these findings for our overall project.

13.1 Negative causation and mental causation

I take cases of negative causation to be cases where an absence causes an effect. I take absences to be phenomena whose essence is the non-occurrence of another phenomenon (cf. Beebe, 2004, p. 291). For example, *my not watering my plants* is an absence. Its essence is the non-occurrence of another phenomenon, i.e. *my watering my plants*.¹ Cases of negative causation are cases where the non-occurrence of a phenomenon causes an effect.

If negative causation is possible, there can be no production requirement on causation. Production is understood as the transfer of energy, but absences have no energy, so they cannot transfer any. Schaffer (2004, p. 204) makes the same point in terms of *persistence*. If causation requires production, there must be a line of persistence between the cause and the effect that allows for the transfer of energy. Given that absences are non-occurrences, there can be no line of persistence from (or via) absences to effects.² Given that the production requirement is a popular objection to lightweight accounts of causation, and therefore also threatens our insensitive interventionism, it would be good news for us if negative causation was indeed possible. In this section, I discuss the evidence in favour of negative causation and its relevance to mental causation debates. It will appear that the truth of a substantial subset of our causal judgments requires possibility of negative

¹This characterization is likely to have its defects, and serves mostly as a placeholder. Characterizing absences is notoriously difficult (see Bernstein (2015, section 2)), but I take the above characterization to be clear enough for our purposes.

²Schaffer refers to a passage by Fair (1979, p. 246) that makes this point. Fair goes on to acknowledge the possibility of negative causation. Dowe (2000, 2001) makes a similar point in terms of *conserved quantities* rather than energy. In contrast to Fair, Dowe does maintain that negative causation is impossible.

causation. In particular, it will appear that our judgments that mental phenomena cause behaviour require that there can be negative causation.

We can start with a familiar observation. We often talk as if negative causation is possible. Suppose I neglected to water my plants over the summer and they consequently withered. The following statement seems straightforwardly true

- (1) My not watering my plants caused them to wither.

Hence, it seems true that an absence caused an effect. Similar examples are easy to provide. Father's absence ruined the picnic. Jeffrey weeps because Clarice did not bring pudding. We often take ourselves and others to cause effects by *not* doing something.

Furthermore, cases of negative causation are not restricted to human omissions or negligence. Consider the following mechanism. The 'early bird feeder' mechanism is designed to feed those birds that are up at sunrise. It consists of saucer underneath a closed food container with a trapdoor at the bottom. The trapdoor opens when the first rays of sunlight hit a solar panel, which uses the solar energy to remove the blockage that held the trap door shut. In the evening one closes the trapdoor and fills the container with bird food. At the break of day, the sunlight causes the trapdoor to open and thereby causes the food to fall down on the saucer. The causal structure is presented in Figure 13.1.

Absences play a central role in this causal mechanism. When sunlight hits the solar panel, it causes the blockage of the lock to be removed, thereby preventing the lock from preventing the trapdoor from opening. The absence of the blockage causes the trapdoor to open, and the food falls down on the saucer. The mechanism is thus organised such that the sun rays result in the food being on the saucer via a double prevention mechanism, and hence via two absences: the absence of the blockage and the trapdoor not stopping the food from falling down. The following claim sounds true:

- (2) The sunlight caused there to be bird food on the saucer.

Like (1), the truth of (2) presupposes the possibility of negative causation.

Opponents of negative causation often propose that the sensed truth of claims like (1) and (2) is due to a confusion between causation and something

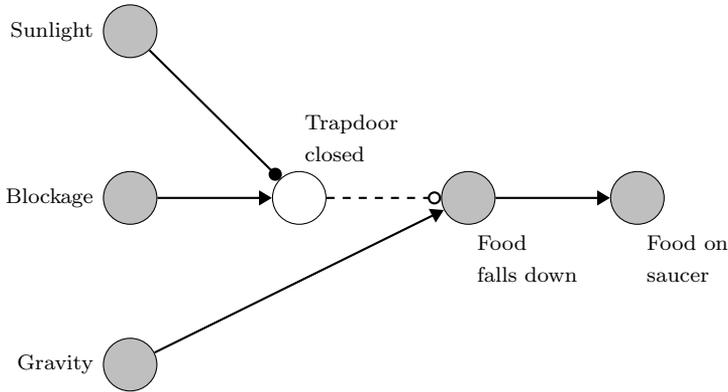


Figure 13.1: The early bird feeder. Arrows ending in a dot represent a prevention relation, empty neurons or arrowheads represent non-actualized phenomena, causings or preventions. The sunlight causes the absence of the blockages, which causes the trapdoor not to be closed. The trapdoor not being closed (in combination with gravity) causes the food to fall down.

like causal relevance (Gibb, 2013b), causal explanation (Beebe, 2004; Tang, 2015), or quasi-causation (Dowe, 2001). Even though such causation-like notions serve us well in everyday communication, planning, and prediction, they do not amount to *real* causation, because *real* causation cannot take absences as its relata. Or so the opponent of negative causation argues.

Such strategies of dealing with negative causation meet with a serious challenge. Schaffer (2000, 2004) demonstrates that many causal claims that we confidently take to be true and do not appear to involve absences in fact require the possibility of negative causation. For example, standard gun mechanisms operate via absences: pulling the trigger causes the *absence* of an obstruction to a seared-up spring uncoiling. When the spring uncoils, it compresses the gun powder. The consequent explosion will then launch the bullet. Such gun mechanisms work by a double prevention structure: pulling the trigger *prevents* a blockade from *preventing* the spring to uncoil. The relevant mechanism is represented in Figure 13.2.³ Therefore, any killing that involves a shooting with such a gun also involves negative causation.

³Cf. Schaffer (2004, p. 200)

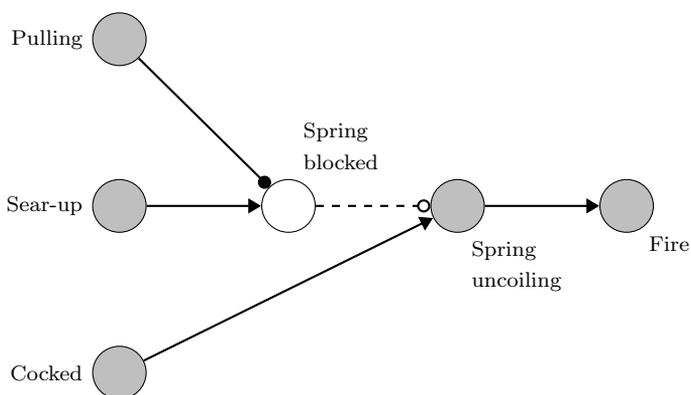


Figure 13.2: Double prevention structure underlying a gun firing

Denying the possibility of negative causation thus amounts to denying that one can cause someone's death by pulling the trigger of a gun that relies on a double prevention structure. More generally, plenty of causal relations turn out to be negative causal relations in disguise. Schaffer (2004) furnishes his argument with examples from legal systems, psychology, biology, chemistry and physics. If negative causation is impossible, then all these disciplines systematically make mistaken causal claims. Accounts that cannot allow for negative causation therefore systematically contradict our causal judgments and scientific causal claims. Schaffer concludes that no theory of causation that is so dismissive of our causal judgments deserves to be considered a theory of causation (2004, p. 205).

We can refrain from making such bold statements here. Nonetheless, the prevalence of negative causation cases provides a powerful response to the production objection against (thoroughly) lightweight accounts of causation. As it turns out, there are very good reasons *not* to impose a production requirement causation, as a substantial number of respectable causal claims require the possibility of negative causation. As examples of such cases multiply, this line of response gains in force. For now, I take it that these cases defend the credibility of insensitive interventionism from the production objection.

Even if I am mistaken in doing so, these cases are still of significant interest to the dualist. Recall that our causal judgments are the main motivation for believing *Mental Causation* (cf. Chapter 3). This means that insisting that causes must produce their effects undercuts the motivation for *Mental Causation*, and therefore relieves some of the pressure on the dualist in the mental causation debate. After all, if our causal judgments are indeed systematically mistaken, it is less clear why it tells against the dualist that she cannot allow for mental causation. The dualist can reasonably demand for an explanation why not allowing for mental causation is more objectionable than not allowing for negative causation.

In fact, negative causation cases might improve the dualist's dialectical position even further. Russo (2016) provides a simple argument for the thesis that all models of mental causation need to allow for negative causation.⁴ If this argument is sound, a production requirement on causation not only undercuts our motivation for *Mental Causation* but straightforwardly contradicts the possibility of mental-to-behavioural causation in creatures with our physical make-up. Here is a (simplified) version of Russo's argument:

1. All cases of mental-to-behavioural causation involve causation via some causal chain connecting nerve signals coming from the brain and muscle contractions.
 2. All causal chains involving a double prevention structure involve negative causation.
 3. All causal chains connecting nerve signals coming from the brain and muscle contractions involve a double prevention structure.
-
4. All cases of mental-to-behavioural causation involve negative causation.

The argument is valid. If all cases of mental-to-behavioural causation involve causation via a causal chain that involves double prevention structures, and all causal chains that involve double prevention structures involve negative

⁴See also Kroedel (2020, section 3.6) for a similar argument

causation, then all cases of mental-to-behavioural causation involve negative causation. The soundness of the argument thus falls squarely on the truth of the premises. 1 seems straightforwardly true: all our physical behaviour requires muscle contractions, consequently mental-to-behavioural causation requires nerve signals coming from the brain to cause muscle contraction. 2 seems straightforwardly true.⁵ In order to conclude that 4. is true, we only need further justification for 3.

Russo derives his justification for 3 from an example provided in Schaffer (2004). Schaffer describes the causal chain connecting nerve signals coming from the brain and muscle contractions. The chain looks as follows. The nerve signal coming from the brain prevents tropomyosin from preventing myosin-actin binding in the muscle fiber. The nerve signal thereby allows myosin to bind with the actin in the muscle fiber. This binding partially constitutes a muscle contraction. The nerve signal thus serves as a double preventer for the muscle contraction. The causal diagram takes the familiar form of a double prevention structure, as is presented in Figure 13.3.⁶ If this

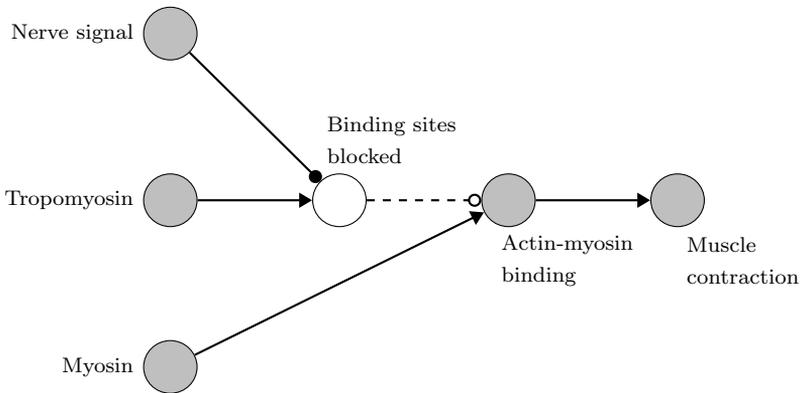


Figure 13.3: Double prevention structure underlying intentional muscle contraction

model is apt, it appears that we have a sound argument for 4. It follows

⁵Or at least, it seems straightforwardly true on the uncontroversial assumption that all steps in a causal chain are causal.

⁶Cf. Schaffer (2004, p. 200)

that any model of mental-to-behavioural causation has to allow for negative causation.

Consequently, physicalists should be careful not to rely on production requirements when arguing against dualist mental causation. True, if real causation is productive, our attempt at providing a credible model of dualist mental causation has failed. However, the dualist's position in the mental causation debate would improve considerably. If there can be no mental-to-behavioural causation for creatures like us regardless what ontology of the mind turns out to be true, the mental causation objection to dualism does not carry much force. This would be a Pyrrhic victory that the dualist should be happy to grant.

For now however, I will assume that the negative causation cases provide a convincing response to the production objection against insensitive interventionism. This means that insensitive interventionism is still in the running as a credible model of dualist mental causation. However, there is more to be said about negative causation. Allowing for absences to be causes has a flip side. Opponents of negative causation often argue that one cannot allow for absences to be causes without permitting too many causes for every effect. Given that our model of dualist mental causation allows for negative causation, it is vulnerable to this objection. In the remainder of this chapter, I spell out this problem in more detail and argue that the correct solution relies on the insensitivity of the relevant correlation patterns.

13.2 Spurious negative causation

Leniency towards negative causation comes at a cost. On the face of it, a potentially infinite number of absences will meet insensitive interventionism's requirements for any effect. In the example of my withering plants, it is not only the case that I did not water them; *nobody or nothing watered them*. The reader did not water them, my kitchen utensils did not water them and the prime minister of Belgium did not water them. For all of these absences, it is the case that if one intervenes on them such that the relevant watering *did* take place, my plants would *not* have withered. Once one allows an absence to cause an effect, it is hard to draw a line between the causally effective

absences and absences that do not cause that effect. In this section, I spell out this ‘spurious negative causation’ problem for insensitive interventionism in greater detail and critically evaluate four solutions from the literature.

We can illustrate the spurious negative causation problem with causal claims like the following:

- (1*) The Belgian prime minister not watering my plants caused them to wither.
- (2*) The Belgian prime minister not holding the trapdoor closed caused the food to land on the saucer.
- (3*) The Belgian prime minister not blocking the spring caused it to uncoil.

All three of these claims seem false and accounts of causation that allow for negative causation have a hard time explaining these data. For example, our own insensitive interventionism predicts that (1*)–(3*) are in fact true. There is a possible intervention on the Belgian prime minister not watering my plants that would have resulted their not withering; namely an intervention that makes it the case that he actually *did* water my plants. *Mutatis mutandis*, the same goes for (2*) and (3*). Productive accounts of causation spuriously treat negative causal claims as false, but insensitive interventionism spuriously treats negative causal claims as true.

The robustness requirement in insensitive interventionism does not alleviate the problem. Generally speaking, the correlations between absences and target effects are robust. For example, if we take the metaphysical base of my not watering my plants to be whatever I am doing instead, it is *not* the case that the correlation between me not watering my plants is asymmetrically screened off by its metaphysical base. It does not matter in virtue of what underlying phenomenon I do not water my plants. I could be out traveling, hospitalized, watching reruns of Seinfeld, . . . My not watering my plants robustly correlates with them withering. The same holds for the unblocked spring. It does not matter *how* the blockage is absent, its mere absence will make the spring uncoil. Typically, the correlation between absences and their target effects are not screened off or exhausted by their metaphysical bases. Unfortunately, these observations generalize to spurious cases as well: the

correlation between the Belgian prime minister not watering my plants and my plants withering is robust relative to the metaphysical base of the relevant absence, as is the correlation between his not blocking the spring and the spring uncoiling, etc. Our current interventionist model of causation thus allows the absences in (1*)–(3*) to be causes.

Arguably, this is as serious an infringement of our causal judgments as treating (1) and (2) as false. For example, Beebee (2004, p. 293) argues that, since both accepting and denying negative causation results in contradicting a substantial subset of our causal judgments, our causal judgments cannot provide convincing evidence for or against accepting negative causation. She further complicates the challenge for the proponent of negative causation by pointing out that our judgments about negative causation are sensitive to factors that are unlikely to be of causal significance. For example, she uses the following case to argue that these judgments are affected by moral norms (Beebee, 2004, p. 294):⁷

Z's dog is bitten by an insect, contracting an eye disease as a result, which Z ignores. The dog loses its sight. Intuitively, Z's negligence caused the dog's blindness.

We seem to agree that Z causes the blindness because she *ought* to have looked after her dog. After all, we do not take the failure of Z's neighbour, who does not have the moral responsibility to take care of Z's dog, to cause the blindness. Beebee adds that this judgment remains stable even if we stipulate that Z's neighbour was more likely to take the dog to the vet than Z. This sensitivity to seemingly causally irrelevant factors can be taken to indicate that one cannot provide a principled distinction between spurious cases and good cases of negative causation.

Before discussing the possible replies to this challenge, it is worth making a comment on the relevance of this problem for the dualist. Suppose that Beebee is right, and there is no approach to negative causation that comes close to respecting our causal judgments. Given the prevalence of purported

⁷In fact, sensitivity to norms has been shown to be a feature of causal judgments in general, rather than being specific to negative causal judgments (e.g. Hitchcock and Knobe, 2009; Willemsen and Kirfel, 2019). I think this finding undermines the relevance of Beebee's observation, but postpone a detailed discussion of this point to the section on what I will call the 'abnormality response' to spurious negative causation.

cases of negative causation reported by Schaffer (2000, 2004), this would mean that our causal judgments must be systematically unreliable. Once again, this would take the pressure off the dualist in the mental causation debate, as it undermines our evidence for *Mental Causation*. Consequently, it would not be disastrous for the dualist's position in the mental causation debate if there is no appropriate response to spurious negative causation.

Moreover, even if we focus on a model of dualist mental causation, rather than just defusing the mental causation objection to dualism, the dualist's demands on responses to spurious negative causation are relatively modest. Implementing the solution in insensitive interventionism should not impede with its thoroughly lightweight character. As we shall see when discussing different responses from the literature, none of these impose any heavyweight restrictions on causation. The dualist could thus adopt any of the following solutions without endangering the model of dualist mental causation we proposed in the previous chapter. Eventually, I will argue that the dualist can do better than just responding to spurious negative causation. The most promising response to the problem crucially relies on the insensitivity of the relevant correlation patterns. I take the viability of this approach to further demonstrate the importance of insensitivity for causation. Rather than merely not being endangered, our overall approach to causation would thereby be corroborated. If my arguments succeed, insensitive interventionism, and thereby the dualist, would thus be in a comfortable position. However, even if my arguments fail and my proposed solution in terms of insensitivity remains unconvincing, spurious negative causation would not put the dualist in any serious peril.

13.3 Replies to spurious negative causation

In this section, I discuss four replies to the spurious negative problem from the literature: Lewis's (2000) proposal in terms of conversational pragmatics, Schaffer's (2005) proposal in terms of salient contrast phenomena, the proposal inspired by McGrath (2005) in terms of abnormality, and Woodward's (2006) proposal in terms of the stability of the dependencies involved. While the first three strategies correctly predict that our causal judgments

are sensitive to seemingly irrelevant factors, they do not predict the correct truth values for (spurious) negative causation claims. The fourth reply by Woodward fares better with regards to predicting the correct truth values, preserves the promising aspects of the other three proposals, and respects some of the ideas driving these proposals. Or so I will argue.

Inappropriateness

Lewis (2000) denies that proponents and opponents of negative causation are equally challenged by our causal judgments. True, both positions conflict with a substantial subset of our causal judgments. But whereas the opponent has to deny seemingly true claims, the proponent of negative causation has to insist that seemingly false claims are in fact true. According to Lewis, this puts the proponent in an advantageous position, as she can maintain that the seemingly false claims are merely *inappropriate* rather than actually false. After all (Lewis, 2000, p. 196):

[t]here are ever so many reasons why it might be inappropriate to say something true. It might be irrelevant to the conversation, it might convey a false hint, it might be known already to all concerned. . .

For example, when you ask me why my plants died and I answer ‘France is a republic’, I have not said anything false, but my claim will still be deemed inappropriate. Whether or not France is a republic is *irrelevant* to the death of my plants. Lewis proposes that the seeming falsity of (1*)–(3*) could be explained by their inappropriateness.

The advantage of this explanation is that it predicts that our negative causation judgments can appear unprincipled. That is to say, it predicts that the acceptability of negative causal claims will be sensitive to factors that one would not have thought to be of causal significance. Conversational pragmatics is unpredictable territory. As Lewis mentions, there are many reasons why a claim might be inappropriate, and we are a long way from fully mapping the interactions between states of affairs, conversational contexts and appropriateness. It would come as no surprise that our moral norms, or other factors that seem irrelevant to the causal relations in question, affect

the appropriateness of negative causal claims.

The drawback of Lewis's proposal is that it treats claims like (1*)–(3*) as *true*. This conflicts with some salient data about such claims. First of all, such spurious negative causation claims strike us as all-out *false* rather than inappropriate but true. Even though claims can be inappropriate but true, we do distinguish between such merely inappropriate claims and all-out false claims. For example, we recognize that my claim that 'France is a republic' is in fact true, even though it is inappropriate in a conversation about my withering plants due to its irrelevance. By contrast, (1*)–(3*) relate directly to the topic at hand; they just happen to make a false claim about that topic. At the very least, we feel that these claims are false. But we have more than just our intuitions to go on here. Spurious negative causation claims exhibit two further tell-tale signs of being false rather than merely inappropriate.

First, as argued by McGrath (2005, p. 128–129), we are quite confident in asserting negations of spurious negative causation claims. As opposed to (1*) and (2*),

(1-neg) The Belgian prime minister not watering my plants did not cause them to wither.

(2-neg) The Belgian prime minister not blocking the spring did not cause it to uncoil.

strike us as true. Whereas,

(4*) France is not a republic.

strikes us as false. We take the denial of spurious negative causation claims to be true, this indicates that the spurious claims themselves are false (cf. DeRose, 2002).

Second, the inappropriateness of spurious negative causal claims cannot be cancelled. Sometimes, true claims are inappropriate because they suggest the truth of a further claim that is false. For example, consider a letter of recommendation that only appraises the applicant's excellent command of English and regular seminar attendance (cf. Grice, 1975). The letter writer *claims* that the applicant has an excellent command of English and attends

seminars regularly, but further *implicates* that the applicant is not a fit candidate. After all, if the letter writer considered the applicant fit for the position, she would supply more relevant information. In such cases, the inappropriateness of the claim can be ‘cancelled’ by explicitly denying the implicated claim. Such a cancellation does not strike us as contradictory and the conjunction of the irrelevant claim and the cancellation does not strike us as false. Consider the following example.

- (5) The applicant has an excellent grasp of the English language and attends seminars regularly. But I do not want to suggest that she is a bad philosopher. Her term papers are brilliant without exception and her contributions to seminar discussions are invaluable. I am confident that she will prove an asset to any philosophy department.

(5) would not strike us as contradictory or, assuming the candidate is in fact as excellent as promised, false.

If we object to spurious negative causation claims because, despite being literally true, they suggest the falsity of more relevant omission claims, we should be able to *cancel* this inappropriateness accordingly. However, ‘Both the Belgian prime minister’s negligence and my negligence to water my plants caused them to wither’ does not strike us as true any more than (1*) does.⁸ Spurious negative causation claims do not behave as inappropriate true claims, but as all-out false claims.

Based on these considerations, it appears that we cannot simply relegate the problem of spurious negative causation to conversational pragmatics. However, I would like to note here that intuitions differed among readers of the above discussion. Some are not convinced that (1*)–(3*) are false and take them to be merely irrelevant. Overall, this would be good news for insensitive interventionism. Even so, it would improve our position even further if we can provide an explanation why claims like (1*)–(3*) are systematically irrelevant and thus inappropriate, whereas (1)–(3) are acceptable. If Lewis’s

⁸There are likely to be other candidates for claims that are suggested as false by (1*), such as ‘The Belgian prime minister was responsible for the withering of the plants’ or ‘I am not to blame for the withering of my plants’. However, denying these claims does not cancel the inappropriateness either.

response to spurious negative causation indeed works, the remaining three proposals can be interpreted as providing such an explanation.⁹ For now however, we will assume that spurious negative causal claims are false. Consequently, our account of causation should predict that (1*)–(3*) are false, rather than merely inappropriate.

Contextualism

Schaffer (2005) provides a proposal that can do just that. Like Lewis, Schaffer suggests that the sensed falsity of spurious negative causation claims could be due to pragmatic considerations (2005, p. 332). Unlike Lewis, Schaffer proposes that pragmatic considerations could set the context of causal utterances such that spurious negative causal claims are in fact *false*. It might be the case, he argues, that the relevant contrasts are not contextually salient. This approach holds some promise of retaining the advantages of Lewis's response whilst avoiding its drawbacks.

Schaffer's proposal starts from the idea that causal claims make sense only against the background of contextually salient contrast phenomena for the cause as well as the effect.¹⁰ For example, (1) makes sense only because there is a contrast class for my not watering my plants as well as for my plants withering, namely my watering my plants and their thriving. Most often, these contrast classes remain unexpressed and are provided by the conversational context instead. Consequently, they can differ between speakers that utter the same sentence to refer to the same phenomena in virtue of these speakers having different background assumptions. For example, Schaffer (2012, p. 37) argues that, in a conversation between two firefighters discussing a raging forest fire, the following claim is likely to be false:¹¹

(6) The presence of oxygen caused this forest fire.

After all, firefighters do not hold the absence of oxygen around forests to be a relevant alternative possibility. That is to say, even though the presence of oxygen and the forest fire exhibit the right pattern of dependence, an absence

⁹I am grateful to Gunnar Björnsson for this suggestion.

¹⁰Schaffer assumes *events* rather than *phenomena* or *variables* as the causal relata, but the difference is irrelevant to the matters at hand.

¹¹Schaffer borrows this example from Putnam (1982, p. 150).

of oxygen surrounding the forest is not a live possibility in the firefighters's conversational context. It would be pragmatically obtuse for the firefighters to consider such an alternative. Two Venusians however, who think the presence of oxygen is quite remarkable and find its absence a salient alternative, could look at the same forest fire and utter (6) in good faith. Even though both the firefighters and the Venusians refer to the same phenomena when uttering (6) it is false when the firefighters utter it, but true when the Venusians utter it. This is because the absence of oxygen is a salient contrast phenomenon for the Venusians, but not for the firefighters.

According to the resulting picture, sentences like (6) are incomplete expressions of more complex propositions. In order to provide a more complete expression of causal propositions one needs to add a contrast class for both the purported cause and the purported effect. For example, a more complete expression of the proposition expressed by the Venusians would read:

(6-V) Oxygen being present, rather than absent, caused this forest to be on fire, rather than not being on fire.

We rarely use such complete expressions of causal propositions. The relevant contrast classes often remain unexpressed and are instead delivered by the conversational context. That is why (6) is true in the context of two Venusians discussing the forest fire, but not in the context of two firefighters discussing that same fire: in the case of the Venusians, the conversational context delivers the apt contrast classes, in the case of firefighters it does not.

Perhaps, Schaffer suggests, the context-dependence of contrast classes can help solve the spurious negative causation problem. It seems reasonable to posit that spurious negative causation claims are false because the causally relevant contrast phenomena are not contextually salient. Neither the Belgian prime minister watering my plants nor his blocking the coiled spring in a gun mechanism are contextually salient contrasts in most natural scenarios. Moreover, the relation between context and the content of propositions is notoriously difficult to systematize.¹² Consequently, Schaffer's proposal also predicts that the line between good cases of negative causation and spurious

¹²See for example Nowak (forthcoming) for an overview of the difficulties of mapping the relation between contexts and the content of demonstratives.

cases will appear to be unprincipled. For example, Z being the dog's owner is likely to set the context such that her taking the dog to the vet is a salient contrast phenomenon, whereas Z's neighbour taking the dog to the vet is not a salient contrast phenomenon, even if the latter was more likely to occur than the former and both of these would have prevented the dog from going blind. At a first pass, Schaffer's proposal delivers the right results.

However, the opponent of negative causation will have no problem making spurious negative causation reappear. Even if Schaffer's contrastive account of causation is correct, one can still decouple the truth value of causal claims from the contextual salience of the contrast phenomena. One can do so by explicitly mentioning the relevant contrast phenomenon, rather than leaving it up to the conversational context. Consequently, the following negative causation claims come out as true on Schaffer's account:

- (1*) The Belgian prime minister not watering my plants rather than watering them, caused them to wither.
- (2*) The Belgian prime minister not blocking the spring rather than blocking it, caused it to uncoil.

But this cannot be correct, as we are confident in asserting their negations in good faith:

- (1-neg') The Belgian prime minister not watering my plants rather than watering them, did not cause them to wither.
- (2-neg') The Belgian prime minister not blocking the spring rather than blocking it, did not cause it to uncoil.

Consequently, this proposal still gets some cases wrong.

Schaffer acknowledges that our willingness to assert the negation of spurious negative causal claims requires explanation (2005, fn. 10). He suggests that the relevant contrast phenomena might just be too outlandish to serve as contrast phenomenon.¹³ We naturally imagine the Belgian prime minister

¹³Perhaps I am taking my liberties with Schaffer's footnote here. He talks of the 'irrelevance' of the queen watering his flowers. Surely he cannot mean the causal irrelevance, as that would be a circular explanation. He goes on to say that we naturally imagine the queen to attend a banquet rather than watering flowers. I take it he means that the contrast phenomenon (the queen watering his flowers) is too irrelevant to really serve as a contrast phenomenon.

sitting in meetings and giving press conferences, rather than watering my plants or blocking springs in gun mechanisms. Even though this much is plausible, there are two pieces of evidence that weigh against such an explanation.

First, when I state (1*) or (1*'), I *do* immediately imagine the Belgian prime minister watering my plants, however outlandish that scenario might seem. I even acknowledge that my plants would be thriving if he in fact had watered my plants. Nevertheless, I still think that the negative causal claim in question is false.

Second, it does not take care of spurious negative causation claims where the contrast phenomena are not outlandish. For example, there seems to be nothing outlandish about Z's neighbour taking her dog to the vet, or my neighbour watering my plants. Even so, the following claims seem false:

(7*) Z's neighbour not taking her dog to the vet, rather than taking her dog to the vet, caused the dog to go blind.

(8*) My neighbour not watering my plants, rather than watering them, caused them to wither.

The proponent of negative causation (and with her, the insensitive interventionist), is therefore still in need of a response to spurious negative causation.

Abnormality

A third solution that has recently gained traction is to appeal to the abnormality of the absences involved. This solution was initially proposed by McGrath (2005) and was further supported with empirical evidence by Henne et al. (2017), Henne et al. (forthcoming), Willemsen and Reuter (2016), and Willemsen and Kirfel (2019). Whereas Beebe specifically argued that its sensitivity to norms is a sign that something is wrong with negative causation, proponents of this solution argue that this sensitivity provides a principled distinction between spurious cases and good cases: in spurious cases, the absences are (in some significant sense) normal, in good cases they are (in that sense) abnormal. These proponents are liberal about what the relevant kinds of norms are. Here is a characterization by McGrath (2005, p. 138):

The notion of the normal I have in mind is highly abstract and applies very generally: to actions, the behaviour of artifacts, and the behaviour of both biological and non-living systems. It may be illustrated by means of examples. It is normal for x to ϕ iff x is *supposed* to ϕ . People are supposed to keep their promises (it is normal for them to keep their promises); alarm clocks are supposed to ring at the set time (it is normal for them to ring at the set time); hearts are supposed to pump blood (it is normal for them to pump blood); the rain is supposed to come in April (it is normal for it to come in April); water is supposed to flow downhill (it is normal for it to flow downhill).

According to the resulting account, absences can be causes only if they deviate from the norm, where norms are understood in the liberal sense described above. That is to say, they can cause only if their corresponding presences were *supposed* to occur. I will call this approach to spurious negative causation the ‘abnormality’ approach.

This approach has several advantages. First, there is empirical evidence that our causal judgments on negative causation in fact track the difference between normal and abnormal absences. For example, Henne et al. (2017) performed experiments which indicate that we operate with an account of causation according to which “[...] unselected [i.e. spurious] omissions are not simply irrelevant, but they are denied causal status because *they do not violate a norm*” (Henne et al., 2017, p. 12). Indeed, some of our own examples seem to fit this pattern: my not watering my plants is abnormal and a cause of their withering, whereas the prime minister not watering my plants is normal and not a cause of their withering, Z’s not taking her dog to the vet is abnormal, and a cause of the dog’s blindness, Z’s neighbour not taking Z’s dog to the vet is normal, and not a cause of the dog’s blindness.

Second, there is independent evidence for the central role of norms in our causal selection procedures. For example, Hitchcock and Knobe (2009, p. 594) used the following vignette to demonstrate that we typically single out abnormal phenomena as the cause from a set of (seemingly) sufficient conditions:

The receptionist in the philosophy department keeps her desk

stocked with pens. The administrative assistants are allowed to take pens, but faculty members are supposed to buy their own.

The administrative assistants typically do take the pens. Unfortunately, so do the faculty members. The receptionist repeatedly e-mails them reminders that only administrators are allowed to take the pens.

On Monday morning, one of the administrative assistants encounters Professor Smith walking past the receptionist's desk. Both take pens. Later that day, the receptionist needs to take an important message. . . but she has a problem. There are no pens left on her desk.

The lack of pens is counterfactually dependent on both of the pen-takings. However, it seems like we take professor Smith's action to be the cause because it breaks a norm, whereas we do not take the administrator's action to be the cause because it does *not* break the norm.¹⁴ Indeed, recent empirical research indicates that the abnormality of a cause plays a more decisive role in our judgments than the cause being an absence rather than a presence (Clarke et al., 2015; Willemsen and Reuter, 2016). This indicates that our negative causation judgments and our non-negative causation judgments are not as different as they might have seemed. If, as Beebe suggests, sensitivity to norms is a reason to dismiss our causal judgments, we should dismiss our causal judgments wholesale. Which, as we have repeated time and again, is not an attractive option for the dualist's opponent.

The abnormality approach appears to be in good shape. It provides a principled distinction between good cases and spurious cases of negative causation and, in doing so, relies on a guiding principle for causal selection more generally. Consequently, it counteracts the contention that negative causation is some anomalous form of causation. However, this proposal can only be successful if all absences that are causes are abnormal and this is not the case. Normal absences frequently cause effects as well. This is especially salient in mechanisms whose proper functioning relies on a double prevention structure. For example, the blockage not blocking the spring after one pulls

¹⁴See also Willemsen and Kirfel (2019) for further evidence of for the central role of norms in our causal judgments.

the trigger is normal, and so is the absence of the blockage of the trapdoor when the first rays of sunlight hit the sensor of the early bird feeder. These absences are *supposed* to occur and the occurrence of the corresponding presences would have been abnormal. Even so, these absences still cause the target effects, i.e. the uncoiling of the spring and the opening of the trapdoor. On the assumption that such double prevention mechanisms are quite prevalent, the abnormality proposal leaves a substantial number of negative causation cases unexplained.

Moreover, causation by normal absences is not unique to double prevention mechanisms. To see this, consider the causal work done by so-called ‘electron holes’ in semi-conductors, which play a central role in the functioning of transistors and LED lights. P-type semi-conductors rely on so-called electron holes, i.e. *absences* of electrons, to generate light. These electron holes move across the semi-conductor by attracting electrons. When we turn on the light switch of a LED light, the electron holes start moving which causes the room to be lit. Once again, these absences are *supposed* to occur and their occurrence is essential to the normal functioning of LED light. The causal work done by these electron holes thus makes for the proper functioning of transistors and LED lights.¹⁵ Here as well, the relevant absences are both normal and causal.

At this point, the proponent of the abnormality approach can choose to rely on her liberal notion of norm. Being abnormal only requires the breaking of *some* norm, and certainly all of these absence must be breaking *some* norm. For example, the absences of the trapdoor blockage and the spring blockage are normal according to the proper functioning norms, but perhaps they are statistically abnormal: more often than not, these blockages are in place. However, this is not a particularly promising way forward. Even if the blockage only stops the spring when the gun is cocked, and most of the guns are in a non-cocked state more often than a cocked state, the pulling the trigger would still cause the gun to fire in virtue of causing the absence of the blockage.¹⁶ To look for some way to categorize these absences as abnormal

¹⁵Schaffer (2005, p. 202–203) also makes note of the causal work attributed to electron holes in physics.

¹⁶Some superficial internet research suggests that this is more or less how some guns work. For double-action revolvers, pulling the trigger causes both the cocking and the releasing of the hammer, which suggests that the absence of the blockage is statistically

would be to miss the point that these examples bear out. It does not seem like the absences in normally operating double prevention mechanisms are causal because they are breaking a norm. Quite to the contrary, it seems as if they are causal in virtue of doing exactly what they should be doing. Even if we find a way to classify them as abnormal, this would not explain their being causal.

Moreover, the problem of spurious negative causation will simply reappear if we are too liberal in ascribing abnormality. To see this, consider my church-going behaviour. Every Sunday at 11 AM sharp, I am not going to church. My not going to church at that time certainly breaks *some* norm. Nevertheless, intervening on my not going to church at 11 AM sharp on any given Sunday would prevent the occurrence of many effects that are not *caused* by my not being in church at that time. If I cycle over an ice patch and subsequently break my arm at 11 AM on a Sunday, my not being in church is not the cause of my breaking my arm. Only religious fanatics would deny that. Similarly, if I stumble upon a briefcase containing a million dollars at 11 AM on a Sunday, my not being in church is not the cause of my instantaneously becoming a millionaire. Only anti-religious fanatics would deny that. The underlying point is quite general; we are breaking norms by not doing things all of the time. We are not pursuing our dreams, we are not writing on our thesis, we are not living according to a maxim that we could reasonably will to be a universal law, etc. For all of these breakings of norms, there are effects that would not have occurred if we *did* follow those norms, but which are not caused by those breakings of norms. Not just any old breaking of a norm makes for negative causation.

Perhaps there are elegant ways of meeting the above challenges. Once again, it would probably be good news for our overall project if there are. After all, it could provide the insensitive interventionist with a solution that *prima facie* does not require it to give up on its thoroughly lightweight character. However, I think that there is a more promising approach in the vicinity. Both the normal and abnormal absences result in their target effects against a broad variety of normal changes in the background conditions, whereas the effects would not occur given some natural changes in the background con-

normal in such revolvers.

ditions of spurious cases of negative causation. Introducing some new jargon, we can say that the correlation between the purported cause and the target effect is *stable* in the good cases, and *unstable* in the spurious cases. In the next section, I argue that this feature allows us to draw a principled distinction between good cases of negative causation and spurious cases, whilst respecting the intuitions driving the above three proposals.

Stability

Woodward (2006) points out an important difference in the dependencies involved in spurious negative causation cases as opposed to good negative causation cases.¹⁷ The dependencies underlying good cases of negative causation are *stable* in a way that the dependencies underlying spurious cases are not.¹⁸ This notion of stability can provide the right truth values for negative causation claims and can predict the seeming unprincipledness of negative causation claims. The subsequent solution thus provides a significant improvement over the three proposals discussed above. Given that the stability of dependencies can be cashed out in terms of stability of correlation patterns, this solution can also be neatly implemented in our insensitive interventionist model of causation.

To say that a causal claim is *unstable* is to say that the counterfactual dependency underlying it would break down under natural variations in the background conditions. Recall our example cases of spurious negative causation:

- (1*) The Belgian prime minister not watering my plants caused them to wither.
- (2*) The Belgian prime minister not holding the trapdoor closed caused the food to land on the saucer.

¹⁷Whilst Woodward explicitly remains neutral on the question whether spurious causal claims are in fact false or merely inappropriate in this text, he appears more inclined to treat spurious negative causal claims as false elsewhere (e.g. Woodward, 2003, p. 91). Moreover his frequently repeated *dictum* that causal relations ought to be ‘available for manipulation and control’ suggests that stability plays a central role for interventionist causation (e.g. Woodward, 2007, 2014, *passim*).

¹⁸Woodward uses ‘insensitive’ rather than ‘stable’. However, I will reserve the term ‘insensitive’ for a more general feature of correlation patterns.

- (3*) The Belgian prime minister not blocking the spring caused it to uncoil.

In the case of (1*), for example, it *is* true that¹⁹

- (i) NOT WATERING $\square \rightarrow$ WITHERED
 (ii) \sim NOT WATERING $\square \rightarrow$ \sim WITHERED

However, very slight and quite natural changes in the background circumstances would have interfered with the truth of (i). Imagine, as is not too unnatural, that I actually *did* water my plants. In such a case, the Belgian prime minister not watering my plants would not have resulted in their withering. Consequently, (i) would have been false.

Similarly, for (2*) it holds that

- (iii) NOT BLOCKING $\square \rightarrow$ UNCOIL
 (iv) \sim NOT BLOCKING $\square \rightarrow$ \sim UNCOIL

However, very slight and quite natural changes in the background circumstances would have interfered with the truth of (iii). Namely, if the trigger had not been pulled and the mechanisms blockage had remained in place, then the prime minister's not blocking it would not have resulted in the uncoiling of the spring. Consequently, (iii) would have been false. In both cases, the positive counterfactual is true, but very unstable.

Woodward argues that this instability explains the sensed falsity of the corresponding negative causal claim. He goes on to remark that causal claims whose negative counterfactuals are unstable are often less objectionable, but cases in which both the positive and the negative counterfactual are unstable are often extra objectionable. In what follows, we will focus mostly on the stability of the positive dependency. The stability of the positive dependence in negative causation claims tracks the truth of those claims. In good cases of negative causation, like (1)–(3), the positive dependency relation *is* stable. The absence results in the target effect against a broad variety of background conditions. This contrasts with clearly spurious cases like (1*)–(3*), which

¹⁹I assume a Lewisian reading of these counterfactuals, but nothing turns on this. They are used to bear out a general point about these dependencies, similar tests would do just as well.

we used to illustrate the spurious negative causation problem. In all of these cases, the Belgian prime minister not performing the relevant action would not have resulted in the target effect given some natural variations in the background conditions. If I had watered the plants, if the sun had risen a few minutes later, and if the shooter had not pulled the trigger, the prime minister not performing those actions would not have resulted in any of these effects. The stability of the dependence relation thus predicts the correct truth values for negative causation claims.

We can also see the tracking relation between stability and the truth of negative causation claims in action by considering grey-area cases of negative causation. There are certain cases where we feel torn about whether or not a dependency relation involving an absence amounts to causation or not. In such cases, the perceived stability tracks our readiness or reluctance to accept the relevant causal claim. Consider Hall's example of double prevention (Hall, 2004, p. 241):

Suzy is piloting a bomber on a mission to blow up an enemy target, and Billy is piloting a fighter as her lone escort. Along comes an enemy fighter plane, piloted by Enemy. Sharp-eyed Billy spots Enemy, zooms in, pulls the trigger, and Enemy's plane goes down in flames. Suzy's mission is undisturbed, and the bombing takes place as planned. If Billy hadn't pulled the trigger, Enemy would have eluded him and shot down Suzy, and the bombing would not have happened.

Hall notices that we feel torn about the following claim (2004, p. 242):

(9) Billy's pulling the trigger is a cause of the bombing

On the one hand, we certainly feel that Billy is in some sense responsible for the bombing, and it is unclear how he could be responsible if he did not cause the bombing. On the other hand, we are reluctant to state that Billy is involved in the causal process leading up to the bombing, as this process seems to initiate in Suzy's aircraft and for all we know the air battle between Billy and Enemy has no effect on her. (9) appears to land in the grey area of negative causal claims.

Information affecting the perceived stability of the dependence between Billy's shooting and the bombing affects our intuitions on (9). Suppose that Enemy is in fact an unreliable fighter pilot, and would only have intercepted Suzy out of sheer luck. Suppose also that the air battle between Billy and Enemy took place 1000 miles from Suzy's position, and Enemy was in fact likely to get lost along the way. These qualifications make us reconsider Billy's causal contribution to the bombing. A lot can happen to an unreliable enemy pilot crossing 1000 miles and a lot can happen to Suzy during several hours. That is to say, both the positive and negative dependence relation between Billy's shot and the bombing now appear less stable. Conversely, Halpern and Pearl (2000) pointed out that we are likely to consider Billy's contribution more favourable if the involved air fighters were automata whose workings we perceive to be predictable and reliable. That is to say, if we have good evidence that the dependence relation between Billy's shot and the bombing is stable, we are more likely to accept (9). The stability of negative causation relations thus tracks our judgments in good cases, bad cases and those cases in between.

The stability proposal fits well with our model of causation developed in the previous chapter. Stability is a kind of insensitivity, and insensitive interventionism was supposed to isolate insensitive correlations as causal. To some extent, stability is the analogue of robustness applied to background conditions instead of base variables. What it means for a dependency between two phenomena A and B to be stable is for the correlation between occurrences of A and occurrences of B to persist across scenarios that contain normal changes in the background conditions. What it means for the correlation between the values of variable B and variable A to be robust, is for it to persist across scenarios that contain changes in A 's base variables. In both cases, this property of the correlation patterns informs us that the positive correlation is not restricted to scenarios where very specific or unnatural circumstances hold. Instead, information about stable and robust correlations are applicable to a broad variety of scenarios we deem natural.

Consequently, stable and robust correlations are, all else equal, more available for prediction, manipulation and control than unstable and non-robust correlations. If, for example, someone wants to see my plants withered, she

will be more effective in her fiendish plans by preventing me from watering them, than by preventing the Belgian prime minister from watering them. After all, preventing me from watering them results in their withering against a variety of normal variations in the circumstances, whereas preventing the prime minister from watering them only does so in a very specific circumstance. Moreover, the stability of correlations are used to explain that we tend to causally select abnormal phenomena (Hitchcock and Knobe, 2009, p. 607–608). Stability therefore provides a unified explanation of negative causation by abnormal as well as normal absences. Aside from this affinity with our model of causation and the explanation of our interest in abnormal absences, the stability proposal also respects the ideas driving the three proposals from the previous section.

First of all, normality and abnormality play an important role in the stability proposal. Correlations are stable if and only if they hold up under *natural* changes in the background conditions. On the plausible assumption that naturelness is normative according to the liberal notion of normativity that is used in the abnormality proposal, it follows that, just like in the abnormality proposal, norm-breaking plays a central role for (negative) causation in the stability proposal. The stability proposal just locates the proper role of norm-breaking with the background conditions rather than with the absences itself. Consequently, normal absences can be causes, but only if they result in their effect across normal changes in the background conditions.

Second, the stability approach reserves an essential role for context and pragmatic considerations. As was borne out by the case of the Venusians and the forest fire, the conversational context and pragmatic considerations can affect which variations in background conditions are considered to be natural. Earthly firefighters hold the absence of oxygen to be an abnormal change in the background conditions, whereas Venusians do not. Woodward provides a more realistic example of the same phenomenon. Economists and psychologists make different assumptions about which background conditions should be held fixed and which should be allowed to vary (Woodward, 2006, p. 13–14):²⁰

[F]or the purposes of doing economics, “abnormal” changes in

²⁰Note that Woodward himself uses the term ‘stability’ rather than ‘insensitivity’ here.

neurological processing are taken to involve irrelevant or ignorable departures from actuality while changes in information or relative costs are regarded as highly relevant to the assessment of stability.

Consequently, a dependency that is stable in the context of economics might not be stable in the context of psychology or pharmacology, and *vice versa*. It thus appears that which background conditions we consider alterable or fixed tight is often context-dependent. Given that the stability of a causal claim depends on which background conditions one considers alterable or fixed tight, the stability of causal claims is thus context-dependent. Consequently, the seeming unprincipledness of negative causation claims can be explained by relying on the capriciousness of conversational context and pragmatic considerations.

Before concluding this chapter on negative causation, it is worth addressing one potential worry for the stability approach. If accounts of causation are to respect our causal judgments, they must allow for seeming cases of unstable causation. That is to say, they must allow for cases of causation where the cause only resulted in the effect given very specific background conditions. Consider the following example. A marksman tries to shoot a practice target from a long distance on a gusty day. He hopes to get the shot in there between the gusts and succeeds in hitting the target. It is highly salient to the marksman — and to us — that the chances of success are minimal and the slightest change in the weather conditions could affect the outcome. Nevertheless, it is true that, if the bullet hits, the marksman's shot is the cause of the target shattering.²¹ One might worry that, by adopting the stability response to spurious negative causation, one cannot allow for such cases of unstable causation. I think there are two compatible lines of response to this worry.

First and foremost, allowing absences to be causes in virtue of exhibiting stable patterns of correlations does *not* require us to maintain that stability is a necessary condition on causation. There could be other ways of causing effects than exhibiting a stable pattern of correlation with that effect. For example, perhaps one phenomenon can cause another in virtue of producing

²¹I am grateful to Gunnar Björnsson for providing this example.

or physically necessitating the effect.²² This point is of particular relevance for our overall project. Recall that we need to deny that production or physical necessitation are *necessary* for causation, but that we do not need to deny that they suffice for causation. We do not have to provide necessary conditions for causation in order for our overall project to succeed, only sufficient ones. Consequently, we can hold that absences can cause effects in virtue of exhibiting stable patterns of correlation with these effects without requiring that all causes do so.

Second, the context-dependence of stability provides some leeway in explaining seeming cases of unstable causation. Even when cases are explicitly presented as unstable, as in the case of the marksman, there might still be cues that make us ignore the sensitivity to background conditions. For example, the use of words like ‘marksman’, ‘shooting’ and ‘target’ paint the picture of a competent agent undertaking an action to achieve a goal. These cues might prompt us to ignore even slight and natural variations in the background conditions and focus on the actions of the agent instead. Hence, setting the context thus that there are no relevant changes in the background conditions and only changes in the marksman’s actions are considered. In general, intervening on the intentions of competent agents is an effective strategy of manipulation, and it would not be surprising if this fact is reflected in our conversational contexts.

13.4 The upshot

In this chapter, we discussed the case of negative causation and its relevance for insensitive interventionism. This discussion bore out two points that are crucial to our project.

First, production requirements on causation systematically contradict our causal judgments. Many respectable causal claims require the possibility of negative causation. In particular, mental-to-behavioural causation in creatures like us requires the possibility of negative causation. It does so even on a physicalist ontology. The prevalence of negative causation provides us with

²²Or by exhibiting a stable pattern of correlation with a phenomenon that produces or physically necessitates the effect. A proposal along these lines can be found in Schaffer (2001b) and Loew (2019).

a powerful response to the production objection against our model of dualist mental causation.

Second, one can respond to the spurious negative causation worry by relying on a particular kind of insensitivity of the correlation patterns involved in good cases of negative causation. In particular, one can rely on the stability of the relevant correlation pattern. If we insert this extra condition in our definition of insensitive interventionism, we arrive at the following formulation:

- (IM-Stability)** X is a type-level direct cause of Y with respect to a variable set V if and only if
- (i) there is a possible intervention on X that will change Y or the probability distribution of Y when one holds fixed at some value all other variables Z_i in V *that do not stand in a nomic necessitation relation or a tight relation to X or Y ,*
 - (ii) the correlation between X and Y under such interventions is robust relative to X 's base variables, and
 - (iii) the positive dependency of Y on X is stable.

Save for the addition of the stability clause (iii), (IM-Stability) is identical to (IM), the central definition in insensitive interventionism.

(IM-Stability) is a definition of direct type-level causation. However, definitions of both indirect causation and token causation follow naturally from this definition. I will not provide a definition of indirect causation, but, as most of the cases in the negative causation debate concern token causation, it is well worth it provide a definition of token causation based on (IM-Stability). Here is an initial proposal:

- Insensitive Token Causation - Stability** X taking value x is a direct token cause of Y taking value y relative to variable set V if and only if, X meets (IM-Stability) relative to Y and V , and there is a possible intervention on x that is followed by a change in y .

This initial proposal might give rise to a problem. Perhaps there are stable

correlations between token absences and their target effect across the relevant alternative scenarios, without there being a stable correlation between the relevant types. This would be the case if there are background conditions that are specific to a certain scenario and that, for whatever reason, are considered normal in this scenario but not in general. Suppose for example, that everyone in the family knows that Amber always throws a tantrum when father cannot make it to family events. Any change in that behaviour would be considered abnormal and therefore not a contextually salient option. Suppose also that father cannot make it to the picnic and Amber throws a tantrum impressive enough to ruin the entire day. In this case, father's absence can be said to ruin the picnic, even if there might not be a stable correlation between the 'paternal absence'-type and the 'ruined picnic'-type. Perhaps cases like these can be solved by finding a better level of abstraction when selecting the relevant type variables, but finding a better level of abstraction is likely to be a complicated and arduous task in many cases.

Another solution is to impose the stability condition at the token-level for token causation and at the type-level for type-causation. (IM-Stability) would thus remain identical, but (Insensitive Token Causation - Stability) would be reformulated as follows:

Insensitive Token Causation - Stability X taking value x is a direct token cause of Y taking value y relative to variable set V if and only if, X meets (IM) relative to Y and V , and there is a possible intervention on x that stably correlates with a change in y .

Here, (IM) is identical to (IM-Stability), *modulo* the type-level stability condition contained in (iii). Given this, token absences can be causes if they exhibit a stable pattern of correlation with the target effect, even if there is no stable correlation between the respective types. There might be further complications to address, but I take it that a proposal of roughly this form can address the problem of spurious negative causation.

The viability of the stability solution to spurious negative causation demonstrates the importance of insensitivity for causation. This is good news, as insensitivity plays a central role in our account of causation. Moreover, it does not appear to interfere with insensitive interventionism's promise to model

dualist mental causation. As said however, the addition of stability might not be essential to our model of dualist mental causation. All we require is a solution that does not impose heavyweight requirements on causation. Consequently, I will not defend the stability approach any more than I have done already. Nonetheless, I would like to make a final note on the implications of adding stability to insensitive interventionism.

By inserting the notion of stability in our definition of causation, we are committed to causal statements being context-dependent. The stability of a correlation is a measure of its sensitivity to natural changes in background conditions, and which changes in background conditions we consider natural depends on the context. Some might see a commitment to contextualism about causal statements as a cost (e.g. Montminy and Russo, 2016), others are convinced that such a *causal contextualism* is plausible anyway (e.g. Schaffer, 2005, 2012). Given limitations of space I cannot satisfactorily address the advantages and drawbacks of contextualism about causal statements here. A more perspicuous formulation of the stability version of insensitive interventionism would probably make this context-dependence more explicit. Given that we will not crucially rely on stability here, I leave such a formulation for another occasion. Note however, that causal contextualism does not appear to conflict with our overall project.

In fact, Maslen (2005) argues that contextualist accounts of causation hold a particular promise for non-reductionist accounts of mental causation.²³ If the relevance of alternative scenarios depends on the conversational context, and the conversational context is determined (at least partially) by the interests of the participants, then scenarios in which the necessitation relation between mental phenomena and their physical bases are severed should plausibly be ignored when evaluating mental causation claims. For example, we can ignore cases where my pain is absent but its physical base is present, as such cases are of no practical importance to us. Even though Maslen primarily focuses on non-reductionist physicalist accounts of mental causation, she appears to be at least open to the possibility of such an approach to dualist

²³To be more precise, Maslen makes a claim about causal *relevance* rather than causation *simpliciter*. I ignore the distinction here because *A* being causally relevant for *B* presupposes that *A* causes *B* on interventionist accounts of causation and causal relevance (Woodward, 2008, p. 227).

mental causation.²⁴

I am sympathetic to such considerations and they could be used to bolster my case for dualist mental causation. As mentioned in our discussion on robustness in the case of dualist mental causation, it would strengthen our case if we can ignore nomically impossible scenarios. However, I do not want to put too much weight on the relevance of causal contextualism for mental causation here. As mentioned, what sets conversational context is an unpredictable matter, and these issues are beyond the scope of this dissertation. Moreover, I take it that even if this approach works, a robustness condition on causation will still have to be in place.

This concludes our discussion of negative causation. In the next chapter, we discuss the relation between causation and fundamental physics.

²⁴As opposed to the proposals by Woodward (2015) and Shapiro (2010), Maslen does not take the *impossibility* of such scenarios to be the relevant feature, but focuses on their contextual irrelevance instead.

Chapter 14

Causation and Physics

In this chapter, I discuss recent work on the relation between causation and fundamental physics.¹ The purpose of this discussion is twofold. First, it will indicate that the physical necessitation requirement on causation conflicts with everyday and scientific causal judgments. A proper understanding of physical necessitation teaches us that the phenomena that are causes according to our ordinary and scientific causal judgments rarely, if ever, physically necessitate their effects (against background conditions or not). Consequently, the physical necessitation objection to insensitive interventionism is mistaken. In relation to this first point, I will argue that some of the attempts to explain the occurrence of causation in the light of the inhospitality of fundamental physics result in accounts of causation that resonate well with insensitive interventionism. In particular, so-called ‘Neo-Russellian’ views on causation agree with the insensitive interventionist’s view that causation consists of robust and stable patterns of correlation available for manipulation and control.

Second, this chapter will bear out a further worry for insensitive interventionism. Given the time-symmetry of the laws of physics, accounts of causation have a difficult time explaining why causes typically precede their effects. Our model of causation runs the risk of predicting that many actual

¹I will often drop the qualifier ‘fundamental’ throughout the rest of this chapter. I understand fundamental to mean the same thing here as in Chapter 2: The laws and facts in fundamental physics are fundamental in the sense that they are not explained by any further laws or facts, but rather explain other laws or facts.

phenomena have causes that lie in their future. If our model predicts this result, its sufficiency conditions for causation are most probably too permissive, and its credibility as a model of dualist mental causation is therefore threatened. I survey some solutions to the time-asymmetry problem from the literature and argue that these solutions are compatible with insensitive interventionism.

The structure of this chapter looks as follows. First, I introduce the three central mismatches between causation and physics that were discussed by Russell (1912): the mismatch in size, grain and time-asymmetry. I also show how they can be used to argue against the (sophisticated) physical necessitation requirement on causation. Second, I discuss the Neo-Russellian approach to reconciling these three mismatches with a realist account of causation, and I argue that the resulting view of causation resonates well with insensitive interventionism. Third, I discuss how, in the light of the first two sections, the mismatch in time-asymmetry might still pose a worry for insensitive interventionism and survey some of the solutions from the literature.

14.1 Mismatches between causation and physics

For the past century, it has been a vexed issue whether or not causation can be reconciled with fundamental physics. Russell (1912) famously held that it cannot be done. In this section, I discuss the three mismatches between causation and physics that Russell took to underlie the incompatibility of causation and physics,² and argue that these mismatches undermine physical necessitation requirements on causation, rather than establishing the incompatibility of causation and physics.

We take causation to relate relatively local and coarse-grained phenomena, such as smoking behaviour and lung cancer. Moreover, we take the causal relation to be asymmetrical: smoking behaviour causes lung cancer, but not the other way around. Russell (1912) pointed out that fundamental physics makes no reference to causes, but instead provides exceptionless

²Russell's text raises further objections to the notion of causation, but these are irrelevant to the matters at hand.

symmetric laws that exclusively relate maximally fine-grained total states of the universe at different times. The leading theories of fundamental physics of his day thus appeared particularly inhospitable to the notion of causation. Russell famously concluded that (1912, p. 1):

The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm.

Though physics has developed considerably since 1912, the mismatches between causation and fundamental physics persist (cf. Field, 2003). As we have seen in Chapter 7, a phenomenon has to be maximally fine-grained and span the entire cross-section of the backwards light cone of a target effect in order to physically necessitate that effect. Physically sufficient phenomena are therefore radically different from the local coarse-grained phenomena we consider to be causes. There is a clear mismatch between the size and grain of typically recognized causal relata on the one hand, and the phenomena covered by the laws of fundamental physics on the other.

Furthermore, the laws of fundamental physics are *time-symmetric*.³ According to contemporary fundamental physics, future phenomena necessitate past phenomena to the same extent that past phenomena necessitate future phenomena.⁴ By contrast, we take there to be a salient time-asymmetry in causation: causes typically precede their effects. The above considerations thus provide us with three mismatches between causation and the fundamental laws of physics: one in the size of the relata, one in the grain of the relata, and one in the time-asymmetry of the relation. If Russell was right that these mismatches render causation and physics incompatible, and we want to hold

³This is not uncontroversial. Collapse interpretations of quantum mechanics, such as the Ghirardi-Rimini-Weber interpretation and the Copenhagen interpretation, *do* have time-asymmetric laws. See Albert (2000, Ch. 7) for a proposal of how the fundamental laws of GRW can play a role in underwriting the time-asymmetry of causation. Given that the purported time symmetry of the fundamental laws of physics will spell trouble for us rather than helping us out (cf. Section 14.3, there is no cheating involved in following Field (2003) and Albert (2000, 2015) here in assuming that the laws *are* time-symmetric.

⁴This is *not* the same as the laws being *time-reversible*. In order for laws to be time-reversible laws any process that is possible in one temporal direction is possible in the other temporal direction as well. Both Field and Albert remark that laws can be time-symmetric without being time-reversible.

on to physics, we should let go of causation.

Cartwright (1979) turned Russell's argument on its head. She argued that the notion of causation plays an indispensable role in underwriting the objective distinction between effective and ineffective strategies. It is true that to stop smoking is an effective strategy to decrease one's risk of incurring lung cancer. It is also true that to bleach one's teeth is an ineffective strategy to decrease one's risk of incurring lung cancer. There is a simple explanation for this difference in effectivity. To stop smoking is an effective strategy because smoking causes lung cancer. To bleach one's teeth is an ineffective strategy because the colour of one's teeth does not cause lung cancer. If fundamental physics cannot allow for causation, it cannot allow for such explanations. Moreover, the fundamental laws of physics appear to be in a poor position to provide alternative explanations. After all, coarse-grained and local phenomena, such as smoking and lung cancer, do not physically necessitate their effects. So if, as Russell appeared to maintain, all physics cares about are lawful necessitation relations, it cannot capture the relevant distinctions. Cartwright concluded that, if the laws of physics cannot allow for causation, then the laws of physics must be defective in some sense.

Note that if the laws of physics are indeed incompatible with causation, this dialectic poses a considerable problem for the dualist's opponent. After all, if, as Russell would have it, there is no causation, then *Mental Causation* is false regardless of the status of dualism. Alternatively, if, as Cartwright would have it, the laws of physics are defective, then it is at least unclear what motivates *Physical Completeness*.⁵ Either way, one of the central premises of the exclusion arguments against dualism turns out to be false, or at least questionable.

Even so, I do not think these are the right lessons to draw from the mismatches between causation and physics. Both Russell's approach and Cartwright's approach are radical. Giving up on causation means giving up on the objective distinction between effective and ineffective strategies for manipulation and deliberating actions. Giving up on fundamental physics means giving up on the leading scientific theory of our universe. All else equal, one should prefer a less radical solution and a less radical solution is

⁵And indeed, Cartwright herself is skeptical of *Physical Completeness*. See Cartwright (2010).

available: one can deny that physical necessitation is necessary or sufficient for causation. As we have seen in Chapters 4 and 7, such a denial is quite standard in contemporary philosophy of causation and, as I will argue below, it allows us to acknowledge the three mismatches between causation and physics without concluding that we have to give up on either one.

When arguing for the incompatibility of causation and physics, Russell crucially relies on the assumption that causes must physically necessitate their effects. The argument Russell provides based on the mismatches in size and grain can be reconstructed as follows:

1. All causes physically necessitate their effects.
 2. If fundamental physics is correct, anything that physically necessitates an effect is enormous and maximally fine-grained.⁶
 3. If our notion of causation is correct, some causes are not both enormous and maximally fine-grained.
-
4. Either fundamental physics is incorrect or our notion of causation is incorrect.

According to the conclusion, we must choose between rejecting fundamental physics and rejecting our notion of cause. However, the conclusion only follows if 1 is true. If 1 is false, causes are allowed to be local and coarse-grained phenomena, because they do not have to physically necessitate their effects. Many have thus concluded that 1 is false (e.g. Blanchard (2016, p. 257); Eagle (2007, p. 171)).

One can formulate a further argument for the incompatibility of causation and physics by focusing on the mismatch in time-asymmetry. It follows from physics that every phenomenon is physically necessitated by enormous and fine-grained phenomena that occur at some *later* point in time. However, causes always *precede* their effects in our world. Therefore, the existence of causation is in contradiction with physics. Here is a rendering of the second argument:

⁶In fact, the phenomenon in question need not be enormous if it were to take place at a minuscule time-interval before or after the effect. However, we rarely if ever are interested evolutions over such minuscule time-intervals, so I ignore that possibility here.

- 1.* For any two distinct phenomena A and B , if A physically necessitates B , then A causes B .
 - 2.* If fundamental physics is correct, then for any physical phenomenon A , there is a physical phenomenon B such that: A precedes B in time, and B physically necessitates A .
 - 3.* If our notion of causation is correct, then for any two actual physical phenomena A and B : if A causes B , then B does not precede A in time.⁷
-
4. Either fundamental physics is incorrect or our notion of causation is incorrect.

Once again, the conclusion only follows if 1* is true. If not all physically sufficient conditions for an effect are causes of that effect, there can be physically sufficient conditions for an effect that occur after that effect without causing it. Even though 1* is probably more contentious than 1, I believe the time-asymmetry argument for 4 poses a more serious challenge than the argument that relies on the mismatches in size and grain. However, I postpone this point to Section 14.3. What matters for now is that one can avoid giving up on either physics or the notion of causation by dissociating between causation and physical necessitation. One can defuse the argument from grain and size by denying that physical necessitation is *necessary* for causation and one can defuse the argument from time-asymmetry by denying that physical necessitation is *sufficient* for causation.

Consequently, there can be no physical necessitation requirement on causation. Recall that we characterized the sophisticated variety of such a requirement as follows in Section 4.1:

The Sophisticated Physical Necessitation Requirement For any two phenomena A and B , A causes B only if there is some set of background conditions c that includes neither B nor anything that physically necessitates B , such that A and c together physically necessitate B .

⁷I restrict this claim to actual physical phenomena, because our notion of causation arguably allows for science fiction cases of backwards causation, such as time travel or precognition (cf. Lewis (1973b, p. 566), Lewis (1976, p. 148), and Mackie (1974, p. 161)).

If we take the mismatch in grain observed by Russell seriously, we can see that such a requirement effectively excludes coarse-grained phenomena from being causally effective. After all, coarse-grained phenomena cannot physically necessitate their target effects even if one is allowed to add background conditions.

To see this, consider again the case of Yue from Chapter 7. Yue threw a rock at a window, causing it to shatter. Now take the background conditions against which Yue's throw caused the window to shatter and assume, for the sake of the argument, that we are allowed to individuate these background conditions in a maximally fine grain. The occurrence of Yue's throw will not physically necessitate the shattering of the windows against these background conditions. There are all kinds of physically possible scenarios in which both the throw and the background conditions occur, but the window remains intact. For example, the fundamental particles making up the rock might be arranged such that it suddenly emits a particle at an immense acceleration that is orthogonal to the rock's anticipated trajectory, causing the rock to make a sudden turn and therefore miss the window. The chance of the rock being realized in this way is vanishingly small, and none of the actual rock throws will ever involve a mid-air turn due to the immensely accelerated ejection of a fundamental particle. However, such trajectories are not physically impossible. Consequently, Yue's rock throw does not physically necessitate the shattering when taken together with its background conditions.

The underlying point generalizes. Typical causes like hurricanes, banking crises and infections are not fine-grained enough to physically necessitate their target effects, even given the background conditions against which they occur. Consequently, if one imposes this requirement on causation, one is forced to conclude that either our causal judgments are systematically mistaken or fundamental physics is mistaken. I take it that neither of these options is attractive to the dualist's opponent.

I will consider two ways of objecting to this this line of reasoning. Both of these argue that there is an understanding of 'background conditions' that does allow coarse-grained phenomena to physically necessitate their target effects against background conditions.

First, one can argue that there *is* a set of background conditions against

which Yue's throw physically necessitates the shattering of the window, but that this set includes *disjunctive* conditions. Such an objection can be developed as follows. Yue's throw metaphysically necessitates a disjunction of maximally fine-grained physical realizers, $\{PR_1, \dots, PR_n\}$. For any one of these physical realizers PR_i , there is a physical background condition BC_i , such that Pr_i and BC_i together physically necessitate the target effect. For example, the thermodynamically abnormal realization of Yue's throw where the ejection of a particle throws the rock off course, will *still* physically necessitate the shattering of the window if taken together with a background condition where Yue's right shoe is physically realized in such a way that it emits a particle at the exact acceleration and angle required to knock the rock back *on* course. The underlying idea is that for any thermodynamically abnormal realization of Yue's throw that would prevent the target effect given thermodynamically normal background conditions, there will be a thermodynamically abnormal realization of the background conditions that will ascertain that the effect occurs after all. If we allow our background conditions to consist of a disjunction of all the members of $\{BC_1, \dots, BC_n\}$, coarse-grained phenomena can, and plausibly quite often do, physically necessitate target effects.⁸

I take it that such an understanding of background conditions is atypical. Necessitation requirements on causation are naturally understood as demanding that the cause necessitated the effect given its *actual* background conditions, rather than a disjunction of the actual background conditions and some further alternatives. It is unlikely that such an understanding of background conditions captures what those who posit necessitation requirements on causation have in mind.

Note further that such a reading of 'background conditions' renders the *Sophisticated Physical Necessitation Requirement* particularly weak. True, dualist mental phenomena will not meet this requirement. Non-physical phenomena cannot be non-redundant parts of physically sufficient conditions for target effects if *Physical Completeness* is true (cf. Section 4.1). However, almost all physical phenomena, coarse-grained or fine-grained, in the backward or forward light cone of the target effect will meet this requirement. For

⁸I am grateful to Pär Sundström for formulating this objection.

example, the physical realization of Yue's shoes will meet this requirement relative to the shattering of the window.⁹ Of course, *Sophisticated Physical Necessitation Requirement* is only a necessary requirement, and one could exclude Yue's shoes by adding further conditions on causation. Nevertheless, this necessary requirement would tell us impressively little about causation and its only purpose appears to be excluding non-physical phenomena like dualist pain. In the absence of independent motivation for such a requirement on causation, we can reasonably insist that this is an *ad hoc* objection to our model of dualist mental causation.

Second, one can argue that there *is* a set of background conditions against which Yue's throw physically necessitates the shattering of the window, but that this set includes some *normality* conditions. If we ignore thermodynamically abnormal scenarios, Yue's throw *does* physically necessitate the shattering of the window when taken together with the other physical phenomena in the backwards light cone of the shattering. Plausibly, this observation will generalize. If we allow our background conditions to include normality conditions, coarse-grained phenomena can, and plausibly quite often do, physically necessitate target effects.¹⁰

However, if we allow a normality condition in our background conditions, it is no longer clear why dualist mental phenomena are excluded from being causes. *Normally*, occurrences of pain are accompanied by a physical nomic base phenomenon. Occurrences of pain without such a physical nomic base are not allowed by the nomic naturalist dualist ontology. Plausibly, there are sets of normal physical conditions that do not physically necessitate my wincing on their own, but *do* physically necessitate my wincing, when combined with me experiencing a *normal* occurrence of pain. Of course, there are ways to characterize normality that excludes thermodynamically abnormal scenarios but allow for nomically impossible scenarios — the term 'thermodynamically normal' seems explanatory enough. But here again the question is whether there is an independently motivated way of doing so. One way of motivating normality conditions is that we can ignore the relevant abnormal

⁹Field (2003, p. 439) makes a similar point: "[...] there would be a big deal if we had to conclude that if c_1 and c_2 are both in the past light cone of e then there is no way of regarding one of them as any more a cause of e than the other."

¹⁰I am grateful to Torfinn Huvenes for suggesting an objection along these lines.

scenarios for all intents and purposes, but this holds for nomically impossible scenarios as well as for thermodynamically abnormal scenarios. This is not to say that normality conditions cannot have a place in accounts of causation. Indeed, they will play a central role throughout the remainder of the chapter. It is just hard to see how normality can be characterized in a non-*ad hoc* way such that it excludes thermodynamically abnormal scenarios but includes nomically impossible scenarios.

I take these considerations to effectively address the physical necessitation objection against insensitive interventionism. Together with the findings from the previous chapter, this provides powerful evidence against heavyweight accounts of causation. The apparent prevalence of causation by absences and coarse-grained phenomena is strong evidence that causation requires neither production nor physical necessitation. However, our discussion on physical necessitation gives rise to a further question. Given that the laws of fundamental physics do not give us much by way of explanation of how there can be causation in the actual world, one wonders what *does* explain causation. How come we are able to manipulate future phenomena by manipulating local coarse-grained phenomena? Given *Physical Completeness* and *Physicalism about the Non-Mental*,¹¹ one expects there to be some explanation that is properly grounded in physics. One could worry that *that* explanation will not allow for dualist mental causation.

In the next section, I discuss one influential account of how physics allows for causation and argue that it does not provide us with reasons to abandon insensitive interventionism, and hence allows for dualist mental causation.

14.2 The Neo-Russellian project

Even though it is now shown that causation and physics are not incompatible, the three mismatches between causation and fundamental physics still demand an explanation. If the laws of physics are so inhospitable to causation and almost all phenomena are in the end determined by physics, then how

¹¹As the reader might recall, this principle states that all occurrences of non-mental phenomena are metaphysically necessitated by physical phenomena.

can coarse-grained and local phenomena cause effects?¹² The Neo-Russellian project aims to explain how causal relations between local and coarse-grained phenomena can occur in a world with physical laws like ours. In this section, I discuss the central tenets of such Neo-Russellian strategies and demonstrate how a central step in such strategies is to equate causation with robust and stable patterns of correlation.

It is worth setting aside one central aspect of the Neo-Russellian project before continuing. Neo-Russellians maintain that there is no causation at the fundamental physical level. However, Frisch (2014, Ch. 6) has forcefully argued that causal reasoning plays a central role in fundamental physics. For our purposes we do not need to take a stance on this issue. Instead we will focus on the Neo-Russellian explanation of causal relations between local coarse-grained variables. Let us now turn to these issues.

Different developments of the Neo-Russellian project can be found in Albert (2000, 2015), Blanchard (2014), Field (2003), Fernandes (2017), Ismael (2016a), Loewer (2012), and Price and Weslake (2009). These accounts can differ considerably in their eventual models of causation, but they all rely heavily on findings from thermodynamics discussed by Albert (2000, 2015). For our purposes, it suffices to understand how these findings from thermodynamics allow us to explain the mismatch in size, grain and time-asymmetry.

According to the Neo-Russellian, thermodynamics presupposes the following principles: *the statistical postulate* and *the past hypothesis*. The statistical postulate provides us with the likelihood of a certain micro-phenomenon occurring given that a certain macro-phenomenon occurred. This postulate allows us to explain difference in grain between causal relata and the phenomena covered by the laws of fundamental physics. The past hypothesis states that our universe was in a particular global low entropy macrostate in the distant past. This hypothesis allows us to solve the asymmetry problem for causation. Let us take them in turn.

In order to understand the statistical postulate, we need to understand the relation between macro-phenomena and micro-phenomena. Very roughly

¹²Eagle (2007, sect. 7.3) remarks that the situation is somewhat similar to the case of causal exclusion. According to the exclusionist, physical causation leaves no room for non-reductionist mental causation. When one takes a closer look at the fundamental laws of physics, it can appear as if these laws leave no room for causation.

speaking, macro-phenomena are coarse-grained, whereas micro-phenomena are maximally fine-grained. Furthermore, the former are multiply realizable by the latter. Macro-phenomena are individuated in terms of macroscopic properties and objects. For example, Cyril's stubbing his toe is a macro-phenomenon, because it is individuated in terms of macroscopic objects and macroscopic properties like toes and being stubbed. Consequently, macro-phenomena are relatively coarse-grained, because their occurrence imposes relatively mild requirements on what is the case. Micro-phenomena are maximally specific phenomena individuated in the terms of objects and properties studied in fundamental physics, like leptons, quarks, and spin. For example, the fundamental physical particles realizing Cyril's toe-stubbing having their exact position, spin, etc. is a micro-phenomenon. Such micro-phenomena are maximally fine-grained, as their occurrence imposes very strict requirements on what is the case. There is of course, consistent with Cyril's stubbing his toe, a wide variety of micro-phenomena that could have occurred at the time and place of Cyril's stubbing his toe. Consequently, macro-phenomena are multiply realizable by micro-phenomena.

The statistical postulate maps that realization relation. It does so by ascribing an equal probability to all the micro-phenomena consistent with a given macro-phenomenon. It thereby provides us with the probability of a certain micro-phenomenon occurring, given the occurrence of a certain macro-phenomenon. For example, it provides us with the probability that a certain micro-phenomenon occurs at a certain time and place, on the assumption that Cyril stubs his toe at that time and place. This allows us to move from claims about the occurrence of coarse-grained phenomena, like Cyril's stubbing his toe, to claims about the probability of the occurrence of a maximally fine-grained phenomena. Consequently, we can move from claims about phenomena with the coarse-grainedness of everyday causes to claims about phenomena with the fine-grainedness of phenomena that can physically necessitate an effect.

Of course, the necessitating phenomenon should also be of the right size. However, this size problem has a natural solution. When evaluating causal claims, we *hold fixed* the conditions surrounding the purported cause. In particular, the Neo-Russellian strategy is to hold the actual background con-

ditions, individuated macroscopically, fixed when making causal claims. By doing so, we can move from claims about local phenomena to claims about global phenomena.

We can now see how local, coarse-grained phenomena can exhibit *regularities* without being physically sufficient for one another. The statistical postulate, together with the fundamental laws of nature, will allow us to calculate how likely a certain macro-phenomenon MP_1 at time t_1 is to evolve into certain other macro-phenomenon MP_2 at time t_2 : we just look at how many of the micro-phenomena realizing MP_1 will evolve into micro-phenomena that realize MP_2 at t_2 . If, by using this method, one could discover regularities between the occurrence of local macro-phenomena, these regularities could be effectively exploited for manipulation and prediction.

As an empirical matter of fact, such regularities do exist. For example, it is likely that the vast majority of global micro-phenomena that realize the macro-phenomenon of Cyril's stubbing his toe, will, together with the fundamental laws of nature and some background conditions, entail the occurrence of a micro-phenomenon that realizes the macro-phenomenon of Cyril's wincing some time later. There is a tiny minority of the micro-phenomena realizing Cyril's stubbing that entail abnormal futures when combined with fundamental laws of nature. Perhaps Cyril stubs his toe and goes on to transform into a book that describes the life of Richard Nixon, or perhaps he goes on to walk out of the room, backwards and undisturbed; but such micro-phenomena really make up but a tiny minority of the set that realizes Cyril's stubbing his toe. It is highly irregular for toe-stubbings to result in such phenomena — so irregular that they will never result in such phenomena in the actual world. Toe-stubbings *do* regularly result in wincing. Similar regularities can be found for all familiar causal relations. Or at least, that is what the Neo-Russellian maintains.

It is worth noting what it means for such a regularity to exist. If Cyril's stubbing his toe exhibits a strong regularity with his wincing, this means that the correlation between the stubbing and his wincing is both *robust* and *stable* in the sense we have outlined earlier. For a wide variety of micro-realizations of both the stubbing and the macroscopic background conditions of his stubbing, it is the case that Cyril will be wincing sometime later.

Similarly, for a wide variety of micro-realizations of both people smoking and their macroscopic background conditions, it is the case that they will incur lung cancer sometime later. These regularities allow us to manipulate future phenomena by manipulating local and coarse-grained phenomena in the present. In Albert's terminology, such regularities provide us with 'causal handles' on the future.

There are no such causal handles on the past. This asymmetry remains unexplained by the conjunction of the fundamental laws and the statistical postulate. If we were infer what most likely *preceded* Cyril's stubbing his toe based on the global macro-phenomenon at the time of the stubbing, the statistical postulate and the fundamental laws, we would have to conclude that Cyril was most likely wincing. Based on just these three data, it would in fact be highly unlikely that Cyril just walked in the room without a semblance of physical discomfort. In fact, it would be just as unlikely as Cyril walking out of the room backwards *after* he stubbed his toe. If we are to explain why we can manipulate the future but not the past, we require a time-asymmetric restriction on what the world is like.

The past hypothesis provides such a restriction. It postulates that some particular global low-entropy macro-phenomenon occurred in the distant past, but that no such phenomenon occurs in our future. Given that entropy is extremely unlikely to decrease in closed systems, and our universe is a closed system, this means that entropy steadily increases towards the future in our universe, whereas it is on a steady decrease in the future-to-past direction. If we restrict the possible scenarios we consider to those that are compatible with the past hypothesis and, using the statistical postulate and the fundamental laws, try predict what would happen if we manipulate local coarse-grained phenomena, it will transpire that, as an empirical matter of fact, such manipulations will exhibit robust and stable regularities with local and macroscopic changes in the *future*, but not in the past.¹³ Or, put simply, there are causal handles on the future, but not on the past.

We now have a rough outline of the Neo-Russellian project. It starts with the fundamental laws of nature and aims to explain the occurrence of causal

¹³Frisch (2007) argues that Albert overestimates the asymmetry that is provided by the past hypothesis. There might still be flukes in other possible worlds that affect macro-phenomena in the past. I postpone this worry to the next section.

relations. It proposes to resolve the grain problem by adding the statistical postulate as a given. It proposes to resolve the symmetry problem by adding the past hypothesis as a given. It proposes to resolve the size problem by holding background conditions fixed — but that was hardly original. There are different strategies to develop this rough outline into a full-blown account of causation, but they all require this fundamental set-up. The fact that there is causation in our world is, according to the Neo-Russellian, ultimately explained by the statistical postulate and the past hypothesis, as well as the fundamental laws of nature.

Note however, that the success of this strategy is contingent on causation consisting of robust and stable patterns of correlation. The Neo-Russellian set-up does *not* provide us with an asymmetry of physical necessitation or flow of energy. It provides us only with a time-asymmetry in the correlation patterns between local and coarse-grained phenomena. These patterns of correlation are taken to explain how I can cause the occurrence of a flame by striking a match: given my current macroscopic background conditions, the fundamental laws of physics, the thermodynamical postulate, and the past hypothesis, my striking a match is most likely to be realized in such a way that it will result in a flame. According to the Neo-Russellian resolution of the mismatches between causation and physics, that is what it means for the striking of a match to cause the flame. It is also what allows us to manipulate occurrences of the latter by manipulating occurrences of the former. Ismael summarizes the situation as follows (2016b, p. 136):

[T]he way in which the dilemma created by the Russell/Cartwright exchange was resolved is that causal structure turns out not to be fundamental but part of a user interface. Causal Pathways highlight emergent regularities, robust enough to support hypothetical reasoning, that can be used as strategic routes to action for appropriately situated agents.¹⁴

Effective strategies are thus available to us despite the fact that the laws of physics make no mention of the kind of phenomena we are in a position to manipulate.

¹⁴Ismael's does not use 'robust' in my technical sense here, but appears to have a more general insensitivity to changes in mind.

In the light of this dialectic, it seems ill-conceived to demand *more* from causes than to exhibit such patterns of correlations with their target effects. In particular, it would be undermotivated to deny the possibility of dualist mental causation. *Prima facie*, our world is such that the occurrence of my pain will, against a wide variety of background conditions, most certainly be instantiated in such a way that it is followed by the presence of a wince. This will be true according to the nomic naturalist dualist ontology we have assumed as well as according to physicalist ontologies.¹⁵ On the plausible assumption that insensitive interventionism picks out such patterns as causal, it seems that a closer study of physics does not motivate us to reject insensitive interventionism. Consequently, a closer study of physics does not motivate the standard assumption that nomic naturalist dualists cannot have mental causation.

Of course, the Neo-Russellian project is just one candidate for providing an account of causation that is properly grounded in physics. Other proposals might not be as friendly towards insensitive interventionism or dualist mental causation. Although I cannot provide an extensive treatment of the alternatives here, I would like to note that some of these are of no help to the dualist's opponent either. For example, Dowe's *Physical Causation* (2000) proposes a productive account of causation, which, as we have seen in Section 13.1, spells trouble for those who want to reject dualism on the grounds of mental causation problems. Others, like Norton (2007) and Kutach (2007), agree with Russell that a proper understanding of physics motivate a skeptical attitude towards causation in general and if there is no such thing as causation, it should not be held against the dualist that she cannot allow for mental causation. So even if the Neo-Russellian project turns out to be misguided, there are other ways to resolve the tension between causation and physics that might be of help to the dualist. For now, I take the relatively broad support for Neo-Russellianism to provide support for insensitive interventionism as well.

¹⁵Although it remains an open empirical question whether or not the relevant correlation patterns are robust (cf. Chapter 12).

14.3 Spurious backwards causation

The Neo-Russellian strategy explains the mismatches between causation and fundamental physics without giving up on either one of them. At first sight, this strategy meshes well with an interventionist conception of causation. Both Neo-Russellians and interventionists conceive of causation as a metaphysically lightweight relation available for manipulation and control.¹⁶ Moreover, robustness and stability, which play a crucial role in the Neo-Russellian strategy to explain the occurrence of causation, helped interventionism avoid two spurious causation problems (cf. Chapters 11 and 13). However, there is a remaining issue. Frisch (2005, 2007, 2014) has argued that the Neo-Russellian strategy does not provide the strict asymmetry in correlation patterns required to back causal asymmetry. In light of this, our insensitive interventionism runs the risk of resulting in spurious backwards causation; i.e. the thesis that many effects have causes that lie in their future. In this section, I briefly expound Frisch's criticism and demonstrate how it might pose a problem for insensitive interventionism. I conclude by surveying some of the strategies to avoid this problem.

We discussed how the Neo-Russellian proposes to back the causal asymmetry with a thermodynamical asymmetry. By relying on the statistical postulate and by maintaining that a particularly low entropy state lies in our past but not in our future we can conclude that local coarse-grained changes in the present reliably correlate with local, coarse-grained changes in the future, but not with local, coarse-grained changes in the past. However, the asymmetry in correlation patterns provided by the statistical postulate and the past hypothesis might not be strict enough to back causal asymmetry. We should be extraordinarily lucky if the correlations provided by the fundamental laws of nature and our two posits from thermodynamics line up neatly with our causal judgments.¹⁷ Can it really be the case that there are *no* stable and robust correlations between changes in local, coarse-grained phenomena in the present and changes in local, coarse-grained phenomena in

¹⁶And indeed, many interventionists are also Neo-Russellians: Blanchard (forthcoming); Hitchcock (2007); Pearl (2000) and Woodward (2007).

¹⁷Albert argues that, if the Neo-Russellian picture is right, it is very likely that we are evolutionarily adapted to picking up on the right patterns and consequently consider these as causal. On such a proposal, the above comment gets things the wrong way around. I set this possibility aside here, as a discussion on evolution would take us too far afield.

the past?

To see the worry more clearly, consider the following example provided by Frisch (2007, p. 383). We are asked to consider a system of moving billiard balls on a table. We know that all the balls were neatly racked in a triangle some ten seconds ago; our knowledge of this past low entropy state will serve as our stand-in for the past hypothesis. Further, we know the state of all the billiard balls right now. In particular, we know that the 5 ball is currently in position p and has velocity v . Now imagine that we are asked what would have happened five seconds earlier if our stand-in past hypothesis was held fixed and everything the current situation was held fixed as well, except that the 5 ball was in a different position and had a different velocity. In other words, we are asked what the past would have looked like given a change in a local, coarse-grained phenomenon. Intuitively, our knowledge of the dynamics of billiard balls — we have knowledge of that as well — would force us to conclude that there were some local, coarse grained differences between the past in the hypothetical situation and the actual situation. Of course, this change would have to occur in a thermodynamically abnormal way, but it is unclear why that abnormal change should have left the past ten seconds unmeddled with but for a brief and sudden change immediately before the current moment. Frisch formulates the problem as follows (2007, p. 383) :

I take it that Albert’s intuition is that something thermodynamically ‘odd’ must have happened to the balls in order for ball 5 to end up at a macroscopically distinct present location and quite plausibly this intuition is correct. But what is difficult to see is why the most plausible past evolution of the counterfactual system of billiard balls is supposed to be one that *exactly* matches the actual macro-evolution until immediately before the present and only then diverges in some thermodynamically unexpected way.

That is to say, in this system, there appear to be patterns of correlations between changes in the present and changes in the past.

Frisch’s example does not discuss robustness or stability, but it is at least unclear that these restrictions will be of help. If stability is what will counter

this case, it seems like there will be an easy work-around. Instead of considering a system of billiard balls, we can consider a ball rolling through a solid steel tube, where the tube protects the trajectory of the ball against any external influences. It seems plausible that the correlations between changes in the present state of the ball and its past are sufficiently stable to qualify as causal. The robustness requirement is not as obvious to accommodate, but it is equally unclear that it will solve the issue. Perhaps there will indeed be no robust correlations between present and past phenomena in such systems, but this will be up to the underlying physics and there appears to be no strong reason to believe that things will go our way. Here is Frisch's assessment of the situation (2007, p. 384):

Now, ultimately the question whether or not it is a consequence of the dynamics, the statistical postulate, and the past-hypothesis that worlds that differ from the actual world locally and macroscopically are overwhelmingly likely to have had exactly the same macroscopic past as the actual world should not be a question that is settled through a battle of intuitions. Whether Albert's thesis is right is a question for the relevant physics.

The upshot is that allowing correlations to be causal in virtue of being stable and robust *might* result in allowing for backwards causation. Only further research in physics will tell.

This poses a challenge for proposals to analyze causation in terms of patterns of correlations. In particular, it poses a challenge for insensitive interventionism. In the absence of further restrictions on causal correlations, we might be committed to the conclusion that some causes have effects that precede them. To see this, consider how insensitive interventionism would treat the case of billiard ball 5. Let us take the variable P_1 to represent its position at the current time and variable P_0 to represent its position five seconds ago. For simplicity, we can say that all of these variables only have an 'on' value and an 'off' value. The I variable represents an intervention as usual. The relevant causal graph is provided in Figure 14.1. In the actual scenario, both variables take the 'on' value. The question is what would happen to the value of P_0 if there had been an intervention on P_1 . Following the above reasoning, it would seem that we have to conclude that there

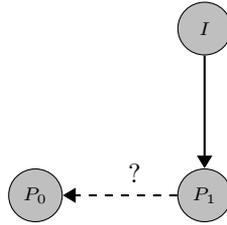


Figure 14.1: Did the 5 ball being in its actual position at $t = 1$ (P_1) cause it to be at its actual position at $t = 0$ (P_0) ?

could very well be robust and stable correlations between interventions on P_1 and changes in P_0 . Consequently, our insensitive interventionism runs the risk of spuriously treating P_1 as a cause of P_0 . In as far as systems sufficiently similar to our billiards case exist in the actual world, insensitive interventionism runs into the problem of spuriously backwards causation: it spuriously treats phenomena as causes of phenomena that preceded them.

There are several proposals to deal with this problem. Some are specific to interventionism, others are more general. It is unclear to me which of these (if any) should be preferred. The main obstacle for selecting the correct response is that the proof will be in the pudding, and figuring out what the pudding is like requires a thoroughly applied understanding of physics. Frisch is skeptical that the thermodynamically abnormal processes will turn out exactly like we would want them to, whereas Albert (2000, 2015) and Loewer (2007a) appear to be more optimistic. Others have proposed extra constraints on causal correlations to deal with time-asymmetry problems. I briefly survey some of these extra constraints below and note that these do not conflict with the thoroughly lightweight aspirations of insensitive interventionism. Implementing any of those strategies would therefore not threaten our overall project.

First, there are two interventionist-specific responses worth considering. Some have proposed that the technical notion of an intervention resolves the problem. Interventions are *causal* interactions with the purported cause variable, and we know that if we were to interact with the position of the billiard ball causally, it would affect its future position but not its past position. In general, we know that causal interactions robustly correlate with phenomena

in their relative future, but not with phenomena in their relative past. This proposal relies on the idea that interventions should model actual *physical* interactions with the purported cause. However, as we have seen in Section 9.2, interventions are highly idealized in order for them to meet the ‘holding fixed’-requirements in (IV), and it is unclear why such highly idealized intervention that causes a change in P_0 and holds all else fixed would behave like any physical interaction we are familiar with. Even so, some believe interventions should be thought of as stand-ins for actual physical interactions (e.g. Blanchard, 2015; Polger et al., 2018),¹⁸ and if this can be made to work, the spurious backwards causation worry disappears indeed. Note also, that if this is how one should think of interventions, our (IIV), which exempts nomic bases, is plausibly a better definition than the standard (IV) definition, which does not exempt nomic bases. After all, actual physical interactions on nomically necessitated phenomena always require changes in the nomic bases of these phenomena.

Another common response is that interventionism avoids spurious the backwards causation problem because it requires interventions to be *exogenous*. The exogeneity requirement on interventions states that interventions must come from outside the causal system one is investigating. Some have argued that intervening on variables that would exhibit the required correlation with phenomena preceding it is therefore impossible. The motivation for this kind of response appears to come from Pearl’s frequently quoted statement that (2000, p. 349–350):

If you wish to include the entire universe in the model, causality disappears because interventions disappear — the manipulator and the manipulated lose their distinction. [...] In most cases the scientist carves a piece from the universe and proclaims that piece in — namely the focus of investigation. The rest of the universe is then considered out or background and is summarized by what we call boundary conditions. This choice of ins and outs creates asymmetry in the way we look at things, and it is this

¹⁸Woodward himself can be hard to pin down on how idealized interventions in fact should be, but he does believe that some actual manipulations might qualify as interventions (2018, p. 15), which suggests that they are not as idealized as we have assumed here (cf. Section 9.3).

asymmetry that permits us to talk about ‘outside intervention’ and hence about causality and cause–effect directionality.

The underlying idea is that, whilst the laws governing the entirety of the universe may be symmetric, asymmetry arises if one focuses on correlations under interventions in local systems. Replies of this kind can be found in Eagle (2007, p. 171), Hitchcock (2007, p. 53–54), and Woodward (2007, p. 92–93). However, it remains unclear how this response deals with cases like billiard ball 5, where the system under investigation appears to be local and intervention from outside seems patently possible. Moreover, even if the only variables that exhibit backwards correlations of the kind Frisch wants to bear out with his example are indeed so enormous as to span, say, the entire cross-section of the forward light cone of the target effect, then outside intervention would still be *metaphysically* possible (cf. Frisch (2014, p. 94), Reutlinger (2013, sect. 2) and Schaffer (2010)).¹⁹ This reply as well, seems to rely on the idea that interventions should not be so idealized as to be physically impossible. Without such a restriction on interventions, the exogeneity requirement on intervention cannot fix the spurious backwards causation problem.

Aside from these interventionist-specific responses, there are several Neo-Russellian responses to such problems. All of these come with their own advantages and drawbacks and many authors choose to combine different strategies. This issue deserves closer attention, but I lack the space to properly defend or evaluate these proposals and will restrict myself to an all too brief discussion of three strands of proposals in the literature.

First, one can rely on other asymmetries aside from the thermodynamic asymmetries. For example, Fernandes (2016) proposes to rely on evidential asymmetries, where an intervention on a cause by a properly deliberating idealized agent is required to constitute evidence for the occurrence of the effect. Blanchard (2016) and Price and Weslake (2009) build on agential asymmetries, where the causal asymmetry derives from the fact that actions by agents like us robustly correlate with changes in the future, but not with changes in the past.²⁰ The latter proposal comes close to the first interventionist-

¹⁹Reutlinger concludes that interventionists cannot be Neo-Russellians. As is apparent from this chapter, I think the two do not exclude one another.

²⁰Albert hints at something in this direction when he states that anything that would fit

specific approach discussed earlier; it relies on the asymmetry of correlations that are due to actual causal interactions, rather than the highly idealized interventions that we have used in this text.

Second, one can rely on the asymmetry of typical causal systems. Menzies and Price (1993); Price (2007) and Price and Weslake (2009) argue that the causal direction is based on the more pervasive asymmetry in patterns of correlations under interventions. Even though there might be specific systems, like our billiards system, where the patterns of correlation are not asymmetric, such systems are uncommon. Once one takes the perspective of agents like us in a world with such pervasive asymmetric patterns, the causal direction is set from past to present even for the occasional system where the actual asymmetry in correlations is absent. Consequently, causal asymmetry is the result of both a pervasive asymmetry in correlation patterns and our perspective as agents.²¹

Third, one can rely on the context-dependence of causation. For example, Eagle (2007, p. 166–167) proposes that those variables that exhibit spurious backwards correlations might be contextually irrelevant. Consequently, due to the context-dependence of causation, there is time-asymmetry between cause and effect. Field (2003) appears to make a similar suggestion when he argues that causation goes forward if we restrict our attention to *salient* variables.²² It is not clear however, why seemingly innocent variables like P_1 and P_0 should be barred from salience or contextual relevance. Field suggests that variables that do exhibit a robust and stable pattern of correlation with variables that precede it are too fine-grained to be salient and Albert makes a similar suggestion when stating that opportunities to affect the past are “[...] rare or impractical or invisible or in some other way beside the point” (Albert, 2015, p. 52). Once again, the answer seems to be up to the underlying physics.

Note that none of these five proposals would require us to give up on the idea that causation is a thoroughly lightweight phenomenon. If anything, the extra requirements they impose provide more grist for the mill of the

our concept of an agent could systematically influence the future, but not the past (Albert, 2015, p. 44).

²¹Ismael (2016a) also seems friendly to such a response.

²²Field can also be read as proposing that causation in fact goes backwards between certain non-salient variables.

dualist. For example, if causation is tightly connected to our capacities as agents, then our capacity to manipulate behaviour by manipulating mental phenomena would speak strongly in favour of mental causation, regardless of one's ontology of mind. That is to say, if the question 'does C cause E ?' corresponds (very roughly) to the question 'could a situated agent that is relevantly like us manipulate E by manipulating C ?',²³ then mental phenomena stand a good chance of being causes, even if dualism is true. Of course, a lot of the hard work will be to spell out 'relevantly like us' in the right way. Even if agents like us cannot affect the celestial movements, we want these movements to still be causal; some abstracting away from our capacities as agents will be necessary. Nonetheless, the tendency towards making causation part of 'the user interface' rather considering it a heavyweight relation that drives the universe is more promising than it is ominous for insensitive interventionism, and hence for the dualist.

Finally, we should note that the spurious backwards causation problem is not specific to insensitive interventionism. Any attempt to analyze causation in terms of patterns of correlation will encounter this problem; it is *not* triggered by the exemption of nomic bases from the 'holding fixed'-requirements. So although spurious backwards causation might pose a challenge for our model of causation, it does not pose a problem that is specific to our overall project.

14.4 The upshot

The goal of this chapter was twofold. First, it was supposed to counter the physical necessitation objection. A closer look at the mismatches between causation and physics demonstrated that causes are typically too coarse-grained to physically necessitate target effects, even given a fixed set of background conditions. Consequently, thoroughly lightweight accounts of causation, like insensitive interventionism, have a forceful reply to such physical necessitation objections.

²³And indeed, that is how Price and Corry summarize their edited volume on causation and physics: "The connecting theme in these essays is that to reconcile causation with physics, we need to put ourselves in the picture: we need to think about why creatures in our situation should represent their world in causal terms" (Price and Corry, 2007, dustcover).

Related to this first point, I also presented the Neo-Russellian picture of the relation between causation and physics. This provided us with a rough outline of how there can be causation in the actual world, despite the apparent inhospitality of fundamental physics. Crucially, the explanation provided by the Neo-Russellians provides us with nothing more than robust and stable patterns of correlation between local, coarse-grained phenomena. It does not appear that a closer study of physics reveals a feature of causation that insensitive interventionism fails to capture.

Second, this chapter discussed the potential worry of spurious backwards causation. Although I did not defend a particular solution to this problem, I remarked that it is not specific to insensitive interventionism, nor due to its thorough lightweighness. Further, I surveyed some of the solutions proposed in the literature and argued that these can be implemented without threatening the thoroughly lightweight character of insensitive interventionism.

Before turning to the conclusion of this dissertation, I address some remaining objections to insensitive interventionism.

Chapter 15

Objections and Replies

The previous two chapters in effect addressed the objection that causation must be heavyweight. The subsequent discussions drew us into issues surrounding negative causation and the relation between causation and physics. In this chapter I address some further objections against insensitive interventionism that require a less extensive treatment.

15.1 Physical equivalence

Even if one sympathizes with the thoroughly lightweight aspirations of insensitive interventionism, one might worry that they are bound to lead the account astray in some cases. Physics contains examples of equivalences between metaphysically distinct physical phenomena with distinct causal roles. These physical equivalences can appear to generate counterexamples to insensitive interventionism. In this section, I provide two candidates for such counterexamples and sample four possible replies.

Consider the following two scenarios. In the first scenario, the pressure of an ideal gas G steadily increases until its container explodes. Plausibly, the pressure of the gas caused the container to be destroyed. According to the ideal gas law, the gas having a given pressure P in this situation physically necessitates it having a certain temperature T , and *vice versa*. That is to say, in this situation there is a physical equivalence between the pressure and the

temperature of the gas. However, we would not want to say that the heat of gas caused the container to be destroyed. The pressure did the causal work and the heat is a side-effect.

In the second scenario, a light switch is turned on and the light bulb emits light because of the electric current in the copper wire connecting the switch to the light bulb. In order to make the causal relation between the current and the emitted light more salient, we can even imagine that the intensity of the emitted light is a measure of the electric current. According to Ampère's law, the electric current flowing through the copper wire in this situation physically necessitates a magnetic field in the surface surrounding the copper wire, and *vice versa*. That is to say, in this situation there is physical equivalence between the magnetic field in a given area surrounding the copper wire and the electric current in the wire. However, we would not want to say that the magnetic field caused the emitted light. The current did the causal work and the magnetic field is a side-effect.

Such cases pose a challenge for our account of causation. Insensitive interventionism runs the risk of counting the magnetic field and the temperature of the gas as causes of the emitted light and the exploding container respectively. *Prima facie*, these cases correspond to the familiar schema represented in Figure 15.1. Moreover, the equivalence ascertains that there can

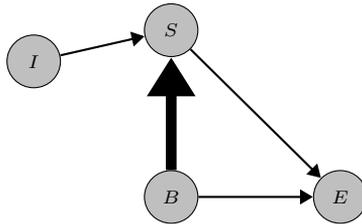


Figure 15.1: Phenomenon (S) is nomically necessitated by its nomic base (B), both cause target effect (E)

be no changes in one variable whilst the other is held fixed. Consequently, the relevant correlations are automatically robust. Such problem cases are likely to multiply, as physics is filled with laws that equate metaphysically distinct phenomena.

There might not be a catch-all response to these kinds of cases. I sample

some of the responses available to the insensitive interventionist when meeting physical equivalence-styled objections. Some of these might work better for some cases than for others, but I take this sample to be varied enough for it to provide a solid line of defense against such objections.

First, we can rehearse our remarks on the restrictive use of ‘nomic’ we are operating under (cf. Chapters 2 and 12). The ‘nomic’ we have used to qualify the necessitation relation between mental and physical phenomena posited by the dualist refers to a very specific kind of relation. A relation is nomic in this sense only if it is a necessitation relation according to the *fundamental* laws of the universe. That is to say, a relation R qualifies as *nomic* if and only if aRb entails that worlds containing a but lacking b (or *vice versa*) would require a lapse in the fundamental laws of the universe. Even though there are many laws, in physics and elsewhere in science, these are rarely of such *nomic* modal strength. For example, the ideal gas laws are merely statistical. That is to say, even though one can safely bet one’s life on the pressure of a gas increasing when heated in an enclosed rigid container, it is nomically possible for the pressure to remain constant.¹ Although physics is probably the right place to look for nomic laws, there are plenty of laws that are physical but not nomic.

Even with this caveat, there will be some physical equivalences left standing. Perhaps Ampère’s law is in fact a good candidate of a nomic law in our restrictive sense. If this is indeed the case we will need a different line of response against the second counterexample. In order to test our three remaining responses, I will assume that both Ampère’s law and the ideal gas law in fact express a nomic relation. This assumption has the further advantage that, if the first reply turns out to be inadequate for other reasons, we can still rely on one of the replies below.

For the second line of response, we need to take a closer look at the physical laws underlying these examples. Even though these laws state that there is a physical equivalence between two variables *in a given situation*, they do *not* state that these two variables are physically equivalent *simpliciter*. That is to say, these equivalences only hold across scenarios where some further factors (other than the fundamental laws of physics) are held fixed.

¹In fact, one should still be careful betting one’s life, as the law breaks down in cases with very high pressures. I set this complication aside here.

For example, here is a more detailed formulation of Ampère's law:

$$\oint_C \vec{B} \cdot d\vec{\ell} = \mu_0 \mu_r I$$

The left hand side refers to the integration of the magnetic field (\vec{B}) along a closed curve (C). The right hand side refers to the product of the permeability of a vacuum (μ_0), the permeability of the surface area (μ_r), and the electric current (I). It follows from Ampère's law that the magnetic field of that surface is nomicallly necessitated by these three factors. However, this does not mean that the magnetic field is nomicallly necessitated by the electric current of the copper wire. There can be changes in the magnetic field in virtue of changes in the medium surrounding the wire rather than changes to the electric current inside the wire. For example, if the wire were surrounded by water or enclosed in rubber, the same current would have given rise to a different magnetic field. The electric current plays a non-redundant part in the nomic base of the magnetic field, but it is, so to speak, not the *full* nomic base of the magnetic field.

Similar observations hold for the ideal gas law. According to this law, the heat of an ideal gas is equivalent to the the pressure of the gas across situations where the volume and the amount of substance of the gas. A more detailed formulation of the ideal gas law looks as follows:

$$T = \frac{PV}{nR}$$

Aside from pressure (P) and temperature (T) it also involves the volume (V), the ideal gas constant (R) and the amount of substance of the gas (n), i.e. the number of atoms or molecules. According to this law, there can be changes in the temperature in virtue of changes in the volume or the amount of substance, *without* changes in the pressure if the gas.² All in all, the pressure of the gas plays a non-redundant part in the nomic base of its temperature, but it is not the *full* nomic base of the temperature.

²Moreover, ideal gasses are a theoretical entity that is composed of point particles, rather than molecules or atoms. Even though one can treat many gasses like an ideal gas for most purposes, one needs to use the Van der Waals equation, which contains further terms that are gas-specific, for better accuracy. Consequently, the temperature of the gas also depends on what kind of gas it is.

These equations indicate that the analogy between the counterexample and purported cases of causation by dualist mental phenomena is not perfect. When modelling the dualist case, we have assumed that the underlying physical phenomenon *is* the full nomic base of the nomically necessitated mental phenomenon. However, this difference does not directly resolve the issue. For one thing, it would be unprincipled to impose the ‘holding fixed’-requirements on non-redundant parts of nomic bases whilst exempting full nomic bases from the ‘holding fixed’-requirements.³ Moreover, imposing the ‘holding fixed’-requirements on non-redundant parts of nomic bases will deliver the wrong results. For example, it would inhibit us from exempting the magnetic field from the ‘holding fixed’-requirements when intervening on the electric current. Consequently, the electric current would not count as a cause of the emitted light. That is obviously the wrong result.

Nevertheless, these observations reveal an important feature of the counterexamples in question. In particular, they allow us to demonstrate that the correlation between the bad candidate for a cause and the target effect is sensitive in a way that the correlation between the good candidate for a cause and the target effect is not. Consider again the case of the electric current and the magnetic field. If we were to change the magnetic field by changing the *medium* rather than the current, there will be no change in the light intensity. Conversely, if we change the electric current, but hold the magnetic field fixed by bringing about changes in the medium, there *will* be a change in the light intensity. The result is a correlation pattern that contains features of both asymmetric screening off patterns and unstable correlation patterns: the current of the wire asymmetrically screens off the correlation between the magnetic field and the light intensity across scenarios that contain changes in the medium. *Mutatis mutandis*, the same holds for the correlation between the temperature of the gas and the explosion of the container: the pressure of the gas asymmetrically screens off the correlation between the temperature of the gas and the explosion across scenarios that contain changes in the background conditions.

The insensitive interventionist can exploit this contrast in correlation patterns. For example, the correlation rate between the magnetic field and the

³Remember that we exempted non-redundant parts of metaphysical bases for similar reasons in Chapter 10.

emitted light is sensitive to changes in the medium whereas the correlation rate between the current and the emitted light is not. Consequently, measuring or manipulating electric current provides a more reliable and easy way to predict and manipulate bulbs emitting light than measuring and manipulating the magnetic field. Given that the (insensitive) interventionist aims to isolate insensitive correlation patterns, i.e. those patterns that are available for reliable manipulation and prediction, it is in line with the motivations of her account to exclude correlations that are asymmetrically screened off across changes in the background conditions. One can do so by adjusting the robustness condition accordingly or by adding an extra condition addressing this kind of correlation pattern separately. I do not provide a worked out fix here, but a reply along these lines is likely to deal with a good share of cases.

A third reply is to point out that our judgments in such cases are sensitive to how one frames the relevant manipulations. When sketching the case of the light bulb and the copper wire, I said that the electric current was activated by the flicking of a switch. This sketch invites the following interpretation: the switch turning on activated an electric current, which gives rise to a magnetic field in the area surrounding the copper wire and causes the light to turn on. The magnetic field seems causally irrelevant to the light turning on. Now consider this scenario: the switch works by creating a magnetic field in the surface along the copper wire. By grace of Ampère's law, this magnetic field results in an electric current that causes the light to turn on. Now it seems right to say that the magnetic field is a cause of the light bulb emitting light. In this situation, the electric current does not causally exclude the magnetic field.

Note that this framing effect appears to be quite general. With some scene setting, we also accept claims like 'the heat of the gas caused the container to explode'. Suppose for example that someone had left the gas container in a car on a hot day. The heat in the car gradually increased the temperature of the gas and hence the pressure it exerted on the container. When filling out the car insurance form, the car owner states that the heat of the gas caused the container to explode, hence causing the damage to his car. It seems wrong to say that the car owner misdescribes the situation. Or to put the same point in a different way, even if the car owner were a physics

professor, it seems wrong to say that she lies when stating that the heat of the gas caused the damage in the car.

Note further that this framing effect is present in mental-physical cases as well. Suppose that Caroline, a brilliant neuroscientist, implants electrodes in Henry's brain. With the use of a remote control, she can make him utter whatever words she wants. For example, if she types 'I am in pain', the electrodes activate the relevant neurons, which cause Henry to say 'I am in pain' and generate a pain sensation along the way. In this case, the neuronal activity seems to cause him to say 'I am in pain', and the actual pain seems to be a side effect. If however, Caroline resorts to simply pricking Henry with a needle to make him say 'I am in pain', we are more inclined to agree that the actual pain was a cause of the utterance, even if the neural phenomena leading up to the utterance are the same as in the electrode case.⁴

The insensitive interventionist can choose to exploit these framing effects in either of the following ways. She can posit that we should make the truth of causal claims sensitive to the nature of the manipulation. Perhaps the manipulation sets the context in a way that allows variations in certain variables but fixes others. Alternatively, she can say that we are easily fooled by framing manipulations differently, but the causal facts remain constant across variations in the nature of the manipulation. Perhaps the first option is more in keeping with suggestions made earlier, but I leave both options open here.

Finally, the insensitive interventionist can bite the bullet. If there are some physical equivalence cases left that cannot be answered in any of the three ways just outlined, the insensitive interventionist can insist that, in such cases, the nomically equivalent phenomena *do* cause the same effects. This of course, is the least elegant of the four strategies, but one hopes it will be required for only a limited number of cases.

⁴In fact, the parallels between mental cases on the one hand and the gas and magnetism cases on the other might even be useful to the dualist. In the case where the heat of the car caused the temperature (and pressure) of the gas to go up, it seems true to say that the temperature of the gas caused the container to explode even though the nearest possible world where the temperature was lower would still have resulted in the explosion. After all, it is metaphysically possible to lower the temperature whilst keeping the pressure constant and all else is held fixed. Given that it is in fact even *nomically* possible to do so, one might be inclined to think that insensitive interventionism is *too* restrictive to deal with such cases. I do not address the worry that insensitive interventionism is not lightweight enough here, as it does not pose a threat for our overall project.

15.2 Gerrymandering

As noted by Franklin-Hall (2016), one can always gerrymander a variable that hits the sweet spot on robustness and stability. One could do so by constructing a variable that takes the disjunction of all the sufficient phenomena for the target effect as its ‘on’ value, and the disjunction of all the sufficient phenomena for the non-occurrence of the effect as its ‘off’ value. For example, if we focus again on Sophie, the pigeon who pecks at red objects, we can construct a variable that takes ‘being presented with a red object in the right circumstances, or being tickled in the right circumstances, or being hungry in the right circumstances, or . . . ’ as one value, and takes a suitably reverse-engineered value for her not pecking as the only alternative.⁵ Interventions on such a variable will stably and robustly correlate with changes in the target effect (cf. Section 11.2).

Such gerrymandered variables can threaten the causal status of plausible cause variables. For any given target effect, one can plausibly gerrymander a variable that asymmetrically screens off the correlation between any plausible cause variable and that effect. For example, the correlation between redness and Sophie’s pecking will be asymmetrically screened off by such gerrymandered variables. If this is the case, the robustness requirement we introduced in insensitive interventionism will result in some form of causal drainage from plausible cause variables to gerrymandered variables.

Woodward (2018) and Blanchard (forthcoming) respond to such challenges. They argue that, far from being within the spirit of interventionism, preferring such overly abstract gerrymandered variables over more specific variables like redness violates the central motivation of interventionism. The interventionist strives to provide an account of causation that isolates those relations that are exploitable for manipulation and control. Gerrymandered variables in fact *hide* such relations by including irrelevant information. For example, by lumping together redness, tickling and hunger, the gerrymandered variable in Sophie’s case hides the fact that there are (at least) three independent causal routes to Sophie’s pecking.⁶ By doing so, it hides the

⁵Of course, ‘the right circumstances’ will require more spelling out, but one gets the general idea.

⁶This is not to say that the presence of these causal routes could not in principle be derived from a model that represents the dependence of pecking on the gerrymandered

fact that one can simply manipulate redness in order to ensure that Sophie pecks, *independently* of any facts about tickling our hunger. Consequently, relying on overly abstract gerrymandered variables would “[...] do a less than satisfactory job at achieving what interventionists regard as the fundamental goal of explanation, namely, identifying explanandum-changing interventions” (Blanchard forthcoming, p. 20).

Woodward and Blanchard respond to the challenges from gerrymandered variables by referring back to the central motivation of interventionism. One of the reasons for developing insensitive interventionism is that it is more in line with that central motivation. The insensitive interventionist can thus safely avail herself of the same response.

15.3 On the Woodward-Baumgartner debate

My account of dualist mental causation relies heavily on Woodward’s (2008; 2015) interventionist account of higher-level causation. However, this account has been subject to a series of criticisms. The most vehement critic of interventionist higher-level causation is probably Baumgartner (2009; 2010; 2013; 2018). His objections also inspired others to criticize Woodward’s proposals. We can distinguish two objections that spawned from this literature. According to the first, interventionist accounts fail to distinguish between the different causal roles of phenomena that stand in a tight relation. According to the second, such accounts fail to distinguish between epiphenomena and higher-level causes. One could worry that insensitive interventionism will be subject to the same objections. I briefly outline a response to both of these objections.

First, Eronen and Brooks (2014) argue that higher-level phenomena will have exactly the same causal role as their metaphysical bases according to (M*) and (IV*). Recall that (M*) and (IV*) are variations of the standard interventionist definitions, with exemption clauses to the ‘holding fixed’-requirements for variables that stand in tight relations. We dubbed the

variable, but rather that such a model would not *represent* these three distinct causal routes. Just like economics might in principle be derivable from fundamental physics, but fundamental physics does not represent the causal dependency of banking crises on real estate prices (cf. Woodward, 2018, p. 21–22).

interventionism that consists of just these two definitions *Minimal Interventionism*. Eronen and Brooks remark that, given that all interventions on higher-level phenomena are interventions on their metaphysical bases as well, the two will have identical causal roles according to *Minimal Interventionism*. For example, redness and being auburn will have the same causal role on this account. As we have seen in the cases of Sophie and Sarah, that is implausible. Sometimes being auburn is a cause when redness is not.

I take this problem to be solved by adding a robustness requirement to *Minimal Interventionism*. If we require that higher-level phenomena result in the target effect across changes in the base variables, then the causal roles of those higher-level phenomena will not be identical to the roles of their metaphysical bases (*pace* Kim (2005)). For example, if an object's being red results in a target effect only if it is also auburn, being auburn will be a cause of that effect, but redness will not. Given that insensitive interventionism contains a robustness requirement, it is not affected by this objection.

Second, Baumgartner (2009, 2010) argues that *Minimal Interventionism* fails to distinguish between epiphenomena and higher-level causes. However, this criticism is clearly misguided. *Minimal Interventionism* makes a clear distinction between which higher-level phenomena are causes, i.e. those that satisfy the (M*) and (IV*) requirements, and which are not, i.e. those that do not satisfy (M*) and (IV*). We can see that there is a significant difference between the two, because dualist mental phenomena *are* epiphenomena according to *Minimal Interventionism*, whereas non-reductionist physicalist mental phenomena are not. True, the difference between nomic necessitation and metaphysical necessitation is empirically untestable, because *in practice* subjecting nomic bases to holding fixed requirements and subjecting metaphysical bases to such requirements are equally unfeasible. However, all this means is that some questions about the metaphysical relations between variables need to be solved before an informative causal graph can be devised. Figuring out whether the relation between variables is tight before drawing conclusions on what causes what is consistent with the interventionist account of causation. The empirical untestability of epiphenomenalism is a well-known fact (e.g. Smart, 1959, 155–156), but it is not the interventionist's problem.

One could worry that insensitive interventionism meets a revised version of this objection. After all, we eliminated the difference between nomic bases and metaphysical bases from the interventionist account. However, the robustness requirement still provides clear guidelines for distinguishing higher-level causes from epiphenomena. If the correlation between a higher-level phenomenon and a target effect is asymmetrically screened off by a base variable, that higher-level phenomenon is an epiphenomenon relative to that effect. If the correlation between a higher-level phenomenon and *all* effects is asymmetrically screened off by its base variables, that higher-level phenomenon is an epiphenomenon *simpliciter*. Consequently, insensitive interventionism does provide a principled distinction between epiphenomena and higher-level causes.

15.4 Conclusion

Now that we have considered some objections to insensitive interventionism's sufficiency conditions on causation, we can assess the credibility of this model. I submit that insensitive interventionism is worth taking seriously. None of the objections we considered provided us with strong reasons to reject this model and its plausibility equals that of standard interventionist models of causation.

First, note that the only objection to insensitive interventionism that specifically addressed the thorough lightweightness of its sufficiency condition was the physical equivalence objection. As we have seen, there are several responses to this objection that do not endanger the model's hospitality to dualist mental causation.

Second, the remaining objections also affect reasonably well-accepted theories of (mental) causation. In particular, they affect the standard interventionist models defended by Campbell (2008, 2010); List and Menzies (2009); Woodward (2008, 2015), and others. These models deny that causation is productive and that causes must physically necessitate their effects. Further, they require a response to the problems of spurious negative causation and spurious backwards causation. We have surveyed some of the responses from the literature, and it appeared that none of these critically rely on causation

not being thoroughly lightweight. That is to say, to the extent that these standard accounts can provide convincing replies to such objections, the insensitive interventionist has convincing replies at her disposal as well. Similar remarks apply to the problem of gerrymandered variables and Baumgartner's criticism of interventionist higher-level causation.

In summary, insensitive interventionism performs as well as standard interventionist models. Given that such standard interventionist models are well-respected in philosophy of causation and a closer look at negative causation and the relation between causation and physics spelled problems for heavyweight accounts of causation, insensitive interventionism appears to be in good shape. If there are good reasons to accept standard interventionist models, there are equally good reasons to accept insensitive interventionism.

Aside from the good reasons we all have to accept insensitive interventionism, the dualist has a strong additional incentive to prefer this model over other accounts of causation. By providing a non-*ad hoc* solution to the exclusion worry and the common cause worry, insensitive interventionism makes for a significant improvement over previous models of dualist mental causation. Consequently, it provides the dualist with a plausible reply to the most pressing objection against her position.

This is not to say that the dualist's claim to mental causation is as secure as the physicalist's. As we have seen in Chapter 10, the exclusion principles that affect non-reductionist physicalist mental phenomena result in universal causal drainage, but there is no similar drainage worry for those who deny dualist mental causation. And as we have seen in Chapter 5, nomic dualist mental causation is not supported by Lewisian counterfactuals, whereas non-reductionist physicalist mental causation *is* supported by such counterfactuals. There is no denying that the physicalist has more tools at her disposal when addressing mental causation worries. Even so, the dualist does not need to acquiesce when being pushed in the epiphenomenalist's corner. She can stand her ground and reasonably maintain that a credible account of causation allows mental phenomena to be causes of our behaviour within her ontology.

Chapter 16

Conclusion

Now that we have responded to some objections, we have time for a brief summary and a few concluding remarks.

I presented and defended a model of causation that allows for dualist mental causation in worlds that are physically complete. In Part I of this dissertation, I explained why such a model is remarkable. The standard view in philosophy of mind is that there cannot be (non-overdetermining) mental causation in worlds where both *Dualism* and *Physical Completeness* is true. A model of causation that allows mental phenomena to be causes in such worlds must deny at least one of the seeming truisms about causation that make up the exclusion argument and will be hard pressed to provide a principled distinction between causation and spurious correlations between effects of a common cause.

In Part II, I argued that the current proposals to establish dualist mental causation leave room for improvement. The objections raised against the exclusion argument can be countered with relative ease (Chapters 5 and 7), and Lowe's models of dualist mental causation do not explain how phenomena that do not stand in tight relations to the physically sufficient conditions of their target effects can (non-overdeterministically) cause those effects (Chapter 8). Kroedel's supernomological dualism *does* provide such an explanation: if the dualist posits that psychophysical laws are modally stronger than regular nomic laws but weaker than metaphysical laws, then the relevant counter-

factuals support the conclusion that mental phenomena cause our behaviour (Chapter 6). However, Kroedel's proposal requires an *ad hoc* posit about the relative distances between possible worlds. The dualist would be better off if she could allow for mental causation without relying on an *ad hoc* posit.

In Part III, I turned our attention to the notion of causation. I presented a minimal version of interventionism and argued that this minimal version needs to be supplemented with a *robustness* condition on causal correlations in order to avoid spurious higher-level causation. The upshot is that higher-level phenomena can be causes in virtue of exhibiting a robust correlation with their target effects. Such a view of higher-level causation is reasonably well-accepted. However, if one assumes this view on causation, it is hard to motivate why merely nomically necessitated phenomena (such as dualist mental phenomena) cannot be causes in virtue of exhibiting such robust correlations with their target effects. Arguably, allowing for causation by nomically necessitated phenomena would be more in keeping with the interventionist party line that those correlations that are available for reliable prediction, manipulation and control are causal. I developed an interventionist model that does allow for causation by nomically necessitated phenomena called 'insensitive interventionism' and showed how it provides a response to the exclusion worry as well as the common cause worry.

Finally, I considered some objections to the sufficiency conditions on causation imposed by insensitive interventionism. If we set aside the objection from physical equivalence (Section 15.1), the objections we considered affect standard varieties of interventionism as well as insensitive interventionism. That is to say, these objections do not specifically target insensitive interventionism's hospitality to dualist mental causation, but concern interventionist accounts more generally. Consequently, the insensitive interventionist can avail herself of the same strategies as standard interventionists when addressing such objections. I surveyed some of the responses from the philosophical literature on causation, argued for my preferences among these if I had some, and showed that insensitive interventionism can incorporate these responses whilst remaining hospitable to dualist mental causation. The upshot is that the standard view on dualism and mental causation is undermotivated, and perhaps even false.

I do not expect to have convinced the dualist's opponent. The idea that there cannot be non-physical causes in a world that is physically complete has an overwhelming intuitive appeal. I too feel the pull of saying that, if dualism is true, the physical nomic bases of mental phenomena must be doing the actual causing in purported cases of mental causation. Even so, I submit that this idea is undermotivated given the current state of philosophy of causation. For comparison, consider the attitude of many physicalists towards the intuition of distinctness we have about mental phenomena and physical phenomena. Physicalists will often admit that they do feel the pull of saying that pain is metaphysically distinct from any physical phenomenon. They then go on to argue that, given some considerations about how phenomenal concepts operate or about the limits of our knowledge, it is undermotivated to maintain that they actually are metaphysically distinct.¹ I think the dualist can make a similar move. She can admit that there is an intuitive thought that her position must result in epiphenomenalism, but that, upon closer scrutiny, these intuitions are not supported by the philosophical studies of causation.

At the end of the introduction, I promised to briefly consider the possibility that my model of mental causation is inadequate. Perhaps a damning objection is around the corner and dualist mental causation will have to be abandoned. Even if this turns out to be the case, the considerations put forward in this dissertation will be of interest to the dualist. It transpired that our causal judgments are at odds with our intuition that causation is heavyweight. Accounts of causation will thus have to sacrifice one or the other. If one chooses to sacrifice our causal judgments, mental phenomena will be in the good company of other phenomena we took to be manifestly causal. If one gives up on the intuition that causation is heavyweight, then the difference between causal correlations and the patterns of correlation exhibited by dualist mental phenomena and their target effects is probably not substantial. Either way, there is no substantive feature that manifest causes have and dualist mental phenomena lack. The dualist could still rest easy. It is not the end of the world.

¹See Papineau (2002) for a defense of the phenomenal concept strategy. See Stoljar (2008b) and Sundström (2018) for defenses of the ignorance strategy.

Bibliography

- Albert, D. Z. (2000). *Time and Chance*. Harvard University Press.
- Albert, D. Z. (2015). *After Physics*. Harvard University Press.
- Andersen, H. (2017). Patterns, information, and causation. *Journal of Philosophy*, 114(11):592–622.
- Andreas, H. and Günther, M. (2019). Causation in terms of production. *Philosophical Studies*.
- Anscombe, G. E. M. (1993). Causality and determination. In Tooley, M. and Sosa, E., editors, *Causation*, pages 88–104. Oxford University Press.
- Anthony, L. M. (2015). Reality and reduction: What’s really at stake in the causal exclusion debate. In Horgan, T., Sabatés, M., and Sosa, D., editors, *Qualia and Mental Causation in a Physical World*, pages 1–24.
- Armstrong, D. M. (1968). *A Materialist Theory of the Mind*. Routledge.
- Armstrong, D. M. (1996). *A World of States of Affairs*. Cambridge University Press.
- Armstrong, D. M. (1999). The open door: Counterfactual versus singularist theories of causation. In Sankey, H., editor, *Causation and Laws of Nature*, pages 175–185. Kluwer Academic Publishers.
- Baron, S. and Miller, K. (2014). Causation in a timeless world. *Synthese*, 191(12):2867–2886.

- Baumgartner, M. (2009). Interventionist causal exclusion and non-reductive physicalism. *International Studies in the Philosophy of Science*, 23(2):161–178.
- Baumgartner, M. (2010). Interventionism and epiphenomenalism. *Canadian Journal of Philosophy*, 40(3):359–383.
- Baumgartner, M. (2013). Rendering interventionism and non-reductive physicalism compatible. *Dialectica*, 67(1):1–27.
- Baumgartner, M. (2018). The inherent empirical underdetermination of mental causation. *Australasian Journal of Philosophy*, 96(2):335–350.
- Beebe, H. (2004). Causing and nothingness. In Paul, L. A., Hall, N., and Collins, J., editors, *Causation and Counterfactuals*, pages 291–308. MIT Press.
- Beebe, H. (2006). Does anything hold the universe together? *Synthese*, 149(3):509–533.
- Bennett, J. (1988). *Events and Their Names*. Oxford University Press.
- Bennett, K. (2003). Why the exclusion problem seems intractable, and how, just maybe, to tract it. *Noûs*, 37(3):471–497.
- Bennett, K. (2007). Mental causation. *Philosophy Compass*, 2(2):316–337.
- Bennett, K. (2008). Exclusion again. In Kallestrup, J. and Hohwy, J., editors, *Being Reduced*, pages 280–307. Oxford University Press.
- Bennett, K. (2017). *Making Things Up*. Oxford University Press.
- Bernstein, S. (2015). The metaphysics of omissions. *Philosophy Compass*, 10(3):208–218.
- Bernstein, S. (2016). Overdetermination underdetermined. *Erkenntnis*, 81(1):17–40.
- Berto, F. and Jago, M. (2019). *Impossible Worlds*. Oxford University Press.
- Bieri, P. (1992). Trying out epiphenomenalism. *Erkenntnis*, 36(3):283–309.

- Björnsson, G. (2007). How effects depend on their causes, why causal transitivity fails, and why we care about causation. *Philosophical Studies*, 133(3):349–390.
- Blanchard, T. (2014). Causation in a physical world. PhD dissertation.
- Blanchard, T. (2015). Review of M. Frisch, *Causal Reasoning in Physics*. *Notre Dame Philosophical Review*.
- Blanchard, T. (2016). Physics and causation. *Philosophy Compass*, 11(5):256–266.
- Blanchard, T. (forthcoming). Explanatory abstraction and the Goldilocks problem: Interventionism gets things just right. *British Journal for the Philosophy of Science*.
- Block, N. (2003). Do causal powers drain away? *Philosophy and Phenomenological Research*, 67(1):133–150.
- Bogardus, T. (2013). Undefeated dualism. *Philosophical Studies*, 165(2):445–466.
- Bontly, T. D. (2005). Proportionality, causation, and exclusion. *Philosophia*, 32(1):331–348.
- Bourget, D. (2019). Anomalous dualism: A new approach to the mind-body problem. In Seager, W., editor, *The Handbook of Panpsychism*. Routledge.
- Brown, C. D. (forthcoming). Exclusion endures: How compatibilism allows dualists to bypass the causal closure argument. *Analysis*.
- Burge, T. (1979). Individualism and the mental. *Midwest Studies in Philosophy*, 4(1):73–122.
- Burge, T. (1986). Individualism and psychology. *Philosophical Review*, 95(1):3–45.
- Burge, T. (1993). Mind-body causation and explanatory practice. In Heil, J. and Mele, A. R., editors, *Mental Causation*, pages 97–120. Oxford University Press.

- Campbell, J. (2008). Interventionism, control variables and causation in the qualitative world. *Philosophical Issues*, 18(1):426–445.
- Campbell, J. (2010). Control variables and mental causation. *Proceedings of the Aristotelean Society*, 110(October 2009):15–30.
- Campbell, K. (1970). *Body and Mind*. Anchor Books, Garden City, N.Y.
- Carroll, J. W. (1991). Property-level causation? *Philosophical Studies*, 63(3):245–270.
- Cartwright, N. (1979). Causal laws and effective strategies. *Noûs*, 13(4):419–437.
- Cartwright, N. (2010). Natural Laws and The Closure of Physics. In Chiao, R., Cohen, M., Leggett, A., Phillips, W., and Harper, C., editors, *Visions of Discovery: New Light on Physics, Cosmology and Consciousness*, pages 612–623. Cambridge University Press.
- Chalmers, D. J. (1996). *The Conscious Mind*. Oxford University Press.
- Chalmers, D. J. (2010). *The Character of Consciousness*. Oxford University Press.
- Chalmers, D. J. (2013). Panpsychism and panprotopsyism. *Amherst Lecture in Philosophy*, 8.
- Chalmers, D. J. and McQueen, K. (2014). Consciousness and the collapse of the wave function. <https://www.youtube.com/watch?v=DIBT6E2GtjA>. Accessed: 2018-11-03.
- Clark, M. J. and Wildman, N. (2018). Grounding, mental causation, and overdetermination. *Synthese*, 195(8):3723–3733.
- Clarke, R., Shepherd, J., Stigall, J., Waller, R. R., and Zarpentine, C. (2015). Causation, norms, and omissions: A study of causal judgments. *Philosophical Psychology*, 28(2):279–293.
- Clutton, P. and Sandgren, A. (2019). A new puzzle for phenomenal intentionality. *Ergo: An Open Access Journal of Philosophy*, 6.

- Crane, T. (1995). The mental causation debate. *Proceedings of the Aristotelian Society*, LXIX:2011–236.
- Crane, T. and Arnadóttir, S. T. (2013). There is no exclusion problem. In Gibb, S. C., Lowe, E. J., and Ingthorsson, R. D., editors, *Mental Causation and Ontology*, pages 248–265. Oxford University Press.
- Davidson, D. (1967). Causal relations. *Journal of Philosophy*, 64(21):691–703.
- Davidson, D. (1970). Mental events. In Foster, L. and Swanson, J. W., editors, *Essays on Actions and Events*, pages 207–224. Clarendon Press.
- Demarest, H. (2015). Fundamental properties and the laws of nature. *Philosophy Compass*, 10(5):334–344.
- Dennett, D. C. (1978). Current issues in the philosophy of mind. *American Philosophical Quarterly*, 15(4):249–261.
- Dennett, D. C. (1988). Quining qualia. In Marcel, A. J. and Bisiach, E., editors, *Consciousness in Contemporary Science*, pages 42–78. Oxford University Press.
- Dennett, D. C. (1991a). *Consciousness Explained*. Little Brown, Boston.
- Dennett, D. C. (1991b). Real patterns. *Journal of Philosophy*, 88(1):27–51.
- Dennett, D. C. (2012). The mystery of David Chalmers. *Journal of Consciousness Studies*, 19(1-2):1–2.
- DeRose, K. (2002). Assertion, knowledge, and context. *Philosophical Review*, 111(2):167–203.
- Descartes, R. (1970). *Descartes: Philosophical Letters*. Clarendon Press.
- Descartes, R. (2003). Extract from meditations ii & vi. In Heil, J., editor, *Philosophy of Mind: A Guide and Anthology*, pages 36–59. Oxford University press.
- Dorr, C. (2016). Against counterfactual miracles. *Philosophical Review*, 125(2):241–286.

- Dowe, P. (2000). *Physical Causation*. Cambridge University Press.
- Dowe, P. (2001). A counterfactual theory of prevention and 'causation' by omission. *Australasian Journal of Philosophy*, 79(2):216–226.
- Dowe, P. (2009). Causal process theories. In Beebe, H., Hitchcock, C., and Menzies, P., editors, *The Oxford Handbook of Causation*. Oxford University Press.
- Dretske, F. (1993). Mental events as structuring causes of behavior. In Heil, J. and Mele, A. R., editors, *Mental Causation*. Oxford University Press.
- Eagle, A. (2007). Pragmatic causation. In Price, H. and Corry, R., editors, *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford University Press.
- Engelhardt, J. (2017). Interactive, inclusive substance dualism. *Philosophia*, 45(3):1149–1165.
- Eronen, M. I. (2012). Pluralistic physicalism and the causal exclusion argument. *European Journal for Philosophy of Science*, 2(2):219–232.
- Eronen, M. I. (2017). Interventionism for the intentional stance: True believers and their brains. *Topoi*.
- Eronen, M. I. and Brooks, D. S. (2014). Interventionism and supervenience: A new problem and provisional solution. *International Studies in the Philosophy of Science*, 28(2):185–202.
- Fair, D. (1979). Causation and the flow of energy. *Erkenntnis*, 14(3):219–250.
- Fernandes, A. (2016). A deliberative account of causation: How the evidence of deliberating agents accounts for causation and its temporal direction. PhD dissertation.
- Fernandes, A. (2017). A deliberative approach to causation. *Philosophy and Phenomenological Research*, 95(3):686–708.
- Field, H. (2003). Causation in a physical world. In Loux, M. J. and Zimmerman, D. W., editors, *The Oxford Handbook of Metaphysics*, pages 435–460. Oxford University Press.

- Fodor, J. A. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthese*, 28(2):97–115.
- Fodor, J. A. (1989). Making mind matter more. In *A Theory of Content and Other Essays*, pages 137–159. MIT Press.
- Foster, J. A. (1991). *The Immaterial Self: A Defense of the Cartesian Dualist Conception of Mind*. Routledge.
- Franklin-Hall, L. R. (2016). High-level explanation and the interventionist's 'variables problem'. *British Journal for the Philosophy of Science*, 67(2):553–577.
- Frisch, M. (2005). Counterfactuals and the past hypothesis. *Philosophy of Science*, 72(5):739–750.
- Frisch, M. (2007). Causation, counterfactuals, and entropy. In Price, H. and Corry, R., editors, *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, pages 351–395. Oxford University Press.
- Frisch, M. (2014). *Causal Reasoning in Physics*. Cambridge University Press.
- Garrett, B. J. (2000). Defending non-epiphenomenal event dualism. *Southern Journal of Philosophy*, 38(3):393–412.
- Gebharder, A. (2017). Causal exclusion and causal bayes nets. *Philosophy and Phenomenological Research*, 95(2):353–375.
- Gibb, S. C. (2013a). Introduction. In Gibb, S., Lowe, E. J., and Ingthorsson, R. D., editors, *Mental Causation and Ontology*, pages 1–17. Oxford University Press.
- Gibb, S. C. (2013b). Mental causation and double prevention. In Gibb, S. C., Lowe, E. J., and Ingthorsson, R. D., editors, *Mental Causation and Ontology*, pages 193–213. Oxford University Press.
- Gibb, S. C. (2014). Mental causation. *Analysis*, 0(0):1–12.
- Gibb, S. C. (2015a). The causal closure principle. *Philosophical Quarterly*, 65(261):626–647.

- Gibb, S. C. (2015b). VIII—Defending dualism. *Proceedings of the Aristotelian Society*, 115(2pt2):131–146.
- Goff, P. (2017a). *Consciousness and Fundamental Reality*. Oxford University Press.
- Goff, P. (2017b). Panpsychism is crazy, but it's also most probably true. <https://aeon.co/ideas/panpsychism-is-crazy-but-its-also-most-probably-true>. Accessed: 2018-10-10.
- Gozzano, S. and Hill, C. S. (2015). *New Perspectives on Type Identity: The Mental and the Physical*. Cambridge University Press.
- Grice, H. P. (1975). Logic and conversation. In *Studies in the Way of Words*, pages 22–40. Harvard University Press.
- Hall, N. (2000). Causation and the price of transitivity. *Journal of Philosophy*, 97(4):198–222.
- Hall, N. (2004). Two concepts of causation. In Collins, J., Hall, N., and Paul, L., editors, *Causation and Counterfactuals*, pages 225–276. MIT Press.
- Halpern, J. and Pearl, J. (2000). *Causes and Explanations: A Structural Model Approach; Technical Report R266*. Los Angeles: Cognitive Systems Laboratory, University of California.
- Handfield, T., Twardy, C. R., Korb, K. B., and Oppy, G. (2008). Where's the biff? *Erkenntnis*, 68(2):149–68.
- Harré, R. and Madden, E. H. (1975). *Causal Powers: A Theory of Natural Necessity*. Blackwell, Oxford.
- Hasker, W. (2010). Persons and the unity of consciousness. In Koons, R. C. and Bealer, G., editors, *The Waning of Materialism: New Essays*, pages 175–190. Oxford University Press.
- Hasker, W. (2014). The dialectic of soul and body. In Lavazza, A. and Robinson, H., editors, *Contemporary Dualism: A Defense*, pages 204–219. Routledge.

- Hattiangadi, A. (2018). Moral supervenience. *Canadian Journal of Philosophy*, 48(3-4):592–615.
- Heil, J. (1992). *The Nature of True Minds*. Cambridge University Press.
- Heil, J. (2013). Mental causation. In Gibb, S. C., Lowe, E. J., and Ingthorson, R. D., editors, *Mental Causation and Ontology*, pages 18–32. Oxford University Press.
- Henne, P., Niemi, L., Pinillos, N. A., Brigard, F. D., and Knobe, J. (forthcoming). A counterfactual explanation for the action effect in causal judgment. *Cognition*.
- Henne, P., Pinillos, A., and Brigard, F. D. (2017). Cause by omission and norm: Not watering plants. *Australasian Journal of Philosophy*, 95(2):270–283.
- Hitchcock, C. (2007). What Russell got right. In Price, H. and Corry, R., editors, *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, pages 44–65. Oxford University Press.
- Hitchcock, C. (2012). Theories of causation and the causal exclusion argument. *Journal of Consciousness Studies*, 19(5-6).
- Hitchcock, C. and Knobe, J. (2009). Cause and norm. *Journal of Philosophy*, 106(11):587–612.
- Hodgson, D. (1991). *The Mind Matters: Consciousness and Choice in a Quantum World*. Oxford University Press.
- Horgan, T. (1987). Supervenient qualia. *Philosophical Review*, 96(October):491–520.
- Horgan, T. (2010). Materialism, minimal emergentism, and the hard problem of consciousness. In Koons, R. C. and Bealer, G., editors, *The Waning of Materialism: New Essays*, pages 309–330. Oxford University Press.
- Horgan, T. and Tienson, J. (2002). The intentionality of phenomenology and the phenomenology of intentionality. In Chalmers, D. J., editor, *Philosophy of Mind: Classical and Contemporary Readings*, pages 502–533. Oxford University Press.

- Huxley, T. (1874). On the hypothesis that animals are automata, and its history. *Fortnightly Review*, 95:555–580.
- Ismael, J. T. (2003). Closed causal loops and the bilking argument. *Synthese*, 136(3):305–320.
- Ismael, J. T. (2016a). How do causes depend on us? the many faces of perspectivalism. *Synthese*, 193(1):245–267.
- Ismael, J. T. (2016b). *How Physics Makes Us Free*. Oxford University Press.
- Jackson, F. (1982). Epiphenomenal qualia. *The Philosophical Quarterly*, 32(127):127–136.
- Jackson, F. (2006). The knowledge argument, diaphanousness, representationalism. In Alter, T. and Walter, S., editors, *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, pages 52–64. Oxford University Press.
- Jackson, F. and Pettit, P. (1990a). Causation and the philosophy of mind. *Philosophy and Phenomenological Research*, 50(n/a):195–214.
- Jackson, F. and Pettit, P. (1990b). Program explanation: A general perspective. *Analysis*, 50(2):107–117.
- Keaton, D. and Polger, T. W. (2014). Exclusion, still not tracted. *Philosophical Studies*, 171(1):135–148.
- Kendler, K. S. and Campbell, J. (2009). Interventionist causal models in psychiatry: repositioning the mind-body problem. *Psychological medicine*, 39(6):881–887.
- Kim, J. (1989). The myth of non-reductive materialism. *Proceedings and Addresses of the American Philosophical Association*, 63(3):31–47.
- Kim, J. (1998). *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. MIT Press.
- Kim, J. (2002). Responses. *Philosophy and Phenomenological Research*, 65(3):671–680.

- Kim, J. (2003). Supervenience, emergence, realization, reduction. In Loux, M. J. and Zimmerman, D. W., editors, *The Oxford Handbook of Metaphysics*. Oxford University Press.
- Kim, J. (2005). *Physicalism, or something near enough*. Princeton University Press.
- Kim, J. (2007). Causation and mental causation. In McLaughlin, B. P. and Cohen, J. D., editors, *Contemporary Debates in Philosophy of Mind*, pages 227–242. Blackwell.
- Kim, J. (2010). *Philosophy of Mind*. Westview Press, Boulder.
- Koons, R. C. and Bealer, G. (2010). Introduction. In Koons, R. C. and Bealer, G., editors, *The Waning of Materialism: New Essays*, pages i–xxxii. Oxford University Press.
- Kriegel, U. (2013). *Phenomenal Intentionality*. Oxford University Press.
- Kroedel, T. (2015). Dualist mental causation and the exclusion problem. *Noûs*, 49(2):357–375.
- Kroedel, T. (2020). *Mental Causation: A Counterfactual Theory*. Cambridge University Press.
- Kroedel, T. and Schulz, M. (2016). Grounding mental causation. *Synthese*, 193(6):1909–1923.
- Kutach, D. (2007). The physical foundations of causation. In Price, H. and Corry, R., editors, *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford University Press.
- Ladyman, J. and Ross, D. (2007). *Every Thing Must Go: Metaphysics Naturalized*. Oxford University Press.
- Lavazza, A. and Robinson, H. (2014). *Contemporary Dualism: A Defense*. Routledge.
- Lee, S. (2016). Occasionalism. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition.

- Leeds, S. (2010). Interventionism in statistical mechanics. <http://philsci-archive.pitt.edu/5464/>. Accessed: 2019-07-07.
- Leibniz, G. W. (1898). *The Monadology and Other Philosophical Writings*. Garland.
- Lepore, E. and Loewer, B. (2011). More on making mind matter. In Lepore, E. and Loewer, B., editors, *Meaning, Mind, and Matter: Philosophical Essays*, pages 179–194. Oxford University Press.
- Levine, J. (2001). *Purple Haze: The Puzzle of Consciousness*. Oxford University Press USA.
- Lewis, D. (1966). An argument for the identity theory. *Journal of Philosophy*, 63(1):17–25.
- Lewis, D. (1973a). Causation. *The Journal of Philosophy*, 70(17):556–567.
- Lewis, D. (1973b). *Counterfactuals*. Oxford University Press.
- Lewis, D. (1976). The paradoxes of time travel. *American Philosophical Quarterly*, 13(2):145–152.
- Lewis, D. (1979). Counterfactual dependence and time’s arrow. *Noûs*, 13(4):455–476.
- Lewis, D. (2000). Causation as influence. *Journal of Philosophy*, 97(4):182–197.
- List, C. and Menzies, P. (2009). Non-reductive physicalism and the limits of the exclusion principle. *The Journal of Philosophy*, 106(9):475–502.
- List, C. and Stoljar, D. (2017). Does the exclusion argument put any pressure on dualism? *Australasian Journal of Philosophy*, 95(1):96–108.
- Loew, C. (2019). Causes as differencemakers for processes. *Philosophy and Phenomenological Research*, 98(1):89–106.
- Loewer, B. M. (1996). Humean supervenience. *Philosophical Topics*, 24(1):101–127.

- Loewer, B. M. (2001). From physics to physicalism. In Gillett, C. and Loewer, B. M., editors, *Physicalism and its Discontents*. Cambridge University Press.
- Loewer, B. M. (2002). Comments on Jaegwon Kim's Mind and the Physical World. *Philosophy and Phenomenological Research*, 65(3):655–662.
- Loewer, B. M. (2007a). Counterfactuals and the second law. In Price, H. and Corry, R., editors, *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, pages 293–326. Oxford University Press.
- Loewer, B. M. (2007b). Mental causation, or something near enough. In McLaughlin, B. P. and Cohen, J. D., editors, *Contemporary Debates in Philosophy of Mind*, pages 243–64. Blackwell.
- Loewer, B. M. (2008). Why there is anything except physics? In Hohwy, J. and Kallestrup, J., editors, *Being Reduced: New Essays on Reduction, Explanation, and Causation*, pages 149–163. Oxford University Press.
- Loewer, B. M. (2012). Two accounts of laws and time. *Philosophical Studies*, 160(1):115–137.
- Loewer, B. M. (2015). Mental causation: The free lunch. In Horgan, T., Sabatés, M., and Sosa, D., editors, *Qualia and Mental Causation in a Physical World*, pages 40–63. Cambridge University Press.
- Lowe, E. J. (1992). The problem of psychophysical causation. *Australasian Journal of Philosophy*, 70(3):263–276.
- Lowe, E. J. (1996). *Subjects of Experience*. Cambridge University Press.
- Lowe, E. J. (1999). Self, agency, and mental causation. *Journal of Consciousness Studies*, 6(8):225–239.
- Lowe, E. J. (2000). Causal closure principles and emergentism. *Philosophy*, 75(294):571–585.
- Lowe, E. J. (2005). Review of Uwe Meixner: the Two Sides of Being: A Reassessment of Psycho-Physical Dualism. *Erkenntnis*, 62(2):290–294.

- Lowe, E. J. (2008). *Personal Agency: The Metaphysics of Mind and Action*. Oxford University Press.
- Lycan, W. G. (2009). Giving dualism its due. *Australasian Journal of Philosophy*, 87(4):551–563.
- Mackie, J. L. (1974). *The Cement of the Universe*. Oxford, Clarendon Press.
- Malcolm, N. (1968). The conceivability of mechanism. *Philosophical Review*, 77(1):45–72.
- Maslen, C. (2005). A new cure for epiphobia: A context-sensitive account of causal relevance. *Southern Journal of Philosophy*, 43(1):131–146.
- Maudlin, T. (2002). *Quantum Non-Locality and Relativity: Metaphysical Intimations of Modern Physics*. Blackwell.
- McDonnell, N. (2017). Causal exclusion and the limits of proportionality. *Philosophical Studies*, 174(6):1459–1474.
- McDonnell, N. (2018). Transitivity and proportionality in causation. *Synthese*, 195(3):1211–1229.
- McGrath, S. (2005). Causation by omission: A dilemma. *Philosophical Studies*, 123(1-2):125–148.
- McLaughlin, B. P. (2010). Consciousness, type-physicalism, and inference to the best explanation. *Philosophical Issues*, 20(1):266–304.
- McLaughlin, B. P. (2015). Does mental causation require psychophysical identities? In Horgan, T., Sabates, M., and Sosa, D., editors, *Qualia and Mental Causation in a Physical World: Essays in Honor of Jaegwon Kim*. Cambridge University Press.
- Meixner, U. (2004). *The Two Sides of Being: A Reassessment of Psychophysical Dualism*. Mentis, Paderborn.
- Meixner, U. (2014). Against physicalism. In Lavazza, A. and Robinson, H., editors, *Contemporary Dualism: A Defense*, pages 17–34. Routledge.
- Mele, A. R. (1992). *Springs of Action: Understanding Intentional Behavior*. Oxford University Press.

- Mellor, D. H. (1995). *The Facts of Causation*. Routledge.
- Mendelovici, A. (2018). *The Phenomenal Basis of Intentionality*. Oxford University Press.
- Menzies, P. (2013). Mental causation in the physical world. In Gibb, S., Lowe, E. J., and Ingthorsson, R., editors, *Mental Causation and Ontology*, pages 58–86. Oxford University Press.
- Menzies, P. and Price, H. (1993). Causation as a secondary quality. *British Journal for the Philosophy of Science*, 44(2):187–203.
- Mill, J. S. (1843). *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence, and the Methods of Scientific Investigation*. Longmans, Green, Reader, and Dyer.
- Molenaar, P. C. M. (2006). Psychophysical dualism from the point of view of a working psychologist. *Erkenntnis*, 65(1):47–69.
- Montague, M. (2016). *The Given: Experience and its Content*. Oxford University Press.
- Montero, B. G. and Brown, C. D. (2018). Making room for a this-worldly physicalism. *Topoi*, 37(3):523–532.
- Montminy, M. and Russo, A. (2016). A defense of causal invariantism. *Analytic Philosophy*, 57(1):49–75.
- Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, 83(October):435–50.
- Ney, A. (2008). Defining physicalism. *Philosophy Compass*, 3(5):1033–1048.
- Ney, A. (2009). Physical causation and difference-making. *British Journal for the Philosophy of Science*, 60(4):737–764.
- Ney, A. (2012). The causal contribution of mental events. In Christopher, H. and Simone, G., editors, *New Perspectives on Type Identity: The Mental and the Physical*, pages 230–250. Cambridge University Press.

- Nolan, D. (2017). Causal counterfactuals and impossible worlds. In Beebe, H., Hitchcock, C., and Price, H., editors, *Making a Difference*, pages 14–32. Oxford University Press.
- Norton, J. D. (2007). Causation as folk science. In Price, H. and Corry, R., editors, *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford University Press.
- Nowak, E. (forthcoming). No context, no content, no problem. *Mind and Language*.
- O'Connor, T. (2000). *Persons and Causes: The Metaphysics of Free Will*. Oxford University Press.
- Papineau, D. (1993). *Philosophical Naturalism*. Blackwell.
- Papineau, D. (2001). The rise of physicalism. In Gillett, C. and Loewer, B. M., editors, *Physicalism and its Discontents*, pages 3–36. Cambridge University Press.
- Papineau, D. (2002). *Thinking about Consciousness*. Oxford University Press.
- Papineau, D. (2013). Causation is macroscopic but not irreducible. In Gibb, S., Lowe, E. J., and Ingthorsson, R., editors, *Mental Causation and Ontology*, pages 126–151. Oxford University Press.
- Paul, L. A. (2000). Aspect causation. *Journal of Philosophy*, 97(4):235.
- Paul, L. A. (2006). Coincidence as overlap. *Noûs*, 40(4):623–659.
- Paul, L. A. (2007). Constitutive overdetermination. In Campbell, J. K., O'Rourke, M., and Silverstein, H. S., editors, *Causation and Explanation*, pages 265–290. MIT Press.
- Pautz, A. (2010). A simple view of consciousness. In Koons, R. C. and Bealer, G., editors, *The Waning of Materialism*, pages 25–66. Oxford University Press.
- Peacocke, C. (1979). *Holistic Explanation: Action, Space, Interpretation*. Clarendon Press, Oxford.

- Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.
- Pereboom, D. (2002). Robust non-reductive materialism. *Journal of Philosophy*, 99(10):499–531.
- Pernu, T. K. (2013). The principle of causal exclusion does not make sense. *Philosophical Forum*, 44(1):89–95.
- Pernu, T. K. (2016). Causal exclusion and downward counterfactuals. *Erkenntnis*, 81(5):1031–1049.
- Pietroski, P. (1994). Mental causation for dualists. *Mind and Language*, 9(3):336–366.
- Polger, T. W., Shapiro, L. A., and Stern, R. (2018). In defense of interventionist solutions to exclusion. *Studies in History and Philosophy of Science Part A*, 68:51–57.
- Popper, K. R. and Eccles, J. C. (1977). *The Self and Its Brain: An Argument for Interactionism*. Springer.
- Price, H. (2007). Causal perspectivalism. In Price, H. and Corry, R., editors, *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, pages 250–292. Oxford University Press.
- Price, H. and Corry, R. (2007). *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford University Press.
- Price, H. and Weslake, B. (2009). The time-asymmetry of causation. In Beebe, H., Menzies, P., and Hitchcock, C., editors, *The Oxford Handbook of Causation*, pages 414–443. Oxford University Press.
- Putnam, H. (1973). Meaning and reference. *Journal of Philosophy*, 70(19):699–711.
- Putnam, H. (1975). The nature of mental states. In *Mind, Language, and Reality*. Cambridge University Press.
- Putnam, H. (1982). Why there isn't a ready-made world. *Synthese*, 51(2):205–228.

- Raatikainen, P. (2010). Causation, exclusion, and the special sciences. *Erkenntnis*, 73(3):349–363.
- Raatikainen, P. (2013). Can the mental be causally efficacious? In Milkowski, K. and Talmont-Kaminski, M., editors, *Regarding the Mind, Naturally: Naturalist Approaches to the Sciences of the Mental*, pages 138–166. Cambridge Scholars Publishing.
- Raatikainen, P. (2018). Kim on causation and mental causation. *E-Logos Electronic Journal for Philosophy*, 25(2):22–47.
- Reutlinger, A. (2013). Can interventionists be Neo-Russellians? Interventionism, the open systems argument, and the arrow of entropy. *International Studies in the Philosophy of Science*, 27(3):273–293.
- Robb, D. (forthcoming). Could mental causation be invisible? In Carruth, A., Gibb, S. C., and Heil, J., editors, *The Metaphysics of E.J. Lowe*. Oxford University Press.
- Robinson, H. (2017). Dualism. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2017 edition.
- Robinson, W. S. (1982). Causation, sensations, and knowledge. *Mind*, 91(October):524–40.
- Robinson, W. S. (1988). *Brains and People: An Essay on Mentality and its Causal Conditions*. Temple University Press.
- Robinson, W. S. (2006). Knowing epiphenomena. *Journal of Consciousness Studies*, 13(1-2):85–100.
- Robinson, W. S. (2015). Epiphenomenalism. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2015 edition.
- Robinson, W. S. (2018). *Epiphenomenal Mind: An Integrated Outlook on Sensations, Beliefs, and Pleasure*. Routledge.

- Rudder Baker, L. (1993). Metaphysics and mental causation. In Heil, J. and Mele, A. R., editors, *Mental Causation*, pages 75–96. Oxford University Press.
- Russell, B. (1912). On the notion of cause. *Proceedings of the Aristotelian Society*, 7:1–26.
- Russo, A. (2016). Kim’s dilemma: Why mental causation is not productive. *Synthese*, 193(7):2185–2203.
- Salmon, W. (1971). *Statistical Explanation & Statistical Relevance*. University of Pittsburgh Press.
- Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton University Press.
- Schaffer, J. (2000). Causation by disconnection. *Philosophy of Science*, 67(2):285–300.
- Schaffer, J. (2001a). Causation, influence, and effluence. *Analysis*, 61(1):11–19.
- Schaffer, J. (2001b). Causes as probability raisers of processes. *Journal of Philosophy*, 98(2):75–92.
- Schaffer, J. (2003). Overdetermining causes. *Philosophical Studies*, 114(1–2):23–45.
- Schaffer, J. (2004). Causes need not be physically connected to their effects: The case for negative causation. In Hitchcock, C. R., editor, *Contemporary Debates in Philosophy of Science*, pages 197–216. Basil Blackwell.
- Schaffer, J. (2005). Contrastive causation. *Philosophical Review*, 114(3):327–358.
- Schaffer, J. (2010). Causation, Physics, and the Constitution of Reality: Russell’s Republic Revisited, Edited by Huw Price and Richard Corry.: Book Reviews. *Mind*, 119(475):844–848.
- Schaffer, J. (2012). Causal contextualisms. In Blaauw, M., editor, *Contrastivism in Philosophy: New Perspectives*, pages 43–71. Routledge.

- Schaffer, J. (2016). The metaphysics of causation. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2016 edition.
- Sellars, W. (1981). Is consciousness physical? *The Monist*, 64(1):66–90.
- Shapiro, L. (2010). Lessons from causal exclusion. *Philosophy and Phenomenological Research*, 81(3):594–604.
- Shapiro, L. and Sober, E. (2007). Epiphenomenalism. The do’s and don’ts. In Wolters, G. and Machamer, P., editors, *Thinking about causes: From Greek philosophy to modern physics*, pages 235–264. University of Pittsburgh Press.
- Shapiro, L. and Sober, E. (2012). Against proportionality. *Analysis*, 72(1):89–93.
- Shoemaker, S. (2007). *Physical Realization*. Oxford University Press.
- Sider, T. (2003). Review: What’s so bad about overdetermination? *Philosophy and Phenomenological Research*, 67(3):719–726.
- Smart, J. J. C. (1959). Sensations and brain processes. *The Philosophical Review*, 68(2):141–156.
- Smart, J. J. C. (2017). The mind/brain identity theory. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2017 edition.
- Stapp, H. P. (1993). *Mind, Matter, and Quantum Mechanics*. Springer Verlag.
- Stapp, H. P. (2001). Quantum theory and the role of mind in nature. *Foundations of Physics*, 31(10):1465–1499.
- Stapp, H. P. (2013). Quantum theory of mind. In Lavazza, A. and Robinson, H., editors, *Contemporary Dualism, A Defense*, pages 98–111. Routledge.
- Stapp, H. P. (2014). Quantum physics and philosophy of mind. In Meixner, U. and Corradini, A., editors, *Quantum Physics Meets the Philosophy of Mind: New Essays on the Mind-Body Relation in Quantum-Theoretical Perspective*, pages 5–16. De Gruyter.

- Statham, G. (2017). The manipulation of chemical reactions: Probing the limits of interventionism. *Synthese*, 194(12):4815–4838.
- Stern, R. (2017). Interventionist decision theory. *Synthese*, 194(10):4133–4153.
- Stern, R. (2019). Decision and intervention. *Erkenntnis*, 84(4):783–804.
- Stoljar, D. (2008a). Distinctions in distinction. In Kallestrup, J. and Hohwy, J., editors, *Being Reduced: New Essays on Causation and Explanation in the Special Sciences*, pages 263–279. Oxford University Press.
- Stoljar, D. (2008b). *Ignorance and Imagination*. Oxford University Press.
- Stoljar, D. (2010). *Physicalism*. Routledge.
- Stoljar, D. (2016). Physicalism. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2016 edition.
- Strawson, G. (1987). Realism and causation. *Philosophical Quarterly*, 37(148):253–277.
- Sundström, P. (2018). How physicalists can—and cannot—explain the seeming “absurdity” of physicalism. *Philosophy and Phenomenological Research*, 97(3):681–703.
- Swinburne, R. (1986). *The Evolution of the Soul*. Oxford University Press.
- Tang, Z. (2015). Absence causation and a liberal theory of causal explanation. *Australasian Journal of Philosophy*, 93(4):688–705.
- Usher, M. (Forthcoming). Agency, teleological control and robust causation. *Philosophy and Phenomenological Research*.
- Vaassen, B. (2019). Dualism and exclusion. *Erkenntnis*, pages 1–10.
- von Wright, G. H. (1971). *Explanation and Understanding*. Cornell University Press.
- Walter, S. (2010). Taking realization seriously: No cure for epiphobia. *Philosophical Studies*, 151(2):207–226.

- Weslake, B. (2011). Exclusion excluded. http://bweslake.s3.amazonaws.com/research/papers/weslake_exclusion.pdf. Accessed: 2019-02-10.
- White, B. (2018). Metaphysical necessity dualism. *Synthese*, 195(4):1779–1798.
- Willemsen, P. and Kirfel, L. (2019). Recent empirical work on the relationship between causal judgements and norms. *Philosophy Compass*, 14(1):e12562.
- Willemsen, P. and Reuter, K. (2016). Is there really an omission effect? *Philosophical Psychology*, 29(8):1142–1159.
- Williamson, T. (2000). *Knowledge and its Limits*. Oxford University Press.
- Williamson, T. (2016). Counterpossibles. *Topoi*, pages 1–12.
- Wilson, J. M. (2009). Determination, realization and mental causation. *Philosophical Studies*, 145(1):149–169.
- Won, C. (2014). Overdetermination, counterfactuals, and mental causation. *Philosophical Review*, 123(2):205–229.
- Woodward, J. (2003). *Making Things Happen: a Theory of Causal Explanation*. Oxford University Press.
- Woodward, J. (2006). Sensitive and insensitive causation. *Philosophical Review*, 115(1):1–50.
- Woodward, J. (2007). Causation with a human face. In Price, H. and Corry, R., editors, *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, pages 66–105. Oxford University Press.
- Woodward, J. (2008). Mental causation and neural mechanisms. In Kallestrup, J. and Hohwy, J., editors, *Being Reduced*, pages 218–263. Oxford University Press.
- Woodward, J. (2009). Agency and interventionist theories. In Beebe, H., Hitchcock, C., and Menzies, P., editors, *The Oxford Handbook of Causation*. Oxford University Press.
- Woodward, J. (2010). Causation in biology: Stability, specificity, and the choice of levels of explanation. *Biology and Philosophy*, 25(3):287–318.

- Woodward, J. (2014). A functional account of causation; or, A defense of the legitimacy of causal thinking by reference to the only standard that matters—usefulness. *Philosophy of Science*, 81(5):691–713.
- Woodward, J. (2015). Interventionism and causal exclusion. *Philosophy and Phenomenological Research*, 91(2):303–347.
- Woodward, J. (2017). Explanation in neurobiology: An interventionist perspective. In Kaplan, D. M., editor, *Explanation and Integration in Mind and Brain Science*, pages 70–100. Oxford University Press.
- Woodward, J. (2018). Explanatory autonomy: The role of proportionality, stability, and conditional irrelevance. *Synthese*, pages 1–29.
- Woolhouse, R. S. (1985). IV—Leibniz’s reaction to Cartesian interaction. *Proceedings of the Aristotelian Society*, 86(1):69–82.
- Yablo, S. (1992). Mental causation. *Philosophical Review*, 101(2):245–280.
- Yablo, S. (2002). De facto dependence. *Journal of Philosophy*, 99(3):130–148.
- Yang, E. (2013). Eliminativism, interventionism and the overdetermination argument. *Philosophical Studies*, 164(2):321–340.
- Zhong, L. (2019). Taking emergentism seriously. *Australasian Journal of Philosophy*, pages 1–16.
- Zwier, K. R. (2017). Interventionist causation in thermodynamics. *Philosophy of Science*, 84(5):1303–1315.

Sammanfattning

Syftet med avhandlingen är att utveckla och försvara en förståelse av kausalitet enligt vilken mentala fenomen kan orsaka fysikaliska fenomen också givet dualism om det mentala och givet att den fysikaliska domänen är fysikaliskt komplett.

I den första delen definierar jag den dualistiska teori som avhandlingen utgår ifrån och identifierar två problem för mental kausalitet som teorin behöver hantera: uteslutningsproblemet och problemet med gemensamma orsaker. Vidare argumenterar jag för att en lösning av dessa problem kräver ett fullständigt lättviktigt kausalitetsbegrepp: ett som tillåter att en orsak till en effekt kan vara skild från fenomenen som producerar eller fysikaliskt nödvändiggör den aktuella effekten.

I den andra delen utvärderar jag i litteraturen förekommande svar på dessa utmaningar: (i) List och Stoljars svar på utelutningsproblemet, (ii) Kroedels alternativa dualistiska ontologi, (iii) en utbredd kritik av idén om 'kausal tillräcklighet', och (iv) Lowes modeller av dualistisk mental kausalitet. Jag argumenterar för att inget av dessa förslag är helt tillfredsställande.

I de första fyra kapitlen av tredje delen utvecklar jag sedan en fullständigt lättviktig kausalitetsmodell som bygger på interventionistiska modeller av kausalitet. Först förklarar jag hur så kallade 'konstanskrav' i gängse interventionism utesluter dualistiskt mental kausalitet. Därefter argumenterar jag för att interventionistiska modeller bör inkludera ett 'robusthetskrav' på kausala korrelationer, och att det finns skäl att göra undantag för de aktuella 'konstanskraven'. Jag kallar den interventionistiska modell som detta utmynnar i 'okänslig interventionism'. Modellen löser både uteslutningsproblemet och problemet med gemensamma orsaker, och medger att dualistisk mental

kausalitet kan förekomma under vissa omständigheter, vilka såvitt vi vet kan råda i vår värld.

I de tre avslutande kapitlen försvarar jag okänslig interventionism mot invändningar. Jag granskar invändningen att kausalitet måste vara produktiv, invändningen att orsaker måste nödvändiggöra sina effekter, och invändningen att okänslig interventionism är för tillåtande. Mina svar baseras på tidigare forskning om hur frånvaron av ett fenomen kan orsaka något och om förhållandet mellan kausalitet och fysik. Sammanfattningsvis påstår jag att okänslig interventionism kan ge lika tillfredsställande svar på de tre invändningarna som gängse interventionistiska modeller. Jag drar slutsatsen att okänslig interventionism är en lovande modell för kausalitet.

UMEÅ STUDIES IN PHILOSOPHY

Department of Historical, Philosophical and Religious Studies
Umeå University, Sweden

The series UMEÅ STUDIES IN PHILOSOPHY consists mainly of research communications (monographs, dissertations and collections of essays) in philosophy from the Department of Historical, Philosophical and Religious Studies at Umeå University. The series is not distributed through the bookstores, but individual volumes can be obtained from the department. Orders should be sent to Umeå Studies in Philosophy, Department of Historical, Philosophical and Religious Studies, Umeå University, SE-901 87 Umeå, Sweden.

1. JONAS NILSSON: *Rationality in Inquiry: On the Revisability of Cognitive Standards*, 2000.
2. PER NILSSON: *Naturen, vetenskapen & förnuftet: Upplýsnings dialektik och det andra moderna*, 2001.
3. JAYNE M. WATERWORTH: *Living in the Light of Hope: An Investigation into Agency and Meaning*, 2001.
4. ANDERS ODENSTEDT: *Cognition and Cultural Context: An Inquiry Into Gadamer's Theory of Context-Dependence*, 2001.
5. RÖGNVALDUR INGTHORSSON: *Time, Persistence, and Causality: Towards a Dynamic View of Temporal Reality*, 2002.
6. BENGT LILIEQUIST: *Ludwik Flecks jämförande kunskapssteori*, 2003.
7. PETER NILSSON: *Empathy and Emotions: On the Notion of Empathy as Emotional Sharing*, 2003.
8. ANDERS BERGLUND: *From Conceivability to Possibility: An Essay in Modal Epistemology*, 2005.

9. LARS SAMUELSSON: *The Moral Status of Nature: Reasons to Care for the Natural World*, 2008.
10. EBBA GULLBERG: *Objects and Objectivity: Alternatives to Mathematical Realism*, 2011.
11. JESPER ÖSTMAN: *It's All in the Brain: A Theory of the Qualities of Perception*, 2013.
12. EMMA BECKMAN: *Mistaken Morality? An Essay on Moral Error Theory*, 2018
13. BRAM VAASSEN: *Causal After All: A Model of Mental Causation for Dualists*, 2019