Interventionism and the Exclusion Problem

By Yasmin Bassi

Thesis submitted for the degree of Doctor of Philosophy

University of Warwick
Department of Philosophy
June 2013

Table of Contents

1. Introduction 1
1.1 Background 1
1.1.1 The Current Problem 9
1.1.2 Aim of Thesis 11
1.1.2.1 Why Interventionism? 12
1.2 Outline of Thesis 16

2. The Exclusion Problem and Non-Reductive Physicalism 21
2.1 Introduction 21
2.2 Commitments of Non-Reductive Physicalism: Causal Closure 24
2.2.1 What Does 'Physical' Mean? 25
2.2.1.1 An Argument Against Microphysical Reduction 28
2.2.1.2 The 'Object View' 30
2.2.1.3 The 'Via Negativa' View 32
2.2.2 Is Causal Closure True? 34
2.2.2.1 Implications for Non-Reductive Physicalism 40
2.3 Commitments of Non-Reductive Physicalism: Supervenience 42
2.3.1 Weak Supervenience 43
2.3.2 Strong Supervenience 47
2.3.2.1 The Implications of SS*mn* 49
2.3.2.2 Some Potential Problems with SS*mn* 51
2.4 Conclusion 55

3. The Exclusion Problem and Its Assumptions 57
3.1 Introduction 57
3.2 The SP Concept of Causation 58
3.2.1 Kim and the Assumption of SP 61
3.3 The Assumption of SP and the Exclusion Problem 65
3.3.1 The Assumption of SP and the Exclusion Principle 71
3.3.2 Overdetermination: Some Further Issues 80
3.3.2.1 Bennett's Argument against Overdetermination 81
3.3.2.2 Exclusion All Over Again 86
3.4 Conclusion 96

4. Interventionism 98
4.1 Introduction 98
4.2 Interventionism Outlined and Clarified 100
4.2.1 Invariance 106
4.2.2 Interventions 116
4.2.3 Causal Explanation 121
4.3 Interventionism versus the SP Concept of Causation: Problems for the SP Concept 131
4.3.1 Is Sufficient Production Necessary for Causation? 133
4.3.2 Is Sufficient Production Sufficient for Causation? 145
4.4 Interventionism versus the SP Concept of Causation: Problems for Interventionism 149
4.4.1 Overdetermination 150
4.4.2 Non-Paradigmatic Causation and Causation by Omissions (and insensitivity as a solution to these problems) 155
4.5 Remaining Problems 168
4.5.1 The Problem of Anthropocentricism 168
4.5.2 The Problem of Realism 170
4.5.3 The Problem of Circularity 175
4.6 Conclusion 177

5. Interventionism and Mental Causation 179
5.1 Introduction 179
5.2 An Interventionist Account of Mental Causation 181
5.2.1 Invariance and Realization Independent Dependency Relations (RIDR) 184
5.2.2 Contrastive Focus 197
5.3 A Solution to the Exclusion Problem 202
5.3.1 A *Physicalist* Solution? 205
5.3.2 A *Satisfactory* Solution? 212
5.4 Alternative Manipulationist Accounts of Mental Causation and the Problem of Realism 217
5.4.1 List and Menzies and the Problem of Realism 220
5.4.2 Campbell and the Problem of Realism 232
5.5 Conclusion 240

6. Interventionist Causal Exclusion and the Underdetermination Argument 242
6.1 Introduction 242
6.2 The Interventionist Exclusion Argument 243

6.2.1 The Interventionist Solution to the Interventionist Exclusion Problem 251
6.2.1.1 Modifying (M) and (IV) 261
6.2.1.2 Some Further Worries 265
6.3 The Underdetermination Argument 266
6.4 Conclusion 280

7. Conclusion 282
7.1 Summary 282
7.2 Implications for Mental Causation 286

Bibliography 290

List of Abbreviations 302

List of Illustrations

Figure 3.1: Exclusion ……71

Figure 5.1: RIDR/Intention I_1 ……190

Figure 5.2: Non-RIDR/Fear ……..192

Figure 5.3: Minimal RIDR/Positive thinking ……194

Figure 5.4: The Scale of Invariance and RIDR ……..195

Figure 6.1: Supervenient Mental Causation ……..247

Figure 6.2: Cholesterol ……..256

Figure 5.2: Non-RIDR/Fear ……..…….270

Figure 5.1: RIDR/Intention I_1 ……...271

Acknowledgments

Many thanks to my supervisor, Christoph Hoerl, for providing invaluable feedback and guidance that has helped to develop this thesis. My thanks also go to Johannes Roessler and Bill Brewer, who I was fortunate enough to be supervised by for a period of time. Many thanks also to Naomi Eilan, Matthew Soteriou, Guy Longworth, Peter Poellner and Michael Baumgartner and many fellow graduate students in the philosophy department for helpful comments and discussions.

My greatest thanks go to my family: to my parents for providing endless support throughout the many years of my education. To my sisters, my cousin and to my friends for providing support and often welcome distractions over the years. To Neal, for providing endless support and tireless encouragement throughout this entire process. Finally, to my late grandfather, for making me passionate about philosophy before I knew what philosophy was.

Declaration

This thesis is my own work and does not include any collaborative research. This thesis has not been submitted for a degree at another university.

Abstract

Jaegwon Kim (1998a, 2005) claims that his exclusion problem follows a priori for the non-reductive physicalist given her commitment to five apparently inconsistent theses: mental causation, non-identity, supervenience, causal closure and non-overdetermination. For Kim, the combination of these theses entails that mental properties are a priori excluded as causes, forcing the non-reductive physicalist to accept either epiphenomenalism, or some form of reduction. In this thesis, I argue that Kim's exclusion problem depends on a particular conception of causation, namely sufficient production, and that when causation is understood in interventionist terms, the non-reductive physicalist can avoid the exclusion problem. I argue that Woodward's (2003, 2008a, 2011a) version of interventionism not only provides an account of mental causation that avoids the exclusion problem, but argue that it also upholds all of the minimal commitments of non-reductive physicalism, thereby providing a successful non-reductive *physicalist* solution to the exclusion problem.

In Chapter 2, I argue that all five theses are minimal commitments of non-reductive physicalism that cannot be rejected in order to avoid the exclusion problem. Chapter 3 identifies the assumptions that I take to underlie the exclusion problem. Chapter 4 introduces and outlines the central features of Woodward's (2003) interventionism and Chapter 5 argues that Woodward's interventionist account of mental causation provides a solution to the exclusion problem. I examine two alternative interventionist accounts of mental causation[1] that fail to provide *satisfactory* solutions to the exclusion problem and conclude that Woodward's account therefore provides the *only* satisfactory account of mental causation and solution to the exclusion problem. Chapter 6 addresses some challenges proposed by Michael Baumgartner (2009, 2010) and argues that the interventionist is able to defend her position against these objections and uphold the interventionist solution to the exclusion problem outlined in this thesis.

[1] Proposed by List and Menzies (2009) and Campbell (2007, 2008a, 2008b, 2010).

1. Introduction

1.1 Background

Since Descartes and the emergence of the modern 'mind-body' problem, questions surrounding the causal role of the mental have been commonplace in the philosophical literature. The problem of mental causation for Descartes is well known: if the mental and the physical are *distinct* substances, where the former is immaterial and un-extended and the latter is material and spatially extended, how is causal interaction between such distinct substances possible? It is widely accepted that the thesis of causal closure - the thesis that every physical effect has a sufficient physical cause - makes Cartesian dualism an untenable thesis. However, as we will see, the problem of mental causation did not disappear with dualism, but rather, a new set of problems emerged that are still widely debated today.

With the thesis of causal closure accepted as part of common scientific understanding by the middle of the twentieth century, physicalism - the view that there are *no* non-physical entities or substances and that everything[1] is either

[1] In Chapter 2, I add a caveat to this definition of physicalism, since as Papineau (2001) points out, the formulation of physicalism that follows from the thesis of causal closure does not necessarily entail the view that *everything* is physical, but need only entail the view that everything that has a physical effect must be identified with, or supervene on the physical. This is because the thesis of causal closure states that every physical *effect* has a physical *cause* and hence leaves open the possibility that there may be non-physical properties, such as mathematical properties, so long as those properties do not have any physical effects.

physical, or else is dependent upon the physical - became the received view[2]. It was the work of those such as Place (1956), Feigl (1958), Smart (1959), Putnam (1960), Davidson (1963), Lewis (1966) and Armstrong (1968) (Papineau, 2001) that contributed to physicalism becoming the near orthodox position in philosophy that it is today.

According to Papineau (Ibid), this emergence of physicalism in the 1950's and 1960's can be attributed to the fact that the thesis of causal closure became available as a key premise in the arguments for physicalism at that time. Although there are different formulations of physicalism that follow from the thesis of causal closure, each of which have different implications for mental causation, I focus on the theory of physicalism as it follows from the Causal Argument, advocated, for instance, by Papineau (2004). Moreover, following recent discussion on mental causation, I refer, for the most part, to causation as between properties, or more accurately, property instantiations (more on this below), rather than, for example, as between events or states. With this in mind, the Causal Argument for physicalism can be formulated as follows:

1. Mental properties have physical effects.
2. All physical effects have sufficient physical causes.
3. The physical effects of mental properties aren't always overdetermined by metaphysically distinct, sufficient causes.

Conclusion: Mental properties are identical to, or supervene on physical properties.

[2] I discuss the thesis of causal closure and its specific implications for physicalism in further detail in Chapter 2.

As an illustration, consider a paradigmatic case of mental causation: my conscious desire for a cup of tea causes me to reach for the kettle. Now, according to causal closure, this physical effect, namely my reaching for the kettle, already has a sufficient physical cause, which by definition is enough to bring about the occurrence of the effect. If we then want to avoid the systematic overdetermination of physical effects (by two metaphysically distinct, sufficient causes) we must either identify mental causes with physical causes or accept the supervenience of mental causes on physical causes, hence the physicalist conclusion of the Causal Argument.

Now, as it is formulated above, the Causal Argument generates two broadly physicalist conclusions, one being that mental properties are *identical* to physical properties and the other being that mental properties *supervene*[3] on physical properties. Before exploring these two physicalist positions further, it is important to emphasise that what is common to *all* physicalist positions is their physicalist ontology: all physicalists hold that every concrete particular and entity is physical, therefore disavowing the existence of disembodied souls, spirits, and so on. As we shall see, the difference between these two physicalist positions concerns the question of whether there can be genuinely distinct, non-physical, e.g. mental, *properties*.

The first of the physicalist positions, namely *reductive* physicalism, holds that there are no distinctively mental properties. There may of course be mental predicates and mental levels of description, but these do not correspond to mental properties, but correspond instead to physical properties (Baker, 2009: 2).

[3] I discuss the notion of supervenience below and in detail in Chapter 2.

According to reductive physicalists, mental property types, such as beliefs, intentions and desires are *identical* and *reducible* to physical property types. What does reduction in this context amount to? There are a variety of forms of reduction discussed in the literature[4], but for the purposes of this thesis, we may simply appeal to the notion of explanatory reduction, which involves the idea that a set of properties, *S*, reduces to another set of properties, *P*, if *S* can be exhaustively explained in terms of *P*.

We can then see that the issue of mental causation under reductive physicalism becomes fairly straightforward: if mental properties just *are* physical properties, then it is no wonder that they can have physical effects. However, in giving up on the idea that there are distinct mental properties that can have distinct causal roles in relation to physical effects, the reductive physicalist essentially gives up on mental causation; under reductive physicalism, mental causation just collapses into physical causation.

For many, reductive physicalism is too strong. If reductive physicalism entails giving up on the idea that what we think can and does have a real and distinct effect in the physical world, then for many, this is so much the worse for reductive physicalism. The second physicalist position that can be generated from the Causal Argument is therefore *non-reductive* physicalism, which upholds the physicalist ontology, whilst positing the existence of distinct mental properties that can have distinct causal roles in the physical world.

[4] For instance, as Daniel Stoljar (Fall 2009) points out, reduction is often taken to involve the idea that one theory reduces to another if it is possible to logically derive the first from the second with appropriate 'bridge laws' (see, for example, Nagel (1961)). Alternatively, reduction is often taken to involve the idea that "the properties expressed by the predicates of (say) a psychological theory are identical to the properties expressed by the predicates of (say) a neurological theory." (Stoljar, Fall 2009).

It was the multiple realization arguments of Putnam (1975a) in particular[5] that provided the support for non-reductive physicalism. Very roughly, these arguments suggest that the idea that mental property types are identical to physical property types is implausible, given that it seems possible for the same mental property type to be 'realized'[6] by a wide variety of physical property types. For example, it seems possible that both a human being and an octopus could share the property of being in pain, while it is unlikely that the physical properties that 'realize' the property of being in pain in those animals share any common physical features. These arguments suggested that it is not possible to identify a mental property type with one physical property type and hence supported the existence of irreducible and hence *distinct* mental properties.

However, multiple realization arguments on their own are not sufficient for non-reductive physicalism. This is because even if one thinks that multiple realization makes reductive physicalism implausible and that there are therefore distinct mental properties, those mental properties may simply be mere epiphenomena, i.e. contrary to our intuitions, mental properties do not actually causally influence anything. Moreover, if they are mere epiphenomena then this would seem to undermine the idea that they are genuinely distinct; if mental properties do not contribute anything new in addition to physical properties, then in what sense can they be considered as 'real', distinct properties? Central to non-reductive physicalism then is the idea that if mental properties are to be

[5] The work of Fodor (1974), Nagel (1974) and Jackson (1986) was also influential in the rise of non-reductive physicalism.
[6] I discuss the notion of realization below.

considered as genuine and irreducible features of the world, they must have genuine and irreducible causal powers.[7] As Kim puts it,

> "For unless mentality made causal contributions that are genuinely novel, the claim that it is a distinct and irreducible phenomenon over and beyond physical-biological phenomena would be hollow and empty. To be real, Alexander has said, is to have causal powers; *to be real, new, and irreducible, therefore, must be to have new, irreducible causal powers.*" (Kim, 2003b: 203-204)

If these distinct and irreducible properties are not identical to physical properties then what relationship do they bear to the physical that respects the physicalist ontology, whilst avoiding reduction? The answer for most non-reductive physicalists is some form of supervenience. I examine the thesis of supervenience in detail in Chapter 2 and establish exactly which form of supervenience the non-reductive physicalist is committed to, but for now, it will be helpful to point out that all forms of supervenience that are consistent with physicalism entail the *metaphysical dependence* of mental properties on physical properties. (As it is often put, supervenience entails that there can be no difference at the mental level without a difference at the physical level.)

[7] This view contrasts with that of Shapiro (2010, 2011) and Shapiro and Sober (2007), for example, who argue that mental properties can be considered as real and irreducible causes of physical effects, in addition to their physical realizers, even though it turns out that on their account of causation, the causal powers of mental properties are identical and hence reducible to those of their physical realizers. I will not provide an independent argument for it here, but it is a plausible assumption that in order for mental properties to be considered as real and irreducible features of the world, they must have real and irreducible causal powers. Given that Shapiro and Sober accept that the causal powers of mental properties are identical to those of their physical realizers, it is therefore fair to assume that their account of causation would fail to provide a genuine *non-reductive* physicalist solution to the exclusion problem. I will return to these issues later in the thesis (especially in Chapters 5 and 6).

In virtue of what does the supervenience relationship hold? The idea that mental properties are 'realized' by physical properties is an idea that gained popularity with the multiple realization arguments of the 1960's, but what does realization entail? For now, we can simply think of 'realization' as synonymous with 'instantiated', or 'implemented' (Kim, 2003b: 194) and on this reading, it is plausible to assume that physical realization *entails* supervenience (Ibid: 195). The specific nature of this 'instantiation', in particular, the modal force with which it holds, will become clearer when we look at supervenience in more detail in Chapter 2, but for now, we can appeal to the idea that mental properties supervene on and are realized by physical properties in the sense described above.[8] Thus, supervenience (and realization) allow the non-reductive physicalist to uphold the physicalist ontology, whilst making room for irreducible and distinct mental properties. Kim sums up these physicalist commitments in the following passage:

> "Stated as a thesis about properties, physical primacy in this sense comes to this: all mental properties are instantiated in physical particulars. Thus, although there can be, and presumably are, objects and events that have only physical properties, there can be none that have only mental properties alone; mentality must be instantiated in physical systems." (Ibid: 193)

[8] Although it was the theory of functionalism that first emerged from the multiple realization arguments against reduction, it need not be assumed that all non-reductive physicalists who endorse supervenience and multiple realization are thereby committed to functionalism. The discussion in this thesis is neutral as to whether mental properties are purely 'functional' properties.

What then are the key commitments of non-reductive physicalism? As I discuss in Chapter 2, non-reductive physicalists can be defined as minimally committed to the following five theses:

1. Mental causation, the thesis that mental properties have physical effects.
2. Non-identity, the thesis that mental properties are not identical to physical properties.
3. Supervenience, the thesis that mental properties supervene on physical properties.
4. Causal closure, the thesis that every physical effect has a sufficient physical cause.
5. Non-overdetermination, the thesis that the effects of mental causes are not systematically overdetermined by two metaphysically distinct, sufficient causes.

I will not discuss the thesis of non-overdetermination in detail until Chapter 3, but for now, it is worth pointing out why overdetermination would not be a plausible model for mental causation. Consider a classic case of overdetermination: two riflemen each shoot a victim simultaneously, causing his death. In this case, the effect, (namely the death of the victim), would still have occurred if the first rifleman had failed to fire and vice versa for the second rifleman. Although possible, it seems that such cases of overdetermination would be extremely rare. The idea then that the physical effects of mental causes are routinely overdetermined in this way, i.e. frequently caused 'twice over' by two

metaphysically distinct causes, each of which is sufficient for its occurrence, seems highly unlikely and implausible.

1.1.1 The Current Problem

As a theory that is physicalist in its ontology, whilst positing the existence of irreducible and *distinct* mental properties, thereby making room for the possibility of mental causation, non-reductive physicalism is the favoured view among physicalists today. Ironically though, it is out of non-reductive physicalism that the current problem concerning mental causation emerged. This is the exclusion problem, advocated most widely by Jaegwon Kim (1998a, 2005). So, what exactly is the exclusion problem?

As an illustration, consider again the apparently paradigmatic example of mental causation introduced above: my conscious desire for a cup of tea causes a physical effect, namely my reaching for the kettle. Now, the problem according to Kim is that the thesis of causal closure states that this physical effect already has a sufficient physical cause. Moreover, supervenience states that the supposed mental cause of this effect necessarily supervenes on this physical cause. Given thesis 2, the thesis of non-identity, it seems that we are left with two causes (one mental and one physical) of the same physical effect. Then, in order to avoid overdetermination, (thesis 5), we would be forced to exclude one of the causes and according to Kim we are *a priori*[9] forced to exclude the mental cause, since the physical cause must be preserved in order to uphold causal closure. For Kim, it follows that the non-reductive physicalist must accept either

[9] A priori in the sense that it follows from theses 1-5, rather than, say, because the thesis of causal closure is an a priori thesis, which I deny in Chapter 2.

epiphenomenalism or reductionism and since Kim claims that the latter is more plausible than the former, the non-reductive physicalist is apparently forced to accept some form of reduction.

If it is correct, the exclusion problem does not therefore just render mental causation and non-reductive physicalism incompatible, but it also provides an argument against non-reductive physicalism itself. As Kim writes, "Non-reductive physicalism, like Cartesianism, founders on the rocks of mental causation." (Kim, 2003b: 193). Moreover, this conclusion is thought to be so forceful because it is thought to follow *a priori* from the minimal commitments of non-reductive physicalism. Since I argue that all of these theses are in fact minimal commitments, what solution is available to the non-reductive physicalist?

Before answering this, let us consider why it is important to find a solution to this problem. Note that what is at stake in the exclusion problem is the idea that our mental states, such as our intentions, beliefs and desires, *in virtue of* their mental properties, have real effects in the physical world. Without this idea, concepts such as free will, autonomy and moral responsibility would seem empty of content (if what we think and choose does not, after all, have a real effect in the world, then in what sense can we be held responsible for the actions that we perform?). Moreover, mental causation forms an intrinsic part of our concept of rational agency. Given the strong reasons that we have for being physicalists and for holding onto mental causation, providing a coherent, non-reductive physicalist account of mental causation that avoids the threat of epiphenomenalism, or reduction, is therefore an important task to undertake and has implications beyond the scope of philosophy of mind.

1.1.2 Aim of Thesis

In this thesis, I aim to provide a non-reductive physicalist solution to the exclusion problem. I argue that Kim's exclusion problem depends on a particular conception of causation as sufficient production[10] and that when causation is understood in interventionist terms, the exclusion problem can be avoided. After identifying the assumptions that I take to underlie the exclusion problem, I argue that Woodward's (2003, 2008a, 2011a) version of interventionism not only provides an account of mental causation that avoids the exclusion problem, but argue that it also upholds all of the minimal commitments of non-reductive physicalism, thereby providing a successful non-reductive *physicalist* solution to the exclusion problem. An important theme of this thesis is therefore providing a solution to the exclusion problem that is genuinely *physicalist*. Moreover, I argue that other attempts to solve the exclusion problem that also appeal to broadly interventionist theories of causation, such as those proposed by List and Menzies (2009) and Campbell (2007, 2008a, 2008b, 2010) fail to provide *satisfactory* solutions to the exclusion problem, since these theories introduce a problematic kind of anti-realism into their theories. I conclude that Woodward's interventionist account of mental causation therefore provides the *only* satisfactory non-reductive physicalist account of mental causation and solution to the exclusion problem.

In order to support this hypothesis, I will need to establish what the minimal commitments of non-reductive physicalism are and what their implications are for the status of mental causation and the exclusion problem. I

[10] I examine this concept of causation in detail in Chapter 3.

will need to identify the assumptions that underlie the exclusion problem and demonstrate how they contribute to this problem. I will also need to present a convincing argument to undermine these assumptions. Finally, I will need to demonstrate that interventionism provides an account of mental causation that not only avoids the exclusion problem, but also upholds all of the minimal commitments of non-reductive physicalism and address any objections that arise for this theory.

1.1.2.1 Why Interventionism?

Before moving on to present the outline of this thesis, it will be helpful to make clear why it is that I appeal specifically to the theory of interventionism to provide a solution to the exclusion problem. I examine the theory of interventionism in detail in Chapter 4, but for now it will be helpful to make the following points.[11] Firstly, note that when I refer to interventionism throughout this thesis, unless otherwise stated, I refer to the specific version of interventionism proposed by James Woodward in his (2003)[12]. According to Woodward, the distinguishing feature of all causal relationships is that they are potentially exploitable for the purposes of control and manipulation. Very roughly, in order for X to cause Y it is necessary and sufficient that there is some intervention on X that changes Y. As Woodward puts it[13], "(M) *X* causes *Y* if and

[11] It is important to be clear that in discussing interventionism, my aim in this thesis is not to provide a comprehensive analysis of causation itself (any such analysis would be well beyond the scope of this thesis), but rather, in appealing to interventionism I hope to shed light on the exclusion problem and demonstrate that it provides a successful non-reductive physicalist solution to this problem.

[12] In Chapter 6, I appeal to a slightly modified version of interventionism, presented in Woodward (2011a).

[13] Woodward designates this definition with the letter 'M' to capture the idea that interventionism is a *manipulationist* theory of causation.

only if there are background circumstances *B* such that if some (single) intervention that changes the value of *X* (and no other variable) were to occur in *B*, then *Y* would change." (Woodward, 2008a: 222)

Why do I appeal to this theory of causation? Firstly, and perhaps most importantly, I demonstrate that interventionism is able to uphold all of the minimal commitments of non-reductive physicalism and therefore provide a viable non-reductive *physicalist* account of mental causation and solution to the exclusion problem. Secondly, I demonstrate that Woodward's *specific* version of interventionism is the *only* version of interventionism that provides a satisfactory solution to the exclusion problem, since it is the only interventionist account that is able to avoid serious problems concerning realism. More generally, I appeal to interventionism over other counterfactual theories of causation because it provides the most coherent account of causation in general and is able to best deal with common objections raised against counterfactual theories (for example, the problem of overdetermination).[14]

As we will see, this interventionist account of causation (and mental causation) is "metaphysically modest" (Woodward, 2003: 121)[15], for example, in comparison to a conception of causation that posits the transfer of some

[14] In comparison to, for example, Lewis' (1973b, 2000) counterfactual theory of causation. There have been many responses in recent years (for example, Loewer (2007), Shapiro and Sober (2007), Yablo (1992), List and Menzies (2009), Raatikainen (2010), Campbell (2007, 2008a, 2008b, 2010) to name but a few) that argue somewhat similarly that when causation is understood in counterfactual terms (or more specifically, in broadly interventionist terms), the exclusion problem can be blocked. I will not discuss all of these alternative theories in this thesis (I do discuss List and Menzies' (2009) and Campbell's (2007, 2008a, 2008b, 2010) theories in Chapter 5 in relation to the problem of realism), but again appeal to Woodward's specific version of interventionism because it provides the most independently coherent account of causation and because these theories fail to either provide genuinely *non-reductive physicalist* solutions to the exclusion problem, or fail to provide *satisfactory* solutions to the exclusion problem.

[15] I explore this issue in detail throughout the thesis.

conserved physical quantity as a necessary condition for causation[16]. However, I will argue that this "metaphysically modest" (Ibid) account is the only one we can give as serious *physicalists*, given that it is only by being "metaphysically modest" (Ibid) that this account is able to uphold all of the minimal commitments of non-reductive physicalism. For example, I demonstrate that interventionism provides an account of causation by which *supervenient* mental properties can count as genuine causes, in addition to their physical realizers. I show that this account respects the theses of *causal closure* and *non-overdetermination* by guaranteeing that mental properties cannot contribute to or interact with the sufficient physical causes of physical effects, or qualify as metaphysically distinct productive causes of those effects. I demonstrate that this account also upholds causal closure in the sense that it remains true that every physical effect has a sufficient physical cause, even when causation is understood in interventionist terms. And finally, I demonstrate that this account nonetheless upholds the theses of *non-identity* and *mental causation*, since it assigns genuinely distinct causal roles to mental properties, such as intentions, beliefs and desires.

As I have said, although I argue that this account is the only one we can give, given the minimal commitments of non-reductive physicalism, an important theme of this thesis will nonetheless be to demonstrate that this "metaphysically modest" (Ibid) account of mental causation does provide a *satisfactory* account of mental causation and solution to the exclusion problem.

There is one further aspect of interventionism that is worth discussing at this stage and this concerns the question of what *ontology* interventionism

[16] Such as the theories of Dowe (1999) and Salmon (1984).

operates with, or slightly differently, what the causal relata are on the interventionist account of causation.

According to interventionism, the relata of causation can be best understood as variables that can take different values. For example, in the causal claim 'Smoking causes cancer', the cause variable 'smoking' may take one of only two values, namely 'smokes'/'does not smoke', or may take one of many different values, for example, 'smokes five cigarettes a day'/'smokes ten cigarettes a day', while the effect variable 'cancer' may take one of only two values, namely 'develops cancer'/'does not develop cancer'. What then is the relationship between variables and property instantiations, which as I mentioned above, are most commonly invoked as causal relata in the mental causation debate?

As Woodward explains, we may simply understand variables as properties that can have more than one value. For example, the property of having mass may take the particular values of having a mass of 10kg, 5kg, or 2kg and so on, while the property of having some belief may take one of only two values, relating either to the presence or absence of the belief. It will therefore be possible to speak both in terms of causation between property instantiations (following Kim) and in terms of causation between variables (following Woodward). Variables may also represent events or states, which are also invoked in the mental causation debate, but for the sake of continuity, I will follow Kim and will refer, for the most part, to causation between property instantiations.

Unlike, for example, the theories of Davidson (1967) and Hornsby (2003, 2004), interventionism does not therefore operate with a pre-defined ontology.

For example, we may define an effect on one occasion as a property and on another occasion as a complex action or event. This will often be guided by the goal of the enquirer and the context of the causal claim. However, there is an important sense in which it does 'matter' to interventionism how we 'pick out' causes and effects. This is because according to interventionism, to the extent that two variables enter into exactly the same manipulability relations, or more specifically, the same invariance relations, (the notion of invariance will be explained in detail in Chapter 4[17]) in relation to some effect, it is appropriate, in interventionist terms, to consider them as the *same* cause. By the same token, in so far as two variables enter into distinct manipulability relations (i.e. distinct levels of invariance) in relation to some effect, it is appropriate, in interventionist terms, to consider them as genuinely causally distinct, i.e. as causes that cannot be identified or reduced. These ideas will be explored throughout the thesis.

1.2 Outline of Thesis

The thesis is organised as follows. In Chapter 2, I examine, in detail, the five minimal commitments of non-reductive physicalism that apparently lead to the exclusion problem. After briefly outlining how the exclusion problem is generated from these theses, I examine the thesis of causal closure in detail and argue that despite facing potentially serious problems and despite having had a complex history, it is in fact a minimal commitment of non-reductive physicalism. I also examine the thesis of supervenience in order to establish exactly which form of supervenience the non-reductive physicalist is minimally

[17] For the moment, it will be useful to point out that according to interventionism, invariance is the 'key feature' that a relationship or generalization must possess if it is to qualify as causal.

committed to and examine its implications. Although I will not discuss the thesis of non-overdetermination until Chapter 3, assuming that the idea that the physical effects of mental causes are not systematically overdetermined is fairly plausible, I conclude that all five theses are in fact minimal commitments of non-reductive physicalism that cannot be rejected in order to avoid the exclusion problem.

Chapter 3 examines the assumptions that I take to underlie the exclusion problem. I examine Kim's exclusion argument in detail and argue that despite its apparent inevitability, in order to generate the exclusion problem from these minimal commitments, Kim requires the assumption of sufficient production, i.e. the assumption that causation is *identical* to sufficient production. I demonstrate that when this assumption is combined with the theses of causal closure, supervenience and non-overdetermination, the exclusion problem becomes inevitable. However, I show that without this assumption, these minimal commitments *do not* a priori lead to the exclusion problem. Since I go on to undermine this assumption, this allows the non-reductive physicalist to put forward a positive account of mental causation that upholds all of the minimal commitments of non-reductive physicalism, whilst avoiding the *a priori* threat of causal exclusion that follows once one accepts this assumption.

In Chapter 4, I introduce Woodward's (2003) interventionist theory of causation. The main objectives of this chapter are as follows: 1. To present interventionism as a coherent theory of causation and in particular, to examine those features of this theory that are especially relevant to my argument in Chapter 5, in which I present the interventionist account of mental causation as a solution to the exclusion problem. 2. To present an argument that undermines the

assumption of sufficient production. 3. To address any potential objections to the theory. The goal of this chapter is therefore to be left with a coherent theory of causation that can be used to provide a satisfactory solution to Kim's exclusion problem and which undermines the assumption of sufficient production, thereby demonstrating that the non-reductive physicalist need not accept Kim's a priori exclusion problem.

In Chapter 5, I outline Woodward's interventionist account of mental causation and demonstrate that it not only avoids the exclusion problem, but also upholds all of the minimal commitments of non-reductive physicalism, thereby providing a successful non-reductive *physicalist* solution to the exclusion problem.

More specifically, I demonstrate that interventionism not only provides an account of mental causation by which *both* mental and physical properties can qualify as causes of the same effect, but that when causation is understood in interventionist terms, mental properties can actually be considered as preferable causes of their effects, in comparison to their subvenient physical realizers. Most importantly, I demonstrate that when causation is understood in interventionist terms, the question of mental causation becomes an entirely *a posteriori*, not *a priori* question. Moreover, I demonstrate that this account provides a *satisfactory* account of mental causation and solution to the exclusion problem and compare this account to two alternative manipulationist accounts of mental causation[18], which I argue fail to provide satisfactory solutions to the exclusion problem. I conclude that Woodward's interventionist account of mental causation therefore

[18] List and Menzies (2009) and Campbell (2007, 2008a, 2008b, 2010).

provides the *only* satisfactory account of mental causation and solution to the exclusion problem.

Chapter 6 then addresses a series of challenges proposed by Michael Baumgartner (2009, 2010) to the interventionist response to the exclusion problem. Baumgartner claims that far from securing the causal status of mental properties and providing a non-reductive physicalist solution to the exclusion problem, interventionism actually generates a new kind of exclusion problem, which rests on weaker premises than the original Kimian formulation of the problem. Moreover, Baumgartner (2010) argues that the proposed interventionist solution to this novel interventionist exclusion problem leads to an 'underdetermination' of mental causation, making this supposed solution not fit for the purposes of the non-reductive physicalist.

I begin by providing an outline and analysis of the debate between Baumgartner (2009) and Woodward (2011a). I demonstrate that although Woodward's solution involves modifying the definition of interventionism proposed in his (2003), (which I appeal to in Chapters 4 and 5 of this thesis), it does offer a genuine solution to Baumgartner's a priori interventionist exclusion argument. In the second half of the chapter, I present my argument against Baumgartner's (2010) underdetermination argument. I demonstrate that by clarifying the metaphysical implications of interventionist mental causation and by clarifying the conditions under which we can acquire empirical *evidence* for mental causation, the non-reductive physicalist who hopes to use interventionism as a solution to the exclusion problem can avoid Baumgartner's underdetermination argument. In fact, I will demonstrate that this discussion actually provides *further* support for the "metaphysically modest" (Woodward,

2003: 121) account of mental causation that I will outline in Chapter 5. I conclude that the interventionist is therefore able to defend her position against *both* of Baumgartner's objections and uphold the interventionist solution to the exclusion problem outlined in Chapter 5.

Chapter 7 follows with some concluding remarks.

2. The Exclusion Problem and Non-Reductive Physicalism

2.1 Introduction

It is claimed, most notably by Jaegwon Kim (1998a, 2005), that the exclusion problem arises specifically for non-reductive physicalists because of their commitment to the following five apparently inconsistent theses:

1. Mental causation, the thesis that mental properties have physical effects.
2. Non-identity, the thesis that mental properties are not identical to physical properties.
3. Supervenience, the thesis that mental properties supervene on physical properties.
4. Causal closure, the thesis that every physical effect has a sufficient physical cause.
5. Non-overdetermination, the thesis that the effects of mental causes are not systematically overdetermined by two metaphysically distinct, sufficient causes.

How then, according to Kim, does the exclusion problem follow from these five theses of non-reductive physicalism?[1] In order to illustrate this,

[1] I examine Kim's exclusion argument in detail in Chapter 3.

consider again the example introduced in the previous chapter: suppose that on some occasion my desire for a cup of tea causes me to reach for the kettle. This seems to be a paradigmatic case of mental causation; my conscious desire for a cup of tea causes a physical effect, namely my reaching for the kettle. Now, the problem according to Kim is that causal closure (thesis 4) states that this physical effect already has a sufficient physical cause and supervenience (thesis 3) states that the supposed mental cause of this effect necessarily supervenes on this physical cause. Given thesis 2, the thesis of non-identity, it seems that we are left with two causes (one mental and one physical) of the same physical effect. Then, in order to avoid overdetermination (thesis 5), we would be forced to exclude one of the causes and according to Kim we are *a priori* forced to exclude the mental cause, since the physical cause must be preserved in order to uphold causal closure. For Kim, it follows that the non-reductive physicalist must accept either epiphenomenalism or reductionism and since Kim claims that the latter is more plausible than the former, the non-reductive physicalist is apparently forced to accept some form of reduction. There seems, therefore, to be a distinctive problem of mental causation for non-reductive physicalism. How then can the non-reductive physicalist avoid this problem?

One obvious solution would be to argue that one (or more) of the five theses is not actually a minimal commitment of non-reductive physicalism. Given that theses 1 and 2 are the theses that non-reductive physicalists are interested in defending, it would seem that the only option would be to reject one of the remaining theses, namely causal closure, supervenience or non-overdetermination. This could, in theory, provide the non-reductive physicalist with a solution to the exclusion problem. For example, if we were to reject the

thesis of causal closure, mental properties would not *necessarily* stand in causal competition with the physical properties that realize them. Or, without the thesis of supervenience, we would not be forced to accept that mental causation entails physical causation, which appears to lead to the exclusion problem when combined with the other minimal commitments.

However, in the remainder of this chapter I argue that causal closure and supervenience are in fact minimal commitments of non-reductive physicalism that *cannot* be rejected in order to overcome the exclusion problem. The chapter is organised as follows. In Section 2.2, I examine causal closure in detail and argue that despite facing potentially serious problems (Section 2.2.1) and despite having had a complex history (Section 2.2.2), it is in fact a minimal commitment of non-reductive physicalism (Section 2.2.2.1). In Section 2.3, I examine supervenience in detail and establish exactly which form of supervenience the non-reductive physicalist is minimally committed to. In Section 2.3.2.1, I examine the implications of this form of supervenience and in Section 2.3.2.2, I examine some problems that arise. Although I will not discuss the thesis of non-overdetermination until the next chapter, assuming that the idea that the physical effects of mental causes are not systematically overdetermined is fairly plausible, I conclude that all five theses are in fact minimal commitments of non-reductive physicalism that cannot be rejected in order to avoid the exclusion problem. Much of the discussion in this chapter therefore provides the 'framework' for my arguments in subsequent chapters.

2.2 Commitments of Non-Reductive Physicalism: Causal Closure

Before examining the thesis of causal closure in detail, it is important to establish exactly which formulation of causal closure we are dealing with. This is because, as E.J Lowe (2000) observes, there are in fact many formulations of causal closure that can be found in the literature, each of which is not necessarily equivalent to the others. For the purposes of the argument in this thesis, the following formulation will be sufficient:

Causal Closure: every physical effect has a sufficient physical cause.[2]

The implications of this formulation will become clearer in the sections below, but it essentially implies that every physical effect has a physical cause that is sufficient, i.e. enough to determine or explain[3] its occurrence. Lowe provides the following definition of the notion of a 'sufficient' cause:

---

[2] I define causal closure in this specific way, rather than, for example, as the thesis that 'every physical effect is sufficiently determined by purely physical prior occurrences' (which, as will become clear, is entailed by causal closure) because Kim appeals to this formulation in his exclusion argument. So, although the arguments in this thesis do not depend on any particular formulation of causal closure, this formulation will be most useful for the purposes of my argument. However, it is important to emphasise that although I define causal closure in this specific way (as the thesis that 'every physical effect has a sufficient physical *cause*'), it should not be assumed that causal closure thereby entails a particular conception of causation as sufficient production, or determination, since I argue (in Chapter 3) that it is precisely this assumption that leads Kim to generate the exclusion problem. Nevertheless, I suggest that it is still possible (and useful) to define causal closure in this specific way, since I demonstrate in Chapter 5 that it *is* true that every physical effect has a sufficient physical cause, even when causation is understood in interventionist terms. I argue that the crucial difference between the interventionist and Kim is that for the former, the fact that physical causes are, by their very nature, sufficient to determine the occurrence of their effects, is simply an empirical fact about physical causation, rather than something that constitutes their causal status. Importantly, I demonstrate that it is only when one makes this latter assumption that the exclusion problem becomes inevitable for the non-reductive physicalist. These ideas will be explored in further detail throughout the thesis.

[3] By explanation here I *only* mean that physical causes are sufficient to explain the *occurrence* of their effects (i.e. sufficient to explain what brought about, or *determined* those effects) rather than

> "…I understand a sufficient physical cause of a given event to be a non-empty set of physical events, each of which is a cause of the given event and all of which jointly causally necessitate the occurrence of the given event." (Ibid: 575)

One also finds the thesis of causal closure described as the view that the physical domain is 'closed' in the sense that one never has to leave that domain in order to explain any physical effect. For example, it is possible to provide a sufficient explanation of the movement of my arm towards the tea cup on my desk in purely physical terms, for example, by referring to the neuronal processes in my brain, the stimulation of my nerve fibres, the contraction of my muscles and so on. By contrast, for example, note that the economic realm is not closed, since there are economic effects that have non-economic causes, such as the effect of a natural disaster on the state of the economy. With this formulation of causal closure in mind, I will now examine this thesis in detail.

2.2.1 What Does 'Physical' Mean?

To begin, it is first necessary to address the question of exactly what is meant by the term 'physical' in the thesis 'every *physical* effect has a sufficient *physical* cause'. Now, this question needs to be addressed since some (Crane and

---

implying that physical causes are sufficient to explain their effects in any richer sense of causal explanation. The relevance of this point will become clearer in the next chapter. It is also worth pointing out that there is on-going debate as to whether quantum indeterminacy undermines physical determinism (and hence undermines causal closure, as it is formulated above). However, Papineau (2009, especially pp. 59-60) and Lowe (2000) put forward convincing arguments to suggest that quantum indeterminacy would not have such an effect on the truth of physical determinism and causal closure. This is an interesting issue, but is beyond the scope of this thesis.

Mellor, 1990) have argued that it is not actually possible to formulate a coherent definition of the physical on which to ground causal closure (and, as I argue, to consequently ground physicalism). As David Papineau concisely captures the problem, "The causal-closure thesis presupposes some prior concept of the physical realm. Some commentators argue that the unclarity of this concept empties the causal-closure thesis of content…" (Papineau, 2009: 57)

Now, one answer to the question of what the physical is appeals to the idea that we should define the physical by reference to those properties that are expressed by paradigmatic physical theories, such as chemistry and physiology. If we then define the physical by reference to the properties that are expressed by what is arguably *the* paradigm physical theory, namely physics, we seem to have two options. We can either define physical properties as those properties that are expressed by our best *current* theory of physics, or else we could define physical properties as those properties that would be expressed by an ideal or future physics. However, this presents us with a dilemma, which is formulated by Crane and Mellor (1990)[4]: if we define physical properties as those properties that can be expressed by our best current theory of physics, it implies that any properties that we may discover at a later time could not count as physical and also implies that current physics is complete and completely accurate, which most accept is simply not true. On the other hand, if we define physical properties as those properties that would be expressed by an ideal or future physics then we will not know what the physical is until we know which properties such a future theory would cover. Thus, if we try to define the physical in terms of *the* paradigm physical theory, namely physics, we have the choice of either accepting that our

[4] Hempel (1969) presents a similar argument.

best current theory of physics is complete and completely accurate (which most accept is simply not true) or accepting that we do not know what the physical is; hence the dilemma.

It seems that this dilemma would have fairly serious consequences for causal closure, given that this thesis makes essential reference to the notion of the physical. As Papineau put it, without a prior, coherent definition of the physical, this thesis would appear to be empty of content.

In order to avoid this conclusion, what the physicalist will therefore need to prove is that they can provide a definition of the physical that can feature in the thesis of causal closure (and consequently ground physicalism), which *does not* appeal to the properties of either a current or future physics. What this will essentially involve is resisting the idea that *all* physical properties can be reduced to (i.e. explained exhaustively in terms of) the properties of physics[5], an idea that Papineau (2008) calls '*micro*physicalism'[6], as opposed to physicalism. Luckily, I do not think that the physicalist need be committed to any such view.

Before demonstrating this, it is worth considering where the motivation for microphysicalism might come from. As Papineau (Ibid) points out, the idea of physicalism does seem prima facie distinct from the idea of *micro*physicalism; the latter says that all properties that have physical effects are either identical to or supervene on the microphysical properties of physics, whilst the former says that all properties that have physical effects are either identical to or supervene

[5] This follows since we need only be committed to a definition of the physical which appeals to physics if we think that all physical properties ultimately reduce to the properties of physics.
[6] This theory is called *micro*physicalism because it states that all physical properties (biological, chemical, etc.) are reducible to the *micro*physical properties of physics. Papineau (2008) presents a comprehensive argument against this view. However, for the purposes of my argument, the example presented below will be sufficient.

on physical properties, but leaves unspecified which *kinds* of physical properties they are.

Nevertheless, there is something intuitively appealing about the idea that physics can provide a complete explanation of all physical properties, in a way that other physical sciences cannot. This intuition seems to be based on the fact that physics tells us that all matter is constituted by small particles and are subject to fundamental laws, which are the subject matter of physics. There is therefore a sense in which physics deals with the 'ultimate' reality of objects, which in turn are the subject matter of macro-level physical sciences, such as physiology and biology. This appears to give physics a certain ontological authority over other physical sciences and may lend support to the idea that all physical properties ultimately reduce to the properties of physics.[7]

2.2.1.1 An Argument Against Microphysical Reduction

However, many physicalists reject the idea that all physical properties are reducible to the microphysical properties of physics. For example, Kim (1998a: 85) argues that macro-level physical properties, such as the property of having a mass of 10kg, or the property of being a H_2O molecule, have causal powers that do not reduce to the properties of physics.

Is Kim right to argue that macrophysical properties are not reducible to the microphysical properties of physics? In order to demonstrate that Kim is correct, we can appeal to Putnam's famous example of the square peg, captured in the following passage by Bill Brewer (2011):

[7] Crane and Mellor (1990) put forward a similar idea.

> "…compare Putnam's (1978) famous observation that the best explanation of the fact that a given one inch square peg passes through a one inch square hole and not through a one inch round hole is given by citing its size and shape. All other things being equal, it is precisely this property – one inch squareness – whose presence facilitates, and absence obstructs, its passage. Any proposed move in the direction of scientific-physical explanation by appeal to lattices of elementary particles and the like reduces this robust modal generality. For one inch square pegs of quite different materials equally pass through a one inch square hole and not through a one inch round hole, regardless of the fact that the scientific-physical properties involved in explanation of their motion and interaction are quite different; and whatever their scientific-physical differences may be – within reason – appropriately sized pegs that are not square will not pass through a one inch square hole, and square pegs greater than one inch in side will not do so either. Thus, all other things being equal, the scientific-physical differences between pegs that do, and pegs that do not, pass through a one inch square hole but not through a one inch round hole, are explanatorily unified as those in which the peg is one inch square versus those in which it is not." (Brewer, 2011: 78)

What this example illustrates, and what Brewer captures so concisely, is that there are *macro*physical properties, such as the property of 'one-inch squareness', which cannot be reduced to the microphysical properties of physics. For example, there are certain facts about this macrophysical property, such as its "robust modal generality" (Ibid), which simply cannot be reduced to, or

explained in terms of the microphysical properties of physics, thereby undermining the microphysicalist thesis. This example therefore offers a solution to Crane and Mellor's dilemma, since it illustrates that it is possible, in theory, to provide a definition of the physical that does not appeal to the properties of either current or future physics.

How then should we define the physical if it is not by reference to the properties of physics? This question needs to be addressed, since although I have suggested that it is possible to avoid Crane and Mellor's dilemma, we still need to provide a positive definition of the physical if we are to avoid the charge that the theses of causal closure and physicalism are 'empty' of content.

2.2.1.2 The 'Object View'

One promising option is to appeal to what Daniel Stoljar (Fall 2009, 2010) calls the 'Object View' of the physical, which defines physical properties as follows:

> "A property is physical iff: it either is the sort of property required by a complete account of the intrinsic nature of paradigmatic physical objects and their constituents or else is a property which metaphysically (or logically) supervenes on the sort of property required by a complete account of the intrinsic nature of paradigmatic physical objects and their constituents." (Stoljar, Fall 2009)

According to this view, we should define physical properties as those properties that are required to provide a complete account of the intrinsic nature

of paradigmatic physical objects, such as tables, rocks and chairs. For example, if it is true that the properties of having mass and extension are properties required to provide a complete account of the intrinsic nature of tables and rocks, then these properties should be counted as physical.

Now, the immediate appeal of this theory is that it defines physical properties by reference to our ordinary, common sense conception of a physical object and since we do seem to have a pre-theoretical understanding of what physical objects are, for example, they are the ordinary objects that we are presented with in perceptual experience[8], the Object View will accord with our ordinary, intuitive understanding of what it is to be physical. However, two immediate worries arise.

Firstly, there is the worry that this definition is viciously circular, since it appeals to the properties of *physical* objects to define what a *physical* property is. In response, we can appeal to the following point from Stoljar. As Stoljar explains, this definition avoids the problem of being *viciously* circular since we are not strictly using the same definition of the physical in both cases. This is because although we have a clear understanding of what it is for an *object* to be physical, (physical objects are the objects that we are presented with in ordinary perceptual experience), we do not likewise have a clear understanding of what it is for a *property* to be physical. For example, although we can easily understand what it is for an *object* to be square (through perceptual experience), we do not likewise have a clear understanding of what it is for a *property* to be square, since a property is simply not the kind of thing that can *be* square. Stoljar's solution is to suggest that although properties are not physical in exactly the

[8] Brewer (2011) provides a defence of this view.

same way that objects are, they are derivatively similar in the sense that we can define a physical property as one that is a 'distinctive' property of a paradigm physical object. This definition therefore avoids the problem of being viciously circular.

Secondly, there is the worry that this view could not provide us with a *complete* definition of the physical, since there are presumably intrinsic properties of paradigmatic and *non*-paradigmatic physical objects that we are not aware of. Moreover, without such a complete definition, it seems that we would once again run into the problem identified by Crane and Mellor, since one could argue that without a complete definition of the physical, we would not know what the physical is.

Once again, in response, I suggest that we can appeal to the following point from Stoljar. As Stoljar correctly observes, this objection, and the one noted above, would only be fatal to the Object View if we supposed that it could provide a complete and exhaustive definition of the physical. However, there is no reason to suppose that the Object View should provide such a definition. Instead, what Stoljar suggests we should expect from the Object View is that it simply provides a deeper *understanding* of what it is to be physical.

2.2.1.3 The 'Via Negativa' View

There is another way to avoid the problem that the Object View could not provide us with a complete definition of the physical, which is to appeal to the Via Negativa argument proposed by David Spurrett and David Papineau

(Spurrett & Papineau 1999)[9]. According to Spurrett and Papineau, for the purposes of the thesis of causal closure (and for the purpose of appealing to causal closure as a key premise in the argument for physicalism) it is not vital that we know exactly what the physical domain *does* include, but rather it is only vital that we know exactly what it will *not* include. More specifically, it is only vital that we know that the physical will be non-mental. For example, if we begin with an idea of what we mean by 'mental', such as intentional or sentient, we can then define the physical as specifically non-intentional and non-sentient.

This would help us to avoid the problem identified above, since although the Object View cannot provide a complete inventory of all physical properties, the Via Negativa argument claims that this is not necessary in order to provide a definition of the physical that can feature in the thesis of causal closure, since for that purpose, we only need to know that physical properties are non-mental, for example non-intentional, non-sentient, etc.[10]

It is important to be clear that I am not suggesting that this negative definition could provide an exhaustive definition of the physical. Stoljar identifies some obvious problems that would arise from this position. Firstly, it would suggest that everything non-physical is mental, which is not something that the physicalist would want to be committed to, given that there do appear to be genuinely non-physical and non-mental properties, such as mathematical properties. Secondly, if we wanted to avoid this problem by extending the list of

[9] Montero & Papineau (2005) put forward a similar argument.

[10] As Papineau (2001) explains, this definition of the physical is helpful for the purposes of causal closure since if we define physical properties as specifically *non-mental*, then assuming that the physical domain is in this *specific* sense complete, causal closure will be sufficient to ground physicalism. As Papineau puts it, "provided we can be confident that the 'physical' in this sense is complete, that is, that every non-mental effect is fully determined by non-mental antecedents, then we can conclude that all mental states must be identical with something non-mental (otherwise mental states couldn't have non-mental effects)." (Ibid: 11).

all things that are paradigmatically non-physical and use this list to define the physical, then such a list would risk being arbitrary and would not be very helpful in defining the physical for the purposes of causal closure.

Instead, what I have suggested is that the Object View provides a useful and intuitive definition of the physical in terms of the properties that are required to provide a complete account of the intrinsic nature of paradigmatic physical objects and that although this account may not be able to provide a complete inventory of all physical properties, the Via Negativa argument explains that for the purposes of causal closure and physicalism, it is not necessary that we know exactly what the physical domain does include, so long as we know that it will not include mental properties. Moreover, by defining physical properties, at least in part, by reference to the properties of paradigmatic physical objects, this definition has the benefit of being in line with our pre-theoretical, common sense conception of the physical, rather than claiming to somehow follow a priori from the theory of physics. Further still, this definition also avoids the dilemma posed by Crane and Mellor, since it does not refer to the properties of either a current or future physics to define the physical.

2.2.2 Is Causal Closure True?

Now, one consequence of appealing to this definition of the physical, which rejects the view that the physical can be defined in terms of a 'bottom level' physical theory, such as physics, which is closed in the sense that it requires no further explanation and can explain all other physical theories, is that it rules out the idea that causal closure follows *a priori* from the very definition of the physical. In other words, it rules out the idea that the physical can simply

be defined as 'that which is closed'. However, if causal closure does not follow a priori from the very definition of the physical, what reason do we have for accepting causal closure?

In this section, I argue that despite having had a complex history, the thesis of causal closure is a true *a posteriori* thesis, supported by empirical discoveries in science. In order to demonstrate this, I explore the historical account of causal closure provided by Papineau (2001)[11]. Then, in Section 2.2.2.1, I examine the implications of this thesis for non-reductive physicalism and argue that it is a minimal commitment of non-reductive physicalism, since it provides the grounds for physicalism itself. I conclude that causal closure cannot therefore be rejected by the non-reductive physicalist in order to avoid the exclusion problem, but rather that any successful non-reductive physicalist account of mental causation must uphold this thesis.

It is typically thought that causal closure follows from the discovery of the conservation laws of physics, for example, the conservation laws of energy, mass and momentum. As Papineau explains, this seems to follow since these laws tell us that important physical quantities are conserved suggesting that the later states of a physical system are fully determined by prior, purely physical occurrences. However, as Papineau explains, not just any conservation laws will generate this conclusion.

For instance, while the conservation laws of Leibniz, which replaced those of Descartes, guarantee causal closure, the conservation laws proposed by Newtonian physics, which replaced Leibniz's theory, do not.

[11] This discussion draws heavily on the historical account provided by Papineau. See Papineau (2001) for further details.

For example, Descartes' conservation laws specified that the total quantity of *motion* must be conserved within a physical system, but not that the total quantity of *momentum* needs to be conserved, which left open the possibility that non-physical forces (possibly the mind) could interact with and alter the momentum of physical particles in the brain without violating the conservation of quantity of motion.

Leibniz then replaced Descartes' conservation laws with two modern conservation laws, the conservation of linear momentum and kinetic energy. Now, these two modern conservation laws did guarantee the causal closure of the physical domain.[12] This is because the first conservation law, the conservation of linear momentum, guaranteed the preservation of the total sum of quantity of motion for any given direction, which ruled out the possibility that extra mental forces could influence the movement of physical particles. The second conservation law, the conservation of kinetic energy, guaranteed that the speed and direction of these physical particles were fully determined after impact. Very roughly, these laws guaranteed causal closure since they left no room for any non-physical influence on the motion of matter and guaranteed that the later values of any physical quantity were fully determined by the earlier values of that physical quantity.

However, Newtonian physics, which came to replace Leibniz's physics, refuted the 'mechanical philosophy' proposed by Leibniz. Importantly, Newtonian physics supposed that there could be disembodied forces, such as friction and gravity that could exert force on a physical system without any

[12] As Papineau notes, however, only given the standard 17th century assumption of 'no action at a distance'.

impact between physical matter. Newtonian physics also allowed for the possibility of many other disembodied forces, such as magnetic force, chemical force and even vital and mental forces, which could potentially interact with and influence the physical domain. It seems therefore that Newtonian physics undermined causal closure, since it allowed for the possibility that non-physical (possibly mental) forces could influence the physical domain.

Furthermore, as Papineau notes, Newtonian conservation laws did not help to preserve causal closure either. This is because although Newton formulated a conservation law of momentum, he did not formulate a corresponding conservation law of energy and so Newtonian conservation laws did not rule out the possibility that special mental forces could interact with and influence the states of a physical system. Since Newtonian conservation laws do not seem to support causal closure does this mean, contrary to popular thought, that causal closure is not supported by the conservation laws of physics after all? Not necessarily.

This is because the conservation law of energy *was* finally accepted into Newtonian physics in the mid-19th century and this fact, together with other scientific discoveries of the late 18th and 19th centuries did eventually provide support for and lead to the acceptance of the thesis of causal closure.

Firstly, as Papineau explains, the rational mechanics developed by mathematicians in the 18th and 19th centuries helped to develop the Newtonian conservation law of energy by providing mathematical support for the idea that the total sum of kinetic plus potential energy remains constant. Secondly, empirical discoveries in the 19th century, such as the discovery that heat is simply molecular motion, discovered by James Joule, suggested that different natural

processes, such as heat and friction, were simply manifestations of a single underlying quantity, which were subject to conservation laws. Thirdly, these discoveries lent support to the idea that apparently non-conservative, disembodied forces such as friction and gravity do, after all, conserve the total amount of kinetic and potential energy. Papineau describes these three elements as distinct 'strands' that came together to eventually provide support for a universal conservation law of energy. Finally, Papineau explains that it was the work of Hermann von Helmholtz that eventually led to the formulation of the universal conservation law of energy, which Helmholtz took to apply to *all* natural phenomena, including living systems.

However, this was not the end of the story and certainly did not lead to the widespread acceptance of causal closure. This is because the conservation of energy did not necessarily rule out the possibility that special mental or 'vital' forces could exist and influence a physical system, so long as those special forces were deterministic. However, it was during the late 19th and early 20th centuries that evidence finally became available to cast doubt on the existence of these special deterministic forces and to support the acceptance of causal closure. Papineau provides two main arguments that illustrate why vital and mental forces were finally refuted and why causal closure was finally accepted as a part of common scientific knowledge. I outline these arguments below.

The first argument, the 'Argument from Fundamental Forces', appeals to the fact that many so called 'special forces', such as friction, turned out to reduce to a 'small stock' of fundamental forces, which were subject to the conservation of energy. Because these special forces turned out to be nothing more than "macroscopic manifestations" (Ibid: 28) of more fundamental forces, this

provided inductive grounds for thinking that other supposed special forces would similarly turn out to reduce to a small stock of fundamental forces.

The second argument, 'The Argument from Physiology', appeals to the fact that as we have gained increasing knowledge of physical systems and the processes which operate at the most basic level of those systems, nothing like deterministic vital or mental forces have been discovered. Moreover, all of the knowledge that we have gained of the processes operating in living bodies suggests that they can be accounted for by appealing to ordinary physical processes.

So, although it may be true that the first argument on its own, even in combination with the acceptance of Newtonian conservation laws, could not rule out the possibility of vital and mental forces, it is the addition of this second argument which finally provided the evidence needed to rule out the possibility of deterministic vital and mental forces and support causal closure. As Papineau puts it, "In this way, the argument from physiology can be viewed as clinching the case for completeness of physics, against the background provided by the argument from fundamental forces" (Ibid: 31).

What picture of the physical domain does this leave us with? It entails that every physical effect is sufficiently determined by *purely physical* prior occurrences, since it does not leave any room for non-physical forces to interact with or determine physical effects in any way. In other words, this discussion should have demonstrated that causal closure is true and does entail that every physical effect has a *sufficient* physical cause, i.e. a cause that is enough to

determine the occurrence of the effect[13] and that the physical domain is in this specific sense closed. Thus, although causal closure has had a complex history, I hope to have shown that it is in fact supported by the conservation laws of physics, in addition to relatively recent discoveries in science.[14]

2.2.2.1 Implications for Non-Reductive Physicalism

What then are the implications of causal closure for non-reductive physicalism? In the following section I argue that causal closure provides the grounds for physicalism itself and that it is therefore a minimal commitment of non-reductive physicalism. In order to illustrate this, let us again consider the Causal Argument for physicalism that was introduced in the previous chapter:

1. Mental properties have physical effects.
2. All physical effects have sufficient physical causes.
3. The physical effects of mental properties aren't always overdetermined by metaphysically distinct, sufficient causes.

Conclusion: Mental properties are identical to, or supervene on, physical properties.

[13] Again, it should not be assumed that causal closure thereby entails a conception of causation as sufficient production, or determination. (See footnote 2 above.)

[14] As I briefly mentioned in the previous chapter, for Papineau, this progressive emergence of empirical support for causal closure explains the relatively recent rise of physicalism over the last 60 years, since although Newtonian conservation laws were around in the centuries before, the empirical discoveries that supported the adoption of a universal conservation law of energy and the rejection of vital and mental forces were not available until much later. As Papineau himself accepts, although this by no means provides definitive proof for causal closure, or definitive proof for the non-existence of vital or mental forces, these discoveries do nonetheless provide overwhelming *support* for causal closure.

Consider again the example in which my conscious desire for a cup of tea causes me to reach for the kettle. Now, according to causal closure, this physical effect, namely my reaching for the kettle, already has a sufficient physical cause, which is, by definition enough to bring about the occurrence of the effect. If we then want to avoid the systematic overdetermination of physical effects (by two metaphysically distinct, sufficient causes) we must either identify mental causes with physical causes or accept the supervenience of mental causes on physical causes, hence the physicalist conclusion of the Causal Argument.

Now, one way of understanding the role that causal closure plays in the Causal Argument for physicalism is to recognise that causal closure limits the role that mental properties can play in the physical domain. This is because causal closure implies that mental properties cannot affect the energy, mass or momentum of a physical system in order to bring about their physical effects. E.J Lowe captures this point in the following quote,

> "…appeal to [conservation] laws can at best only be used to attack dualist models of psychophysical causation which attribute to the non-physical mind an ability to affect the energy or momentum of a physical system." (Lowe, 2000: 571) [15]

So, if we want to say that mental properties somehow bring about the occurrence of their physical effects we cannot say that they contribute to or interact with the sufficient physical causes of those effects, since this would

[15] This point will be especially important to the argument in the next chapter, since it suggests that causal closure does not imply that the mental cannot play *any* causal role in relation to physical effects, but only rules out the possibility that mental properties can cause physical effects by exerting additional force or energy into a physical system.

violate causal closure. Moreover, if we also want to avoid overdetermination (i.e. avoid the idea that mental properties are metaphysically distinct, sufficient causes of physical effects, in addition to sufficient physical causes) we must accept that mental properties are either identical to physical properties or accept that they are connected via some other dependency relation, for example, supervenience. While the reductive physicalist opts for the type-identity of mental and physical properties and while the non-reductive physicalist, who wants to hold onto irreducible mental properties, opts for the supervenience of mental properties on physical properties, neither position would be generated without the thesis of causal closure. This is because without the idea that physical causes are sufficient to bring about, or determine the occurrence of their effects we would have no reason to accept the identity *or* supervenience of the mental on the physical. Since causal closure provides the grounds for physicalism itself it *is* therefore a minimal commitment of non-reductive physicalism that cannot be rejected by the non-reductive physicalist in order to avoid the exclusion problem.

2.3 Commitments of Non-Reductive Physicalism: Supervenience

As I have explained, the Causal Argument generates two broadly physicalist conclusions: one being that mental properties are *identical* to physical properties (accepted by reductive physicalists) and the other being that mental properties are not type-identical to physical properties, but are related to physical properties via some weaker dependency relation, the most popular option of which for non-reductive physicalists is supervenience. Thus, it is commonly thought that supervenience is a minimal commitment of non-reductive physicalism. However, it is widely accepted that supervenience comes in a

variety of forms and degrees of modal force. It is therefore necessary to establish exactly which form of supervenience the non-reductive physicalist is minimally committed to and what its implications are.

In this final section, I outline and examine two forms of supervenience discussed by Kim (1984). In Section 2.3.1, I examine a weaker formulation of supervenience and argue that it is too weak for the purposes of non-reductive physicalism. In Section 2.3.2, I examine a stronger formulation of supervenience and argue that a version of strong supervenience is a minimal commitment of non-reductive physicalism. In Section 2.3.2.1, I examine the implications of this form of supervenience and in Section 2.3.2.2, examine some potential problems with this strong form of supervenience. I argue that the non-reductive physicalist can avoid these problems and that this form of supervenience is therefore a minimal commitment of non-reductive physicalism that cannot be rejected in order to avoid the exclusion problem.

2.3.1 Weak Supervenience

As Kim (1984) notes, the thesis of supervenience[16] comes in a variety of forms: it can be both 'weak' and 'strong' and both of these forms can vary according to the modal force with which they are thought to hold.[17] For example, both weak and strong forms of supervenience can vary according to whether the

[16] Kim distinguishes between two forms of supervenience, namely individual and global supervenience. As the name suggests, individual supervenience expresses the idea that no two individuals could differ in respect to their mental properties without also differing in respect to their physical properties. In contrast, global supervenience expresses the idea that no two *worlds* could differ in respect to their worldwide distribution of mental properties without also differing with respect to their worldwide distribution of physical properties. In this thesis, I appeal to the notion of *individual* supervenience and hereafter 'supervenience' should be taken to refer to this specific form of supervenience.

[17] I use Lewis-style possible worlds to assess the modal force of the different forms of supervenience.

range of individuals that they cover is limited by either nomological or metaphysical necessity. To begin, let us examine the weaker formulation of supervenience. What exactly does weak supervenience (hereafter WS) state? Kim formulates WS as follows:

> "A *weakly supervenes* on B if and only if necessarily for any x and y if x and y share all properties in B then x and y share all properties in A that is, indiscernibility with respect to B entails indiscernibility with respect to A." (Ibid: 158)

As an illustration, consider the following example from Kim (Ibid): take A to be a set of supervening properties, which contains the property of being a good man (G) and take B to be a set of subvenient properties, which contains the properties of being courageous (C), benevolent (V), and honest (H).[18] We may then ask what it would mean for A to *weakly* supervene on B. According to Kim,

> "This means that if two men share the same properties in B, say, both are honest and benevolent but lack courage (this will insure they share all other properties in B), then they must both be good men or neither is (they of course cannot differ in regard to the tautological or impossible property). Or, what is the same, if one is a good man but the other is not, there must be some property in B with respect to which they differ (say,

[18] Kim specifies that all forms of supervenience only follow if we assume that the sets of supervening and subvening properties are closed under Boolean property-forming operations, for example, complementation, conjunction, and disjunction. Although this is controversial (see McLaughlin and Bennett, Summer 2010), for the purposes of this argument I will follow Kim in assuming that both sets are closed in this way.

> the first is courageous but the second is not). Any differences in A must be accounted for by some difference in B." (Ibid)

Now, we can see from the definition above that WS entails that no two individuals in a *particular* world could differ in respect to their A-properties without also differing in respect to their B-properties. Or slightly differently, it entails that no two individuals in a particular world could share the same B-properties and yet differ in respect to their A-properties.

However, WS *does not* entail that no two individuals could differ in respect to their A-properties without also differing in respect to their B-properties in *another* possible world. As it has been formulated above, WS clearly leaves this possibility open, since it only requires that A supervenes on B *within a particular world*. This degree of modal force, which holds with only nomological necessity (i.e. at worlds with laws of nature similar to our own), leaves open the possibility that in nearby, nomologically distinct worlds, two individuals could share the same B-properties, for example the same physical properties, and yet differ in respect to their A-properties, for example in respect to their mental properties. Similarly, WS leaves open the possibility that two individuals could differ in respect to their (mental) A-properties without also differing with respect to their (physical) B-properties.

The relevant question to answer is whether WS meets the requirements of the non-reductive physicalist. It seems, at least at first glance, that it does not. This is because, as Kim explains, WS fails to meet a basic, presumptive desideratum of supervenience, which is that base properties should determine or fix their supervenient properties in the stronger sense that once an individual's

base properties have been fixed, the supervenient properties follow with *metaphysical* necessity across all possible worlds. As noted above, although WS guarantees supervenient determination within a particular world, it does not guarantee supervenient determination across all possible worlds. Since WS consequently allows for the possibility that in nearby possible worlds, mental and physical properties routinely come apart, it seems that it would not be sufficient for the purposes of the non-reductive physicalist.

Now, although it seems prima facie reasonable that the non-reductive physicalist would require a form of supervenience that holds with metaphysical, rather than merely nomological necessity, could one not argue that for the purposes of non-reductive physicalism, it is in fact sufficient that mental properties are determined or fixed by physical properties at *our* world? Indeed, why should the non-reductive physicalist be committed to any view about the supervenience of mental and physical properties across other possible worlds?

In order to see why, consider the following point made by Brian McLaughlin and Karen Bennett (McLaughlin and Bennett, Summer 2010). As McLaughlin and Bennett explain, the problem for the physicalist is that if one accepts only WS, which states that the mental supervenes on the physical only at a particular world, but denies that this supervenient relationship holds across all other possible worlds, it seems to undermine the status of the supervenient relationship at that particular world. As McLaughlin and Bennett put it, "if there can be things in *different* worlds that are *A*-discernible but not *B*-discernible, why can't there be two such things within a single world? If everything within each world that is *B*-indiscernible is *A*-indiscernible, how can different worlds enforce *different B→A* property pairings?" (Ibid) In other words, the fact that WS fails to

hold across all possible worlds makes the claim that supervenience holds as a matter of *necessity* at a particular world look dubious and it certainly seems that for the purposes of non-reductive physicalism, something stronger would be required.

2.3.2 Strong Supervenience

Although it is clear that the non-reductive physicalist requires a stronger formulation of supervenience than is offered by WS, it is not immediately clear exactly what this stronger version of supervenience must entail. In this section, I examine the strong formulation of supervenience (hereafter SS) put forward by Kim and argue that the non-reductive physicalist requires, and hence is minimally committed to, a form of strong supervenience that holds not merely with nomological necessity, but with metaphysical necessity across all possible worlds, (hereafter SS*mn*).

To begin, note that Kim formulates SS at the most basic level as the following thesis:

> "A *strongly supervenes* on B just in case, necessarily, for each x and each property F in A, if x has F, then there is a property G in B such that x has G, and *necessarily* if any y has G, it has F." (Kim, 1984: 165)

Now, we can see that in order to generate this stronger form of supervenience SS includes an extra modal operator, namely an extra 'necessary', which ensures not only that if any x has some property F in A, it necessarily has some property G in B, but that necessarily if any y has some property G in B it

necessarily has some F in A. This level of bi-directional determination ensures that A necessarily supervenes on B across *all possible worlds*.

However, this leaves open the degree of modal force with which SS holds across all worlds. For example, does SS entail that A-properties supervene on B-properties across all *nomologically* possible worlds? Or does it entail that they supervene as a matter of *metaphysical necessity* across all possible worlds? Moreover, which of these formulations is the non-reductive physicalist committed to?

In order to demonstrate that the non-reductive physicalist is committed to a form of SS that holds with metaphysical necessity, consider the following point made by McLaughlin and Bennett (Summer 2010), in which they claim that SS that holds with only nomological necessity would be consistent with dualism:

> "…Dualists can accept [strong supervenience that holds with only nomological necessity], because dualists can maintain that there are fundamental psychophysical laws…While dualists think that zombies are metaphysically possible, they need not hold that zombies are nomologically possible…Physicalists, of course, do not think that zombies are possible at all. Capturing physicalism therefore requires a supervenience thesis that holds with full-blown metaphysical necessity." (Ibid)

Since this example suggests that SS that holds with only nomologically necessity would be consistent dualism[19], it seems that the non-reductive physicalist would be committed to a version of strong supervenience that holds as a matter of *metaphysical* necessity across those worlds (SS*mn*).[20] In subsequent chapters, unless otherwise stated, I simply take the term 'supervenience' to refer to this specific form of supervenience.

2.3.2.1 The Implications of SS*mn*

However, SS*mn* does not as yet tell us anything about the *nature* of the dependency relation between A-properties and B-properties. As Kim explains, a supervenience claim like SS*mn* simply states a pattern of co-variation between properties, but does not explain the nature, or specific implications of that relationship. For example, it is not clear whether, according to SS*mn,* B-properties necessarily *entail* A-properties, or whether A-properties necessarily *depend* on B-properties.

Firstly, does SS*mn* imply that A-properties are entailed by B-properties? The simple answer appears to be 'yes'. After all, if A-properties supervene on B-properties with metaphysical necessity across all possible worlds, then B-properties will simply entail A-properties.[21]

---

[19] Note that this point is essentially the same one that was raised against WS above. This suggests that SS that holds with only nomological necessity is actually equivalent to WS and that neither are therefore suitable for the purposes of the non-reductive physicalist.

[20] Kim (1987) discusses the way in which SS could accommodate externalism about mental content (as discussed in Putnam (1975b)) by 'widening' the supervenience base of supervening mental properties to include, for example, relational properties. While this is an interesting issue, it is not directly relevant to the argument in this thesis and is beyond the scope of our discussion.

[21] As McLaughlin and Bennett (Summer 2010) point out, the entailment of A-properties by B-properties only follows from SS*mn* if both sets of properties are closed under Boolean operations. Once again, we may assume with Kim that the property sets are closed in this way and that entailment does follow from SS*mn*.

However, it is not equally clear that SS*mn* implies that A-properties metaphysically *depend* on B-properties. As Kim explains,

> "For when we look at the relationship as specified in the definition between a strongly supervenient property and its base property, all that we have is that the base property entails the supervenient property. This alone does not warrant us to say that the supervening property is dependent on, or determined by, the base, or that an object has the supervening property in virtue of having the base property." (Kim, 1984: 166)

Now, this could be potentially problematic for the non-reductive physicalist, since the idea that the supervenient relationship between mental and physical properties is one of asymmetric dependence (whereby physical properties determine mental properties, but not vice versa) is a fairly basic and plausible assumption of physicalism. Since it is not immediately clear that this level of asymmetric dependence follows from SS*mn* (even with entailment), what solution is available for the non-reductive physicalist?

Kim offers the following plausible solution: one can assume the individual supervenient dependence of supervenient properties on their subvenient bases on the grounds that those individual properties belong to a larger set of supervenient properties, which stand in an asymmetric dependence relationship to their subvenient bases. In other words, so long as we have independent grounds for accepting that mental properties, in general, supervene on physical properties, but not vice versa (which the non-reductive physicalist

has in the form of the Causal Argument), we can infer the asymmetric dependence of individual supervenient properties on their base properties. The specific implications of supervenience for mental causation and the role that it plays in the exclusion problem will be made clear in the next chapter.

2.3.2.2 Some Potential Problems with SS*mn*

In this final section, I consider some potential problems that arise for the non-reductive physicalist from this formulation of supervenience. Firstly, one could argue that this formulation of supervenience is too strong. For example, one could argue that as an a posteriori doctrine about the actual world, physicalism should not *a priori* rule out the possibility of things such as Cartesian souls, zombies and ghosts. However, according to the formulation of SS*mn* proposed above, such non-physical entities could not, as a matter of *metaphysical* possibility, exist.

I think that we can solve this problem by appealing to a point that I made in the previous chapter (see footnote 1), which was that physicalism, as it follows from the Causal Argument, does not necessarily entail the view that *everything* is physical, but need only entail the view that everything that interacts *causally* in the world must be identical to, or supervenient on the physical. This is because causal closure, which features as a premise in the Causal Argument, states that every physical *effect* has a sufficient physical cause, but leaves open the possibility that there may be non-physical properties, such as mathematical, or even spiritual properties, so long as these properties do not exert any causal influence on the world. The non-reductive physicalist who is persuaded by the Causal Argument is not therefore committed to the view that non-physical

entities, such as disembodied souls, could not as a matter of metaphysical possibility exist, but is only committed to the view that if these non-physical entities did exist they could exert no causal influence in the world. SS*mn* is not therefore in tension with the a posteriori nature of physicalism.[22]

The second problem that needs to be addressed is whether SS*mn* entails the reduction of supervenient properties on their subvenient base properties. This issue is extremely important for the non-reductive physicalist, since if it turns out that the form of supervenience that I argue is a *minimal* commitment of non-reductive physicalism entails reduction, then non-reductive physicalism will be a priori ruled out by this definition of supervenience.

Now, although I have argued that SS*mn* entails the strong metaphysical dependence and entailment of supervenient properties on their subvenient bases, it is not clear that reduction straightforwardly follows from this form of supervenience. In order to see this, consider the following points.

Firstly, note that SS*mn* is consistent with the multiple realizability of supervenient properties, i.e. consistent with the idea that a supervenient property from set A will supervene on a variety of subvenient bases from set B, such that it will not be possible to type-identify and hence reduce supervenient property *types* with subvenient property *types*. Although the issue of multiple realization is by no means straightforward, it does seem that multiple realization makes type-reduction, even with SS*mn*, implausible.

[22] There is a further worry, which is highlighted by Papineau (2009), which is that if it were true that non-physical properties, such as mathematical properties, did exist but did not exert any causal influence on the world it is not clear how we could acquire any knowledge of them, given the plausible assumption that we normally acquire knowledge of the external world through some kind of causal interaction between properties and our cognitive system. However, as Papineau (Ibid) points out, so long as it is plausible that there are non-causal forms of knowledge, such as a priori knowledge, the physicalist can avoid this problem.

What about reduction at the level of token instantiated properties? SS*mn* states that every particular, or 'token' instantiation of a mental property, such as the property of having a desire for a cup of tea, call it D_1, which causes some physical effect, for example, reaching behaviour B_1, is dependent on and entailed by a token instantiation of a physical property, for example, N_1, which is also a cause of B_1. Does this degree of supervenient dependence and entailment mean that D_1 is thereby reducible to N_1? Not necessarily.

In order to see this, consider Kim's (1984) point that as an *epistemic* activity, reduction does not necessarily follow from this level of entailment and dependence. For example, by knowing that D_1 supervenes on N_1 with metaphysical necessity and that it is entailed by and wholly dependent on N_1 we do not thereby acquire an *explanation* of D_1, for example of its intentional nature.

Moreover, as Kim points out, even though SS*mn* does guarantee that every mental property will be entailed by and wholly dependent on some physical property, it does not follow that those mental-to-physical relationships will be available to analyse for reductive or explanatory purposes. As Kim explains,

> "Where strong supervenience obtains, [this] gives us the assurance that such connections in the form of necessary equivalences are there to be discovered, without of course the further assurance that we shall succeed in discovering them or that they will be representable in an explanatory theory." (Ibid: 176)

What both of these points suggest is that although SS*mn* entails the metaphysical dependence and entailment of supervenient properties on their subvenient bases, this does not necessarily entail the reduction of those supervenient properties. SS*mn* is not therefore incompatible with non-reductive physicalism.

One final point that I will consider is whether SS*mn* entails that supervenient properties are nothing ontologically 'over and above' their subvenient bases. Or as McLaughlin and Bennett (Summer 2010) put it, whether SS*mn* entails that supervenient properties are 'ontologically innocent' with respect to their subvenient bases. It is clear that the non-reductive physicalist requires the 'ontological innocence' of supervenient properties in relation to their subvenient bases, since to accept any kind of ontological distinction between mental and physical properties would be to endorse a form of dualism.

However, if the implications that have been discussed thus far really are implications of SS*mn* it would seem to rule out the possibility that supervening mental properties could be anything ontologically over and above their subvenient bases, in accordance with the requirements of physicalism. This is because the degree of modal force with which SS*mn* holds, which ensures that supervenient properties are entailed by and wholly dependent on physical properties, guarantees that those supervenient properties could not be ontologically distinct from those subvenient base properties.

In this section I have argued that the non-reductive physicalist is minimally committed to a form of strong supervenience, which holds with metaphysical necessity across all possible worlds and which implies that mental properties are entailed by and dependent on physical properties. I argued that any

weaker form of supervenience would not be suitable for the purposes of the non-reductive physicalist, since it would be consistent with dualism. I finally addressed some potential problems with this form of supervenience, but argued that the non-reductive physicalist can avoid them. It is therefore possible to conclude that SS*mn is* a minimal commitment of non-reductive physicalism that cannot be rejected in order to avoid the exclusion problem.

2.4 Conclusion

In this chapter, I began by presenting the exclusion problem as following from five apparently inconsistent theses of non-reductive physicalism and examined two of these theses in detail, namely causal closure and supervenience. I began by examining the thesis of causal closure and argued that despite facing the problem of defining what it means to be physical and despite having had a complex history, causal closure is a true a posteriori thesis that does entail that every physical effect has a sufficient physical cause. Since I argued that this thesis provides the grounds for physicalism itself, I concluded that it is a minimal commitment of non-reductive physicalism that cannot be rejected in order to overcome the exclusion problem.

I then examined the thesis of supervenience in detail in order to determine exactly which formulation of supervenience the non-reductive physicalist is minimally committed to and what its implications are. I argued that the non-reductive physicalist is minimally committed to a form of strong supervenience that holds with metaphysical necessity across all possible worlds and which implies that mental properties are entailed by and dependent on physical properties. After addressing some potential problems with this thesis, I concluded

that it too is a minimal commitment of non-reductive physicalism that cannot be rejected in order to avoid the exclusion problem. As I explained above, although I do not discuss the thesis of non-overdetermination until the next chapter, assuming that this thesis is a plausible one, it is possible to conclude that all five theses are in fact minimal commitments of non-reductive physicalism that cannot therefore be rejected in order to avoid the exclusion problem.

3. The Exclusion Problem and Its Assumptions

3.1 Introduction

In the previous chapter I argued that the exclusion problem appears to follow a priori from five minimal commitments of non-reductive physicalism, namely mental causation, non-identity, supervenience, causal closure and non-overdetermination. Given that I concluded that each of these theses is in fact a minimal commitment of non-reductive physicalism, what solution is available to the non-reductive physicalist? In this chapter I argue that despite its apparent inevitability, the exclusion problem only follows a priori from these minimal commitments when they are combined with an assumption regarding causation, this being the assumption that causation is identical to *production*, *generation* or *determination* and that causes are *sufficient* for the occurrence of their effects.[1] I call this the assumption of sufficient production, (hereafter the assumption of SP). Since I go on, in Chapter 4, to undermine this assumption, this allows me, in Chapter 5, to present interventionism as providing an account of mental causation that not only avoids the exclusion problem, but that also upholds all of

[1] A similar argument is put forward in Woodward (2008a) and in Loewer (2007). Very roughly, Woodward (2008a) argues that Kim's exclusion problem depends on a conception of causation as 'nomological sufficiency' and suggests that when causation is understood in interventionist terms, Kim's exclusion problem does not go through. Loewer (2007) argues that Kim's exclusion problem depends specifically on a conception of causation as 'production' and attempts to provide a solution to the exclusion problem by providing a critique of this productive concept of causation and by proposing a counterfactual approach to causation, roughly along the lines of Lewis' account of counterfactual dependence, in its place. See Woodward (2008a) and Loewer (2007) for further details.

the minimal commitments of non-reductive physicalism, thereby providing a successful non-reductive *physicalist* solution to the exclusion problem.

The chapter is organised as follows. In Section 3.2, I establish exactly what the sufficient production (hereafter SP) concept of causation entails and in Section 3.2.1, I demonstrate that Kim makes the *assumption* of SP[2]. In Section 3.3, I examine Kim's exclusion argument in detail and in Section 3.3.1, I demonstrate that Kim crucially depends on this assumption to generate the exclusion problem. Lastly, in Section 3.3.2, I address some outstanding issues regarding overdetermination, which further demonstrate that Kim crucially depends on the assumption of SP to generate the exclusion problem. It is important to point out that the purpose of this chapter is not to prove that this assumption is false, but only to prove that the exclusion problem crucially depends on this assumption and that without it, the minimal commitments of non-reductive physicalism *do not* lead to the a priori exclusion of the mental.

3.2 The SP Concept of Causation

To begin, we must first establish exactly what the SP concept of causation entails. As I understand it, according to this concept, in order for X to cause Y it is necessary and sufficient that X *produces*, *generates* or *determines* Y's occurrence and that X is a *sufficient* cause of Y, where 'cause' is understood in this productive/generative sense.

[2] In arguing for this I am not claiming that this provides a fully accurate and definitive account of Kim's personal view of causation, since this is not made explicit or clear in Kim's writings. For example, the exclusion problem at times appears to rely on the assumption that causation is identical to nomological sufficiency (see Kim, 2003b: 204), but at other times clearly depends on a conception of causation as sufficient production. So, although the SP concept of causation may not be the concept of causation that Kim advocates in all of his writings, I do argue that Kim advocates this concept in the context of the exclusion problem and relies on it to generate this problem.

How then should we understand the concept of causation as production, generation or determination?[3] Kim defines this concept as follows:

> "On this conception, a cause is something that produces, or generates, or brings about its effects, something from which the effects derive their existence or occurrence." (Kim, 2010a: 235)

Now, this definition is not by itself very illuminating, since one could argue that the notions of 'bringing about' and 'generating' are simply synonymous with causation. Nevertheless, I suggest that we can define this productive concept of causation more precisely by making clear which further features are entailed by this concept. Firstly, there is the idea that cause and effect are connected via a spatiotemporally continuous 'chain' or 'process' of 'causal intermediaries', where these 'causal intermediaries' are understood in the productive sense. (The idea here is presumably that an entire chain or process is sufficient to produce or generate the effect.) Secondly, this concept is closely connected to the idea that causation involves some kind of transfer of energy or momentum, via these productive processes and chains.[4] Lastly, this concept of causation sharply contrasts with and in fact rules out the possibility that omissions and absences can count as genuine causes, on account of their failing to instantiate any such productive chain or process. We can therefore think of these features as the definitive features of the productive concept of causation.

[3] Ned Hall (2004) provides a comprehensive analysis of this productive concept of causation.
[4] This idea can be found in the theories of, for example, Dowe (1999) and Salmon (1984).

This productive concept of causation is common in the literature on causation and is typically associated with particular, well-known examples. For example, it is typically invoked to describe the relationship that obtains when one billiard ball strikes another and *produces* movement in the other ball. Or, it is invoked to describe the relationship that obtains when a baseball strikes a fragile window and produces the resultant physical effect, namely the shattering of the window. Although these examples feature in a wide variety of causal theories, I take it that they are specifically invoked to illustrate the productive concept of causation because they illustrate the primary role that physical processes and transactions play on this account.

The idea that causes are *sufficient* for their effects can then simply be understood as the idea that causes are enough to produce, generate, or determine the occurrence of their effects.[5]

In the remainder of this thesis, I refer to this concept of causation as the *sufficient production*, or SP concept of causation, but it should be noted that the SP concept of causation entails the broader views that causation involves generation or determination, that cause and effect are connected via spatiotemporally continuous productive processes and chains, presumably involving some kind of transfer of energy and that those productive causes are sufficient for their effects. The assumption that is at issue in this chapter is simply the assumption that sufficient production is *identical* to causation. Moreover, as we will see below, there is an explanatory counterpart to this

[5] As Barry Loewer (2007) points out, we should actually understand the idea that causes are sufficient for their effects as entailing the view that causes are *nomologically* sufficient for their effects, since it is only with the laws of nature and the entire physical state of that system that causes can be considered 'sufficient' for their effects. I will not address this nomological issue of the role of laws in Kim's account of causation here, since it is not directly relevant to the argument in this chapter. However, I will return to this issue in the next chapter.

assumption, which assumes that providing a causal *explanation* of some effect is simply a matter of identifying such sufficient conditions for the occurrence of the effect.

Lastly, although I have presented this as one assumption, I will demonstrate that each aspect of the assumption of SP (namely the sufficiency aspect and the productive aspect) plays a distinctive role in Kim's a priori exclusion argument. For example, I will demonstrate that Kim's original formulation of the exclusion problem depends specifically on the sufficiency aspect of the assumption of SP, while Kim's alternative formulation of the exclusion problem, which he advances after acknowledging that overdetermination is not possible between the mental and the physical, depends specifically on the productive aspect of the assumption of SP and even more specifically, on the closely related idea that causation necessary involves some kind of productive process. To be clear, I am not suggesting that each aspect represents a distinct concept of causation and a distinct assumption about causation, but rather, I am suggesting that each aspect plays a distinctive role in Kim's a priori exclusion argument.

3.2.1 Kim and the Assumption of SP

The relevant question to answer is whether Kim makes this assumption about causation and moreover, what the implications of this assumption are for the exclusion problem.

That Kim makes the assumption that causes must be sufficient for their effects is made clear in Section 3.3 below when I outline the way in which this assumption motivates Kim's exclusion problem. Nevertheless, I suggest that

evidence of Kim's assumption can be found elsewhere in his writings. For example, consider the following passage in which Kim explains the causal role that a mental property must play if it is to be considered as a genuine cause of some physical effect:

> "If *M* is a mental property, therefore, *M* must have some new causal powers. This must mean, let us suppose, that *M* manifests its causal powers by being causally efficacious with respect to another property, *N*; that is, a given instance of *M* can cause *N* to be instantiated on that occasion. **We shall assume here a broadly nomological conception of causality, roughly in the following sense: an instance of *M* causes an instance of *N* just in case there is an appropriate causal law that invokes the instantiation of *M* as a sufficient condition for the instantiation of *N***." (My emphasis, Kim, 2003b: 204)

It is clear from this passage that Kim identifies causality with sufficiency, or more accurately, with the idea that causes are *nomologically* sufficient for the occurrence of their effects (see footnote 5 above).[6]

It is also apparent that Kim accepts the explanatory counterpart of the assumption of SP. Consider, for example, the following passage:

[6] What this passage also importantly illustrates is that Kim equates the notion of causal 'efficacy' with the SP concept of causation. Thus, when Kim states, as he often does (see especially Kim, 2003a), that mental properties must be causally 'efficacious' with respect to their effects, rather than merely causally 'relevant', I take it that Kim is implying that mental properties must be sufficient productive causes of their effects.

> "Thus a car accident is explained by a highway designer as having been caused by the incorrect camber of the highway curve, and by a police officer as caused by the inattentive driving of an inexperienced driver. But in a case like this we naturally think of the offered causes as partial causes; they together help make up a full and sufficient cause of the accident." (Kim, 1998a: 66)

As well as further illustrating that for Kim a cause should simply be understood as a sufficient condition for the occurrence of its effect, this passage also suggests that for Kim, causal *explanation* is also simply a matter of providing sufficient conditions for the occurrence of an effect. According to this view, it seems that there is nothing epistemically richer to causal explanation than providing such sufficient conditions. Moreover, given Kim's assumption that causal explanations simply cite 'full' and sufficient conditions for the occurrence of effects, it naturally follows that it is not possible to have more than one causal explanation of a single event, without running into the problem of overdetermination.[7] This leads Kim to accept the following view:

> "…there can be no more than a single complete and independent explanation of any one event, and we may not accept two (or more) explanations of a single event unless we know, or have reason to believe, that they are appropriately related—that is, related in such a way that one

[7] This also naturally leads to the view, captured in the previous quote, that if a cause is not itself sufficient for its effect, we should consider it as a 'part cause', which somehow adds together with other 'part causes' to 'fully' and sufficiently cause and explain the effect. Helen Steward (1997b) puts forward a detailed and convincing critique of this view. I also critique this view in the next chapter.

> of the explanations is either not complete in itself or dependent on the other." (Kim, 2010b: 160)

I discuss the implications of this concept of causal explanation further in the next chapter.

As well as assuming that causes must be *sufficient* for their effects, it is also apparent that Kim assumes that causation necessary involves production, generation and determination. More specifically, he assumes that mental properties must cause their effects in this productive sense if they are to be considered as genuine causes of their effects. This is evident in the following passage:

> "Causation as generation, or effective production and determination, is in many ways a stronger relation than mere counterfactual dependence, and it is causation in this sense that is fundamentally involved in the problem of mental causation." (Kim, 2005: 18)

Kim also makes explicit the fact that he accepts the closely related assumption that causation necessarily involves some kind of continuous productive chain or process and again assumes that mental properties must bring about their supposed effects via such productive chains and processes if they are to be considered as genuine causes. This assumption is captured in the following passage in which Kim summarises the 'worries' or problems of mental causation that supposedly face the non-reductive physicalist:

> "Fundamentally these worries arise, I believe, from the question whether mentality has the power to bring about its effects in a continuous process of generation and production—or the question whether we can show that this is so." (Kim, 2010a: 236)

I make clear the relevance of this assumption to the exclusion problem in Section 3.3 below, but it is important to emphasise that given the strong metaphysical implications of this productive concept of causation, this assumption commits Kim to a fairly metaphysically demanding conception of mental causation. For example, it implies that in order for mental properties to qualify as genuine causes of their effects, those mental properties must be sufficient to produce, generate or determine the occurrence of their effects and presumably do so via *metaphysically distinct* productive chains and processes. I spell out the implications of this metaphysically rich notion of mental causation below (and explore this issue further in Chapter 5).

3.3 The Assumption of SP and the Exclusion Problem

What then are the implications of the assumption of SP on the exclusion problem? In this section, I argue that despite its apparent inevitability, the exclusion problem only follows a priori from the minimal commitments of non-reductive physicalism when they are combined with the assumption of SP. In order to demonstrate this, let us look closely at how Kim formulates his exclusion argument.

Kim presents his exclusion argument in two stages. In stage 1 we are presented with a supposed case of mental-to-mental causation, in which an

instance of mental property M causes an instance of mental property M*. To begin, Kim points out that it is guaranteed by supervenience that M* supervenes on a physical base, P*, which necessitates M*'s occurrence. Kim's next move is to ask what causes M* to be instantiated on this occasion, M or P*?

It is at this stage that Kim introduces 'Edwards' Dictum' into the argument, which states that a tension is created in any case in which there is 'vertical determination', (represented by the metaphysical supervenience of M* on P*), and a claim of 'horizontal causation', (represented by the supposed causal relation between M and M*). For Kim, a tension arises for the supposed causal relationship between M and M* because supervenience guarantees that the instantiation of P* *alone* necessitates M*'s occurrence and would do so regardless of whether M preceded P* as a supposed cause of M*. In fact, Kim goes as far as to claim that "…vertical determination *excludes* horizontal causation." (My emphasis, Kim, 2005: 36) For Kim, it follows that M could have no causal role to play in the instantiation of M*, given M*'s supervenience on P*; that is of course unless M somehow contributes to the occurrence of P*. Kim's solution to this tension is therefore to claim that M can cause M*, but only by causing its subvenient base, P*. In other words, Kim concludes stage 1 of the argument with the claim that supervenience guarantees that mental-to-mental causation *entails* mental-to-physical causation.

At this stage of the argument, Kim claims that no metaphysical assumptions are made and that the conclusion of stage 1 simply follows from the thesis of supervenience, (which I argued in Chapter 2 is a minimal commitment of non-reductive physicalism). However, contrary to Kim's claim, I believe that

this conclusion *does* rely on a metaphysical assumption, this being the assumption that causation is identical to sufficient production.

In order to see this, note that the supposed causal tension created by Edwards' Dictum could not simply arise from the fact that M* supervenes on P*, since it is widely accepted, and Kim himself recognises (Kim, 1998a: 44) that supervenience is not a *causal* relationship. So, even if it is true that P* is sufficient to determine M*'s occurrence on this occasion and would do so whatever else happened to precede P* as a supposed cause of M*, P* *does not cause* M* and could not therefore *causally* exclude any other property from causing M*. In other words, since supervenience is not a causal relation it could not have any such exclusionary causal implications.

I suggest that one would only accept Edwards' Dictum and hence accept that supervenience creates a causal tension for 'horizontal' (i.e. mental) causation if one assumed that *both* supervenience and causation are relations of sufficient determination. This is because once one makes this assumption one could argue that by being sufficient to determine the occurrence of M*, P* would simply capture all there was to causally explain regarding M*'s occurrence and would create a causal tension for any additional purported cause of M*. Moreover, it would also suggest that M* would necessarily be overdetermined by any additional cause, since then M* would apparently be caused by two metaphysically distinct, sufficient causes. Without this assumption, it is difficult to see why one would accept Edwards' Dictum and the supposed causal tension that it creates.

Given that I have suggested that stage 1 of Kim's exclusion argument does depend on a metaphysical assumption, does this mean that Kim's exclusion

argument fails at stage 1? Not necessarily. This is because the conclusion of stage 1 can be reached by appealing to a much simpler argument, which also relies on the thesis of supervenience, but does not rely on the assumption of SP. Consider the following argument:

1. M is thought to cause M*.
2. Because M* supervenes on P*, whatever causes P* also causes M*.[8]
3. If M were to cause M* other than by causing P*, M* would be overdetermined: M would be a cause of M* in addition to whatever causes P*.
4. Therefore, to avoid the overdetermination of M*, M must cause P*.

Thus, it is possible to conclude, in accordance with Kim, that if M is to cause M* it must do so via P*. Or, in other words, we may agree with Kim that mental-to-mental causation does entail mental-to-physical causation.[9]

In any case, it is not until stage 2 that Kim reaches the conclusion of the exclusion argument and demonstrates that he crucially relies on the assumption of SP. In stage 2, Kim explains that according to supervenience it would also be true that M supervenes on a physical base, P, which necessitates M's occurrence. Furthermore, since in stage 1 Kim concluded that if M were to cause M*, it would have to cause P*, he claims that we have good reason to accept that P is also a cause of P*. Very roughly, the reason that Kim offers for this is that since

[8] This follows since if P* necessitates M*'s occurrence, whatever causes P* will also presumably cause M* to be instantiated.

[9] It is important to point out that even if one does not agree with the argument offered above (for example, the concept of causation that I examine in Chapter 4 does not generate the kind of overdetermination that is required in premise 3), so long as one finds the claim that mental-to-mental causation entails mental-to-physical causation at least plausible, (which non-reductive physicalists should do, considering their commitment to causal closure and supervenience), this is sufficient for the purposes of Kim's argument. This is because the rest of Kim's argument is concerned with providing an a priori argument against *mental-to-physical* causation. As Kim explains, he only introduces the case of mental-to-mental causation to begin with in order to show that the exclusion problem also arises for the purely mental case.

M is dependent upon and determined by P on this occasion and since, ex hypothesi, M causes P*, it is plausible to assume that P is also a cause of P*. Thus, it looks as though P* has two causes, M and P.

Now, I take it that this conclusion actually follows from the minimal commitments of non-reductive physicalism, since according to causal closure, P* must have a sufficient physical cause and in order to avoid overdetermination, the non-reductive physicalist claims that the mental cause necessarily supervenes on this sufficient physical cause. Thus, according to non-reductive physicalism, for any case of mental causation we would be left with two causes of the effect under consideration, one mental and one physical. Nevertheless, regardless of how one reaches this conclusion, we may agree with Kim's conclusion at stage 2 that for any case of mental causation, we would be left with two causes of the effect, one mental and one physical.

The crucial move in Kim's argument comes next with the introduction of the 'exclusion principle' (hereafter EP). Kim formulates the EP as follows:

> "If an event *e* has a sufficient cause *c* at *t*, no event at *t* distinct from *c* can be a cause of *e* (unless this is a genuine case of causal overdetermination)." (Kim, 2005: 17)

Now, it does seem, at least at first glance, that when this principle is combined with the minimal commitments of non-reductive physicalism, the exclusion problem becomes inevitable. This is because, as Kim has explained, for any supposed case of mental causation, for example, between mental property M and physical effect P*, supervenience states that M necessarily supervenes on

a physical property, P, which causal closure states is a *sufficient* cause of P*. In any case of mental causation we would therefore be left with two causes of the physical effect, one mental cause and one sufficient physical cause. However, at this point it appears that the EP would kick in and state that unless this was a case of overdetermination, (which the non-reductive physicalist must avoid), one of the causes would have to go. It then looks as though we would be a priori forced to exclude mental property M, since P would have to be preserved as a cause of P* in order to uphold causal closure.

Consider Kim's conclusion of the exclusion argument:

> "The final picture that has emerged is this: P is a cause of P*, with M and M* supervening respectively on P and P*. There is a single underlying causal process in this picture, and this process connects two physical properties, P and P*. The correlations between M and M* and between M and P* are by no means accidental or coincidental; they are lawful and counterfactual sustaining regularities arising out of M's supervenience on the causally linked P and P*. These observed correlations give us an impression of causation; however, that is only an appearance, and there is no more causation here than between two successive shadows cast by a moving car, or two successive symptoms of a developing pathology." (Ibid: 21)

I illustrate this conclusion in Figure 3.1 below.

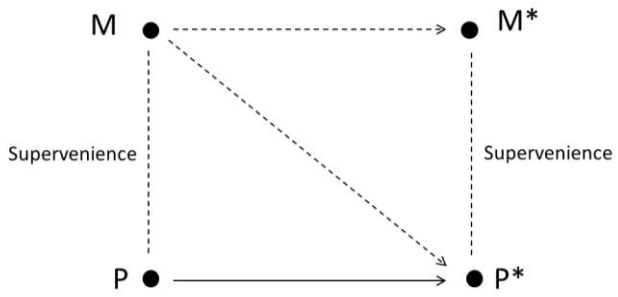

Figure 3.1: Exclusion

Solid arrows represent genuine causal relationships, while the broken arrows represent excluded causal relationships. Broken lines represent supervenient relationships. According to this illustration, the only genuine causal relationship that exists goes from P to P*.

According to Kim, any supposed causal relationship between M and M*, or between M and P* is excluded, or 'pre-empted' by the causal relationship that exists between P and P*. Any attempt to hold onto both M and P as causes of M* or P* would result in the application of the EP and because of the commitment to causal closure, would once again appear to result in the exclusion of the mental cause.

3.3.1 The Assumption of SP and the Exclusion Principle

Now, it is important to recognise that without the EP, the exclusion problem would not follow a priori from these minimal commitments. Remember that Kim's conclusion of stage 2, which I accepted, was only that causal closure and supervenience guarantee that for any case of mental causation we would be left with two causes of the effect, one mental cause and one sufficient physical

cause. Without the introduction of the EP into the argument, we would have no a priori reason to exclude either cause.[10]

Where then does the motivation for the EP come from? Is it, as Kim claims, a "general metaphysical [constraint]" (Ibid: 22) that cannot be 'successfully challenged', or do we have good reason to reject this principle? Remember that without the EP the exclusion problem does not follow from the minimal commitments of non-reductive physicalism, since they only guarantee that for any case of mental causation we would be left with two causes of the effect. We may therefore rightly ask whether the non-reductive physicalist is committed to this principle, which has such serious implications for her theory.

Note that the EP does not follow from the thesis of causal closure. Remember that causal closure only states that every physical effect has a sufficient physical cause, but does not state that if an effect has a sufficient cause at *t* then it could have no other cause at *t* unless it is overdetermined. Causal closure simply states *that* the physical effect has a sufficient cause at *t*. Furthermore, the EP does not follow from the thesis of supervenience. Remember that supervenience only states that mental properties metaphysically supervene on physical properties, but does not state that if those physical properties are sufficient for the purported effects of those supervenient mental properties, then the mental properties could play no causal role in addition to their subvenient bases without resulting in overdetermination.

If the EP is not motivated by causal closure or supervenience, where does the motivation for the EP come from? In the remainder of this section I argue

[10] I demonstrate below that the conclusion of the exclusion argument would not follow here (as Kim seems to assume) simply from the thesis of non-overdetermination, but crucially requires the assumption of SP. I also address a further important issue regarding overdetermination in Section 3.3.2 below.

that in order to motivate the EP and generate the exclusion problem, Kim crucially depends on the assumption of SP.

Take, for example, physical property P, which is a sufficient cause of physical effect P*. The first, more implicit way that I suggest that the assumption of SP motivates the EP is in the sense that if one assumes that causation is *identical* to sufficient production it suggests that by being sufficient to produce its effect, P simply exhausts all there is to cause regarding P*'s occurrence and implies that there would literally be nothing left for any additional property to causally contribute; hence the motivation for the EP, which states precisely that if an effect has a *sufficient* cause at *t*, it can have no other cause at *t whatsoever*, unless it is overdetermined. This motivation is reflected in Kim's appeal to Edwards' Dictum, which Kim admits is his underlying motivation for the exclusion problem (Ibid: 36). Moreover, on the assumption that causal *explanation* is also simply a matter of identifying sufficient conditions for the occurrence of some effect, it implies that by being sufficient, P would simply exhaust all there is to causally explain regarding the occurrence of P*, providing further motivation for excluding any purported additional cause of P*.

It is then easy to see how the EP, motivated by this assumption, causes trouble for the non-reductive physicalist. As I explained above, this is because supervenience guarantees that M necessarily supervenes on a physical property, P, which causal closure states is a *sufficient* cause of P*. Then, given that the non-reductive physicalist must avoid overdetermination, whereby both M and P could count as metaphysically distinct, sufficient causes of P*, it seems that there really would be 'nothing left' for M to causally contribute given P's

occurrence.[11] On the assumption that causation is identical to sufficient production, physical causation, by its very definition would capture all there is to causally contribute and explain regarding the occurrence of some effect and would provide the motivation for excluding the causal role that any additional, mental property might play. Woodward (2008a) captures this point in the following passage:

> "It would seem that physical causation already supplies all of the sufficient conditions (and hence all of the causation) that [is] needed. By definition, a sufficient condition does not require anything "more" to do its work." (Ibid: 252)

The fact that Kim uses the assumption of SP to motivate his argument in this way is, I suggest, evidenced further in the language that he uses to describe the exclusion problem. For example, Kim often frames the exclusion problem in terms of the fact that there is literally 'nothing left' for the mental to cause, given the nature of physical causation. Consider, for example, the following passages[12]:

> "But to acknowledge that *p* has also a physical cause *p**, at *t* is to invite the question: Given that *p* has a physical cause *p**, what causal work is left for *m* to contribute? The physical cause therefore threatens to exclude, and preempt, the mental cause." (Kim, 1998a: 37)

[11] Of course, this relies on the related assumption that M would have to cause P* by being a metaphysically distinct, sufficient cause of P*, which I go on to reject.
[12] The second passage also clearly demonstrates Kim's assumption that causation involves a productive process or chain.

> "…given that *P* is a sufficient physical cause of *P**, how could *M also* be a cause, a sufficient one at that, of *P**? What causal work left is over for *M*, or any other mental property, to do? *M*'s claim as a cause of *P** will be weakened further especially if, as we would expect in real-life neurobiological research, there is a continuous causal chain, a mechanism, connecting *P* with *P**." (Kim, 2003b: 208)

It is difficult to see why one would accept this conclusion unless one assumed that causation is identical to sufficient production. Once one makes this assumption it is easy to see why one would accept the EP, which simply states that if an effect has a *sufficient* cause at *t*, it could have no other cause at *t whatsoever*. Moreover, it is even easier to see how once one accepts the EP and its implicit motivation, the exclusion problem becomes inevitable for the non-reductive physicalist, given that causal closure states that physical causes are, by definition, sufficient to produce, or determine their effects.[13]

The second, more direct way that I suggest the assumption of SP motivates the EP is in the sense that by assuming that causation is *identical* to sufficient production (with its strong metaphysical implications), it suggests that the only way for any property to cause some effect is by being a metaphysically distinct, sufficient productive cause of that effect. According to this assumption then, it would simply not be possible for an effect to have a sufficient cause at *t and* have an additional cause at *t* (where 'cause' is understood in terms of

[13] As I hope to have made clear in the previous chapter, this is again not to imply that causal closure entails the SP concept of causation, since causal closure is a modal claim, which states that every physical effect has a sufficient physical cause and does not entail any particular concept of causation. Rather, the problem that I have identified arises because Kim seems to assume that the physical determinism entailed by causal closure (which guarantees that physical causes are sufficient to produce or determine their effects) is simply *identical* to causation.

sufficient production, with its strong metaphysical implications) without resulting in the overdetermination of that effect; hence the motivation for the EP.

Again, it is easy to see how the EP, motivated by this assumption, causes trouble for the non-reductive physicalist: on the assumption that causation is identical to sufficient production, in order for M to cause P*, it would have to be a metaphysically distinct, sufficient productive cause of P*, *in addition* to P*'s sufficient productive physical cause, P, automatically resulting in a case of overdetermination. Then, the EP would kick in and state that unless we were willing to accept that this is a case of overdetermination, we would be forced to exclude one of the causes and once again, in order to uphold causal closure it looks like we would be forced to *a priori* exclude mental cause M. This is closest to how Kim formulates the exclusion problem himself[14] and it is how the exclusion problem was formulated above.

Without the assumption of SP, it is once again difficult to see why one would accept the EP, which states that if an effect has a sufficient cause at *t* it could have no other cause at *t unless it is overdetermined*. This is because without this assumption, one could claim, for example, that an additional property could cause its effect without being a metaphysically distinct, sufficient cause of that effect and hence without resulting in the automatic overdetermination of that effect, which is necessary to generate the exclusion problem. (This is, in effect, the strategy that I adopt in Chapter 5.)[15] Thus,

[14] See Kim (2005: 42-43).

[15] Crane (1995) argues that by attempting to solve the problem of mental causation by arguing that mental properties do not cause their effects 'in the same way' as physical properties (i.e. by denying what he calls the 'homogeneity' of causation), we lose the original motivation for physicalism and that these accounts, far from solving the problem of mental causation, actually undermine physicalism itself. Very roughly, Crane argues that premise 1 of the Causal Argument for physicalism- the premise of mental causation- must involve the idea that mental properties are

without the assumption of SP, the EP seems unmotivated and without the EP, as Kim himself accepts, the non-reductive physicalist would only be committed to accepting that for every case of mental causation we would be left with two causes of the effect, one mental and one physical, with no a priori reason, however, to exclude either cause.

One might object at this point that I have simply missed the crucial point that Kim is using the EP as a kind of overdetermination principle and that since I have argued that the non-reductive physicalist is minimally committed to the thesis of non-overdetermination, the exclusion problem follows even without the assumption of SP. For example, since I have accepted that for every case of mental causation we would be left with two causes of the effect, one mental and one physical, this would seem to result in a case of overdetermination. In order to avoid overdetermination, the EP, simply understood as an overdetermination principle, would then kick in and state that one of the causes has to go. Then, given the commitment to causal closure, it would seem that we would once again be forced to a priori exclude the mental cause. Thus, it may seem as though the EP is motivated by the thesis of non-overdetermination alone and that it is

---

sufficient to determine the occurrence of their effects, since it is only then do we generate the tension between mental and physical properties when combined with the theses of causal closure and non-overdetermination that is required to motivate the physicalist conclusion of the Causal Argument. However, I do not think that the problem that Crane identifies is a lack of homogeneity per se, since I demonstrate that interventionism provides an account of causation whereby both mental and physical properties cause their effects 'in the same way'. Rather, what Crane's argument highlights is that we do initially require a productive conception of mental causation to motivate physicalism. However, if physicalists then choose to adopt an interventionist conception of mental causation it does not thereby undermine their physicalist position that this concept of causation cannot be used to motivate physicalism, since what the Causal Argument proved is *precisely that* mental properties cannot cause their effects in this productive, generative sense (and that we must therefore accept some form of identity or supervenience between the mental and the physical). In other words, the fact that the interventionist concept of causation cannot be used to motivate the Causal Argument for physicalism merely reflects the fact that this account of mental causation is *constrained* by the commitments of physicalism. It is no wonder then that it cannot be used to generate physicalism. This is an interesting issue, but is one which cannot be pursued further here. See Crane (1995) for further discussion.

therefore possible to generate the exclusion problem without the assumption of SP.

However, despite initial appearances, the EP is *not* equivalent to the thesis of non-overdetermination. In order to see this, remember that the thesis of non-overdetermination only states that the effects of mental causes are not systematically overdetermined by two metaphysically distinct, sufficient causes, but *does not* state that if an effect has a sufficient cause at *t* then it could have no other cause at *t*, unless it is overdetermined. Stated as such, the thesis of non-overdetermination clearly leaves open the possibility that a mental property, such as M, could cause some physical effect, such as P*, in addition to P*'s sufficient physical cause *without overdetermining* P*, so long as M was not a metaphysically distinct, sufficient cause of P*. However, the EP clearly rules out this possibility, since it states that if an effect has a sufficient cause at *t* it could have no other cause at *t* without resulting in overdetermination. I suggested above that the only way to motivate this stronger claim and generate the a priori exclusion of the mental that inevitably follows once one accepts this claim is to assume that causation is identical to sufficient production and hence to assume that the mental cause must be a metaphysically distinct, sufficient cause of its effect, in addition to the sufficient physical cause. Thus, the exclusion problem does not simply follow from the thesis of non-overdetermination, but in order to generate the exclusion problem, Kim requires the stronger claim made by the EP, which appears to be motivated solely by the assumption of SP.

Confusion may arise concerning the connection between the EP and the thesis of non-overdetermination because Kim also formulates the EP in such a

way that it appears equivalent to the thesis of non-overdetermination. Consider Kim's alternative formulation of the EP, which he also advances:

> "No single event can have more than one sufficient cause occurring at any given time- unless it is a genuine case of overdetermination." (Kim, 2005: 42)

Notice, however, that this formulation of the EP is not strictly equivalent to the thesis of non-overdetermination either. As I explained above, this is because the thesis of non-overdetermination leaves open the possibility that a single event *could* have more than one sufficient cause occurring at a given time, so long as the additional cause was not a *metaphysically distinct*, sufficient cause of its effect. By contrast, this formulation of the EP clearly rules out this possibility. Once again, I suggest that the only way to motivate this stronger claim, which inevitably leads to the exclusion problem for the non-reductive physicalist, is to assume that the additional cause *must* be a metaphysically distinct, sufficient cause of its effect (since this makes it impossible for some effect to have more than one sufficient cause without resulting in overdetermination) and I hope to have shown that one would only accept this if one assumed that causation is *identical* to sufficient production.

What this discussion should have demonstrated is that without the assumption of SP, the EP (on either of its formulations) is unmotivated and that without the EP, the exclusion problem does not follow a priori from the minimal commitments of non-reductive physicalism, since as Kim himself accepts, they only commit the non-reductive physicalist to the claim that mental causation

entails physical causation. Thus, without the assumption of SP, the exclusion problem *does not* follow a priori from the minimal commitments of non-reductive physicalism.

3.3.2 Overdetermination: Some Further Issues

One may be wondering whether I have missed an even more important point regarding overdetermination and the exclusion problem, which is that the kind of overdetermination that Kim requires to generate the exclusion problem (which requires the overdetermination of P* by two metaphysically distinct, sufficient causes) is not actually possible given a supervenience relation between mental and physical properties. One could then argue that Kim's exclusion argument can be blocked without having to make any claims about its dependence on the assumption of SP. Karen Bennett (2003) puts forward one such argument in which she claims that one of the necessary conditions for overdetermination, namely that the effect is caused by two *metaphysically distinct*, sufficient causes, cannot be met in the case of mental causation and that the kind of overdetermination that is required for the exclusion problem is simply not possible.

In this final section, I outline Bennett's argument and agree that the kind of overdetermination that is required for the exclusion problem, as it has been outlined above, is not possible in the case of mental causation. However, I go on to demonstrate that rather than providing conclusive proof against the exclusion problem, for Kim, the fact that overdetermination is not possible in the case of mental causation actually provides even greater support for his a priori exclusion

problem. I demonstrate that this conclusion depends, once again, on Kim's assumption that causation is identical to sufficient production.

3.3.2.1 Bennett's Argument against Overdetermination

In a recent paper, Bennett (Ibid) argues that being caused by two properties or events, each of which is sufficient for the occurrence of that effect, is not sufficient for that effect to be overdetermined, since there is a further necessary condition for overdetermination, namely distinctness, which is not met in the case of mental causation. As we shall see, the reason why cases of mental causation fail to meet this requirement and hence fail to result in cases of genuine overdetermination is because of the tight metaphysical connection between the mental and the physical, namely supervenience.

To begin, Bennett discusses what she takes to be a basic presumptive requirement of overdetermination: that it should be possible to consider what the outcome of the effect would have been if one of the causes had occurred without the other. As an illustration, take the classic case of overdetermination involving the firing squad: for the effect (namely the death of the prisoner) to be genuinely overdetermined it *must* be true that if the first rifleman had failed to fire, the prisoner would still have died and vice versa for the second rifleman. For Bennett, this necessary condition for overdetermination can be expressed in the form of a simple counterfactual test:

> "(O1) if *m* had happened without *p*, *e* would still have happened: (*m* & ~*p*)

□→ *e*, and

(O2) if *p* had happened without *m*, *e* would still have happened: (*p* & ~*m*) □→ *e*."[16] (Ibid: 480)

Why should we accept that counterfactuals (O1) and (O2) provide necessary conditions for overdetermination? Bennett's plausible suggestion is that it is because these two counterfactuals capture the natural reasoning that we engage in when we distinguish cases of genuine overdetermination from cases that are not overdetermined, such as cases of joint causation, or exclusionary causation. As Bennett explains,

> "If we needed to decide whether or not the death was overdetermined, we would ask precisely whether these two counterfactuals are true. Would the victim have died if the first gunman had fired without the second? Would he have died if the second gunman had fired without the first? If the answer to both questions is 'no'—if both counterfactuals are false—then the death was not overdetermined, for it was jointly caused by the two gunshots. If only one of the counterfactuals is false, at most one of the gunmen is guilty. So the truth of the counterfactuals does play an important role in our willingness to say that some effect is overdetermined." (Ibid: 477)

[16] Although in this passage Bennett takes m to refer to a mental property and p to refer to the physical property that realizes m, for the moment, we can let m and p represent *any* kinds of properties, since (O1) and (O2) are intended to provide necessary conditions for any case of overdetermination. I refer to m and p as mental and physical properties below.

Bennett is right to suggest that this kind of simple counterfactual reasoning plays an important role in our willingness to state a case of overdetermination. In fact, it is difficult to understand what overdetermination could amount to in the absence of the truth of these counterfactuals. Moreover, one only has to review the way in which overdetermination is discussed in the literature[17] to see that it is widely accepted that the truth of these counterfactuals provide a necessary condition for overdetermination.

The important question to answer of course is whether mental and physical properties meet this necessary requirement. Now, as Bennett explains, in order for this requirement to be met it would need to be true that if mental property m had occurred without physical property p, effect e would still have occurred *and* if p had occurred without m, e would still occur. However, because of the nature of the supervenient relationship between m and p, namely SS*mn*, or strong supervenience that holds with metaphysical necessity, it is *impossible* for p to occur without m and *impossible* for m to occur without p (or more precisely, *some* physical realizer p'). Consequently, at least one of the counterfactuals will turn out false and/or vacuous[18], given that they have impossible antecedents and there is a strong sense in which the vacuity of even one of the counterfactuals means that genuine overdetermination is not possible. As Bennett puts it,

[17] See, for example, Kim (1998a: 44-45), Papineau (2004: 18) and Crane (1995: 5).

[18] I differ here in my opinion from Bennett as to which counterfactual is false and/or vacuous. Although I agree with Bennett that O2 is false and vacuous, Bennett claims that O1 is true, whereas I think we have good reason to think that both counterfactuals are false and vacuous. Very roughly, this is because although because of multiple realization O1 will strictly turn out true, because of the implications of supervenience and causal closure, it is necessary that m supervenes on *some* physical base (call it p') and it would be impossible for m to occur and cause e without p', hence O1 *would* turn out false/vacuous. Nonetheless, for the sake of this argument this issue is not crucial, since I agree with Bennett that the vacuity of even one of the counterfactuals is enough to make overdetermination impossible.

> "To put the point more formally: if one of the causes necessitates the other, if it is at least metaphysically impossible for the one to occur without the other, then one of the overdetermination counterfactuals will come out vacuous. And there is something to be said for the idea that the vacuity of one of them means that the effect is not overdetermined." (Ibid: 479)

What Bennett's argument suggests is that because of the 'tight metaphysical connection' between mental and physical properties, overdetermination is simply not possible in the case of mental causation. What this means is that when we are presented with a case in which some physical effect supposedly has both a mental and a physical cause, such as in the case of P*, we can be sure that although both properties may be sufficient[19] for that effect, they do not run the risk of overdetermining that effect, since they fail to be metaphysically distinct in the way required for genuine overdetermination to occur.

Moreover, it is clear that without the idea that the physical effects of mental causes are always overdetermined by two metaphysically distinct, sufficient causes, Kim cannot reach the conclusion of the exclusion problem[20]. This is because without a claim of overdetermination, there would be no motivation for claiming that when faced with a case of supposed mental causation, involving both a mental and a physical cause, the non-reductive physicalist *must* exclude one of the causes (which, it turns out, must be the

[19] We have, of course, yet to provide a positive account of how to understand the causal relevance of mental property M. For example, I go on to suggest that M can only be considered as a sufficient cause of its effect in virtue of the fact that it supervenes on sufficient physical cause P.
[20] As it has been presented above.

mental cause given the commitment to causal closure) in order to avoid overdetermination. Put slightly differently, once one realises that mental and physical properties cannot genuinely overdetermine their effects, both formulations of Kim's exclusion principle appear to be either irrelevant or simply false. (This is because Bennett's argument shows precisely that it *is* possible for an effect to have a sufficient cause at *t* and have an additional sufficient cause at *t* without that effect being overdetermined, so long as those causes are not *metaphysically distinct* causes of that effect, which they cannot be in the case of mental causation.) And without the exclusion principle, as Kim himself accepts, the non-reductive physicalist would merely be forced to accept that the physical effects of mental causes have both a mental and a physical cause, without, however, facing the threat of a priori exclusion that follows once one assumes that those properties are overdetermining causes.

It is also clear that Kim acknowledges that overdetermination, in the standard sense, is not possible in the case of mental causation. As Kim writes,

> "In standard cases of overdetermination, like two bullets hitting the victim's heart at the same time, the short circuit and the overturned lantern causing a house fire, and so on, each overdetermining cause plays a distinct and distinctive causal role. The usual notion of overdetermination involves two or more separate and independent causal chains intersecting at a common effect. Because of *Supervenience*, however, that is not the kind of situation we have here." (Kim, 2005: 48)

3.3.2.2 Exclusion All Over Again

However, rather than recognising the serious implications that this has for the exclusion problem, Kim claims that the fact that overdetermination is not possible in the case of mental causation actually provides further support for his a priori exclusion argument. In this final section, I demonstrate that this conclusion depends, once again, on Kim's assumption that causation is identical to sufficient production.

The first way that I suggest that the assumption of SP motivates Kim's exclusion argument, even after Kim acknowledges that overdetermination is not possible, can be seen in Kim's (2005) response to Ned Block, who also points out that genuine overdetermination is not possible in the case of mental causation. In this discussion, Kim explains that although it is not strictly true that it is impossible for mental property M to occur without physical property P (since, because of multiple realization, M may be realized by another physical property on another occasion), supervenience and causal closure *do* guarantee that M is realized by *some* physical property, call it P', and according to Kim, the causal exclusion of M follows all over again as a result of M's supervenience on P'. As Kim writes,

> "…we have a replay of exactly the same situation with which we began- M has a physical base, P', threatening to preempt it as a cause of P*. In any world in which Supervenience holds and M causes P*, some physical property, instantiated at the same time, can claim to be a sufficient cause of P*. As long as Supervenience is held constant, there is no world in

> which M by itself, independently of a physical base, brings about P*; whenever M* claims to be a cause of P*, there is some physical property waiting to claim at least an equal causal status." (Ibid: 47)

What Kim seems to be suggesting in this passage is that the occurrence of P' generates a tension for the supposed causal role that supervening mental property M can play in relation to P*. In fact, Kim makes it clear that the mere occurrence of P' (which he correctly observes is guaranteed by supervenience to be instantiated whenever M is instantiated) threatens to 'pre-empt' and exclude M's causal role.[21] As we saw in Kim's argument above, this supposed tension is reflected in his appeal to Edwards' dictum, which states that supervenience *excludes* 'horizontal' (i.e. mental) causation.

Once again, I suggest that this conclusion depends crucially on the assumption of SP. This is because, once one makes this assumption, then by being sufficient to bring about its effect, P', just like P, would simply exhaust all there is to cause and causally explain regarding P*'s occurrence and would provide the motivation for excluding the causal role that any additional, mental property might play. Moreover, given that it has now been recognised that P* *cannot* be overdetermined by M and P' (whereas Kim's original exclusion argument relied on the fact that the non-reductive physicalist must merely *avoid* overdetermination) it seems that there really would be 'nothing left' for mental property M to causally contribute given the occurrence of P'[22]. As Kim puts it elsewhere, "in making a physical cause available to substitute for every mental

[21] More accurately, since P' is a disjunctive physical property, it would be one of the disjuncts of P', instantiated on some particular occasion that causes this supposed tension.
[22] This again relies on the related assumption that M would have to cause P* by being a distinct, sufficient cause of P*.

cause, it appears to make mental causes dispensable in any case." (Kim, 1998a: 44-45) However, as I argued above, one would only reach this conclusion if one assumed that causation is identical to sufficient production; without this assumption, there is no a priori reason that P *or* P' would automatically 'pre-empt' or make 'dispensable' the causal role of mental property M in relation to P*.

The second way that I suggest the assumption of SP motivates Kim's exclusion argument, even after Kim acknowledges that overdetermination is not possible, can be seen in the following passage[23]:

> "In the actual world, we may suppose that a continuous causal chain connects P with P*…And it would be incoherent to suppose there is another causal chain from M to P* that is independent of the causal process connecting P with P*; the only plausible supposition is that if there is a causal path from M to P*, that must coincide with the causal path from P to P*…To be a cause of P*, M must somehow ride piggyback on physical causal chains…And we may ask: In virtue of what relation it bears to physical property P does M earn its entitlement to a free ride on the causal chain from P to P* and to claim this causal chain to be its own? Obviously, the only significant relation M bears to P is supervenience. But why should supervenience confer this right on M? The fact of the matter is that there is only one causal process here, from P to P*, and M's supposed causal contribution to the production of P* is

[23] As we will see, this argument depends specifically on the *productive* aspect of the assumption of SP and more specifically, on the closely related idea that causation necessary involves some kind of productive process or chain.

> totally mysterious…The usual notion of overdetermination involves two or more separate and independent causal chains intersecting at a common effect. Because of *Supervenience*, however, that is not the kind of situation we have here. In this sense, this is not a case of genuine causal overdetermination, and *Exclusion* applies in a straightforward way." (Ibid: 47-48)

What Kim seems to be suggesting in this passage is that since it has been recognised that this supposed case of mental causation could not be a case of overdetermination, whereby M could produce P* via a metaphysically distinct, productive chain or process, M could therefore have no causal role to play in relation to P*, unless it somehow rode 'piggyback' on the only productive chain (and hence the only apparently genuine causal chain) that goes from P to P*.[24] Kim then questions whether we should accept that supervenience can legitimately confer a causal role on M in this way and concludes that given that this cannot be a case of overdetermination, the exclusion of M applies in a more 'straightforward way'.

Now, there is a lot going on in this passage, but it is important to recognise that the exclusion of the mental depends, once again, on the assumption of SP. In order to see this, note that although Kim is correct to point out that this cannot be a case of overdetermination, whereby M could cause P* via a metaphysically distinct, sufficient productive chain or process, without the assumption that this kind of sufficient production is *identical* to causation, there

[24] Remember that M cannot somehow contribute to the causal process that goes from P to P*, given that P is supposed to be sufficient for P* (i.e. given that this would violate causal closure).

would be no reason to conclude that M could have no causal role to play in relation to P* unless it rode 'piggyback' on the productive chain that goes from P to P*. Without this assumption, one could claim, for example, that M's causal relevance to P* should be understood in terms of counterfactual dependence, which claims precisely that properties can be causally relevant to their effects without causing those effects via metaphysically distinct, productive chains or processes. This is, in effect, the strategy that I adopt in Chapter 5, in which I demonstrate that Woodward's interventionist account of mental causation provides an account by which mental properties can be causally relevant to their effects without being metaphysically distinct from the physical causes of those effects.

In response to a paper by Barry Loewer, Kim (2002) does in fact acknowledge that his exclusion problem depends on a conception of causation as 'production' or 'generation', but rejects the possibility that a counterfactual approach to causation could provide a satisfactory account of the causal relevance of mental properties and provide a solution to the exclusion problem. I will not discuss Kim's general worries with the counterfactual approach to causation here, since I demonstrate in Chapters 4 and 5 that interventionism simply avoids these problems. However, it is worth considering *why* Kim thinks that counterfactual dependence could not, in general, provide a satisfactory account of the causal relevance of mental properties, since it sheds light on Kim's reasoning behind his a priori exclusion problem and further suggests how we might avoid this problem.

Why then does Kim think that counterfactual dependence could not provide a satisfactory account of the causal relevance of mental properties and

that the SP concept of causation is required to ground the causal status of mental properties? I suggest that the answer lies in the following passage from Kim:

> "Why should we resort to this "thick" variety of causation in thinking about mental causation? My answer is pretty simple: We care about mental causation because we care about human agency, and agency requires the productive/generative conception of causation. I don't have a knock-down argument to prove that agency requires productive causation; I hope what I will say here makes my claim at least plausible. It seems to me that mere counterfactual dependence is not enough to sustain the causal relation involved in our idea of acting upon the natural course of events and bringing about changes so as to actualize what we desire and intend. An agent is someone who, on account of her beliefs, desires, emotions, intentions, and the like, has the capacity to perform actions in the physical world—that is, to cause her limbs and other bodily parts (e.g., the vocal cords) to move in appropriate ways so as to bring about changes in the arrangement of objects and events around her—open a door, pick up the morning paper, and make a cup of coffee. It seems to me that without productive causation, which respects the locality/contiguity condition, such causal processes are not possible." (Kim, 2010a: 236)

Is Kim right to suggest that without productive causation there would be no agency? The short answer, quite simply, is 'no'. In order to see this, note that the non-reductive physicalist who endorses a counterfactual account of causation

would not be committed to denying that the physical effects of mental causes are also caused by subvenient physical properties, which are sufficient to produce, or determine those effects (presumably via a continuous productive process of some kind), but in fact, given her commitment to causal closure and supervenience, she would be minimally committed to this idea.[25]

When we are presented with a supposed case of mental causation, in which causal relevance is understood in terms of counterfactual dependence, I suggest that we can therefore be certain that the physical effects of those causes, such as the movements involved in picking up the morning paper, or making a cup of coffee, are *still* produced, or determined by the subvenient physical realizers of those mental causes, since this is guaranteed by causal closure and supervenience. The key difference between the non-reductive physicalist in this case and Kim is that the former denies, while the latter insists, that this kind of sufficient production is identical to causation and I hope to have shown that it is only once one makes this assumption that the exclusion problem becomes inevitable.

Of course, given that Kim assumes that this kind of sufficient production is identical to causation and mistakenly assumes that mental properties must be sufficient to produce their effects in order to qualify as genuine causes, it is easy to see why Kim concludes that M could have no causal role to play in relation to P*, other than the one that it acquires by supervening on P, since it really is true that M cannot cause P* in this productive sense, but rather, can only produce P* in virtue of the fact that it supervenes on P. This assumption is reflected in Kim's 'Causal Inheritance Principle', which he defines as follows:

[25] I argued for this at length in the previous chapter.

> "If *M* is instantiated on a given occasion by being realized by *P*, then the causal powers of *this instance of M* are identical with (perhaps, a subset of) the causal powers of *P*." (Kim, 2003b: 208)

According to Kim, the implications of the Causal Inheritance Principle for the non-reductive physicalist are "devastating" (Ibid: 209). For Kim, this is because once we realise that the 'causal powers' of mental properties are identical to those of their subvenient bases, it brings into question the non-reductive physicalist's claim that mental properties are genuinely distinct, irreducible properties. For Kim, the natural consequence of the Causal Inheritance Principle is therefore reduction:

> "…mental events and states have the causal efficacy that they have because their neural/physical realizers have causal efficacy. In fact, a mental state, occurring on a given occasion, in virtue of being realized by a certain neural/physical state, has exactly the causal powers of that physical state…once we are prepared to say what we have just said, the next natural step to take—in my view, a step we are compelled to take—is to reductively identify this particular mental state with its neural/physical realizer. This of course is to jettison the "nonreductive" part of nonreductive physicalism." (Kim, 2010a: 239)

As Kim goes on to explain,

> "To resist the reductive move of identification is to recognize the existence of something whose causal work is at best superfluous, and nonexistent at worst." (Ibid: 263)

Once again, I suggest that this conclusion depends crucially on the assumption of SP. This is because one would only be forced to accept the Causal Inheritance Principle and accept that the causal powers of a supervenient mental property are identical and hence reducible to those of its subvenient base if one assumed that causation is identical to sufficient production, since Kim is right that supervenient mental properties could have no 'new causal powers', independent of their subvenient bases in *this* sense.[26] Without this assumption, however, there would be no reason to conclude that the causal relevance of the mental is at best 'superfluous', or worse still, 'non-existent'.

Lastly, it is important to emphasise that as non-reductive *physicalists* we should not actually be surprised to discover that mental properties can only produce their effects, or be considered as sufficient causes of those effects, in virtue of the fact that they supervene on physical properties. As I explained in

[26] There is, therefore, a sense in which Kim's Causal Inheritance Principle is correct, since it is true that mental properties only have the power to produce or determine their effects in virtue of the fact that they supervene on sufficient physical causes. However, what I have argued is that one would only be forced to accept that the *causal* powers of mental properties are thereby identical and hence reducible to those of their subvenient physical realizers if one assumed that this kind of sufficient production is identical to causation. (Without this assumption, for example, the non-reductive physicalist would merely be committed to accepting that mental properties derive their 'productive power' from their subvenient physical realizers, rather than their 'causal power'.) Consider, for example, Kim's discussion (2010a: 238-239) of Terrence Horgan's non-reductive account of mental causation. According to Horgan, the mental can be said to have genuine causal 'efficacy' in virtue of the fact that mental properties supervene on physical properties, which are sufficient to produce and determine their effects. In this case, it seems Kim is right to claim that the 'causal powers' of Horgan's mental properties would be reducible to those of their subvenient physical realizers. However, it is important to recognise that this is only because Horgan *also* assumes that causation is identical to sufficient production. Without this assumption, there would be no reason to accept this conclusion. See also Kim's worries (Kim, 1998a: 72-77) with Frank Jackson and Philip Pettit's theory of 'Program Explanation' (1990a, 1990b).

detail in Chapter 2, this is because it was our commitment to causal closure (which implies that mental properties cannot exert any force or energy into the physical domain to produce or determine physical effects) and our commitment to the idea that the widespread overdetermination of physical effects by two metaphysically distinct, sufficient causes would be implausible, that we accepted that the mental must supervene on the physical and hence that we should be physicalists in the first place (c.f. the Causal Argument from Chapter 2). In other words, these 'limitations' on mental causation are, I suggest, direct consequences of the minimal commitments of non-reductive physicalism. What I hope to have shown in this chapter is that while for Kim, who assumes that this kind of sufficient production is identical to causation, this is the end of the story for non-reductive physicalists (given that the exclusion of the mental seems inevitable once one accepts this assumption), for non-reductive physicalists who reject this assumption, this is just the beginning of the story.

What I think this discussion therefore suggests is that the real challenge that faces the non-reductive physicalist regarding mental causation is how to provide an account of mental causation that explains how mental properties can have genuinely distinct causal roles (thus avoiding the threat of reduction), whilst being ontologically identical with and metaphysically inseparable from their subvenient physical realizers, which are sufficient to produce their effects. *This*, I believe, is the real remaining 'problem' of mental causation for non-reductive physicalism, but I hope to have shown that there is no *a priori* barrier to providing such an account unless one assumes that causation is identical to

sufficient production.[27] My task in the remainder of this thesis is to provide such an account.

3.4 Conclusion

In this chapter I have argued that the exclusion problem only follows a priori from the minimal commitments of non-reductive physicalism when they are combined with an assumption regarding causation, this being the assumption that causation is identical to sufficient production. I began by examining the SP concept of causation itself and demonstrated that Kim makes the assumption of SP. I then demonstrated how Kim's exclusion problem, as it is most commonly presented, depends crucially on this assumption and that without it, the minimal commitments of non-reductive physicalism *do not* lead to the a priori exclusion of the mental. In the final section, I demonstrated that even once Kim acknowledges that overdetermination is not possible in the case of mental causation, rather than recognising the serious implications that this has for the exclusion problem, Kim once again generates this problem because of the assumption of SP. Thus, it is possible to conclude that the exclusion problem (either in its original formulation, or as Kim presents it after recognising that overdetermination is not possible) does not follow a priori from the minimal

[27] There are some, for example Burge (2003) and Baker (2003), who argue that exclusion arguments place too much importance on metaphysics and not enough importance on actual explanatory practice, which they claim would reveal that the mental is genuinely causal and explanatory. However, I agree with Kim (1998a) that the relevant issue in exclusion argument debates is not *whether* mental causation is real, but *how* mental causation is possible given the metaphysical implications of non-reductive physicalism. Although I argue that the mental plays an important explanatory role that is simply missing on Kim's account of causation, the argument that I present in this thesis does attempt to provide a solution to the exclusion problem that acknowledges the real metaphysical challenges that are posed by a commitment to non-reductive physicalism.

commitments of non-reductive physicalism, but depends crucially on the assumption of SP.

4. Interventionism

4.1 Introduction

In the previous chapter I argued that the exclusion problem only follows a priori from the minimal commitments of non-reductive physicalism when they are combined with an assumption regarding causation, this being the assumption that causation is identical to sufficient production. All of this of course says nothing about whether the non-reductive physicalist is in fact committed to the assumption of SP, or whether the non-reductive physicalist must therefore accept the exclusion problem.

In this chapter I outline and examine an alternative theory of causation, this being the theory of interventionism proposed by James Woodward (2003) and present an argument that undermines the assumption of SP. More specifically, I aim to do the following three things in this chapter: Firstly, I outline Woodward's version of interventionism and in particular, examine those features of this theory that are especially relevant to my argument in Chapter 5, in which I present the interventionist account of mental causation as a solution to the exclusion problem. Secondly, I highlight some problems that the SP concept of causation faces and present interventionism as a coherent alternative theory of causation that avoids these problems, undermining the assumption of SP and hence demonstrating that the non-reductive physicalist need not accept Kim's a priori exclusion problem. Lastly, I address some general objections raised against

counterfactual theories of causation, of which interventionism is an example. It is important to address these general problems if interventionism is to provide a coherent account of mental causation and satisfactory solution to the exclusion problem.

The chapter is organised as follows. In Section 4.2, I outline and examine the central features of interventionism and in particular, examine those features that are especially relevant to my argument in Chapter 5. In Section 4.2.1, I examine the central interventionist notion of invariance, in Section 4.2.2, I examine the technical notion of an intervention and in Section 4.2.3, I explore the interventionist conception of causal explanation. In Section 4.3, I highlight some problems for the SP concept of causation and demonstrate that interventionism avoids these problems, thereby undermining the assumption of SP. In order to demonstrate this, I argue, in Section 4.3.1 that the SP concept of causation does not provide *necessary* conditions for causation. In Section 4.3.2, I argue that providing nomologically sufficient conditions for the occurrence of some effect is not *sufficient* for causal explanation and demonstrate that interventionism is able to avoid these problems. In Section 4.4, I address the worry that despite the problems that the SP concept faces, interventionism fails to provide a viable alternative to this theory and hence fails to undermine the assumption of SP, since it faces serious problems of its own, which the SP concept avoids. These are problems concerning cases of overdetermination (Section 4.4.1), non-paradigmatic causation and causation by omissions (Section 4.4.2). I argue that not only can interventionism overcome these problems, but it is actually able to deal with some of these problems in a more satisfying way than the SP concept. I conclude that interventionism *does*, after all, provide a viable alternative theory

of causation to the SP concept and does therefore undermine the assumption of SP. Lastly, in Section 4.5, I discuss some remaining problems concerning the potentially anthropocentric (Section 4.5.1), anti-realist (Section 4.5.2) and circular nature of interventionism (Section 4.5.3). I conclude that interventionism can avoid these problems and that it can be used to provide a coherent account of mental causation and satisfactory solution to the exclusion problem. An in depth analysis and defence of interventionism is well beyond the scope of this thesis, so I focus only on those features of the theory that are most relevant to my argument.

4.2 Interventionism Outlined and Clarified

According to the interventionist theory of causation proposed by James Woodward (2003, 2008a, 2011a), the distinguishing feature of all causal relationships is that they are potentially exploitable for the purposes of control and manipulation. Very roughly, in order for X to cause Y it is necessary and sufficient that there is some intervention on X that changes Y. Woodward provides the following, more precise definition of interventionism[1]:

[1] Remember that according to interventionism, the relata of causation can be best understood as variables that can take different values and that variables can represent properties, events and states. See Chapter 1 for further details. It is also important to point out that this definition (and a definition relating to the notion of an intervention) will be amended slightly in Chapter 6 in order to deal with an objection raised by Michael Baumgartner (2009, 2010). For now, I continue to use Woodward's original definition, since I demonstrate in Chapter 6 that the validity of the arguments in this chapter and the next are not affected by Baumgartner's objections. Nevertheless, it will become clear that the interventionist must make this amendment in order to address the objections raised by Baumgartner. Lastly, although this definition only provides necessary and sufficient conditions for *type*-level causation, I will discuss the interventionist approach to token-level causation below.

> "**(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**. A necessary and sufficient condition for *X* to be a (type-level) *contributing* cause of *Y* with respect to variable set **V** is that (i) there be a directed path from *X* to *Y* such that each link in this path is a direct causal relationship…and that (ii) there be some intervention on *X* that will change *Y* when all other variables in **V** that are not on this path are fixed at some value. (Woodward, 2003: 59)[2]

The notion of an intervention is examined in more detail below, but for now it will be helpful to illustrate the basic idea of interventionism with an example of Woodward's: it has been found that students who attend private school tend to score higher on tests that measure scholastic achievement than students who attend a government funded school.[3] Now, this raises the question of whether the relationship between attendance at private school and scholastic achievement is genuinely causal, in that private school attendance *causes* scholastic achievement or whether it is merely correlative, in that both attendance at private school and scholastic achievement are joint effects of a common cause, such as parents' attitude to education, or their socio-economic status. According to interventionism, the question of whether attendance at private school causes scholastic achievement, or whether it is merely correlated with it, can be

[2] Woodward differentiates between the notions of direct and contributing causes to accommodate the complexities of causation. However, this distinction is not directly relevant to the argument in this thesis. Woodward (2003), especially pp. 45-61, provides further details.
[3] Based on figures from the US school system.

identified with the question of whether scholastic achievement would change under a suitable intervention on attendance at private school (Woodward, 2008a: 219-220). In general, if intervening on whether a student attends private school is a way of intervening on scholastic achievement, while other causes of scholastic achievement are held fixed, then this relationship will qualify as causal. Conversely, if scholastic achievement does not change under an intervention on private school attendance, while other causes of scholastic achievement are held fixed, then this relationship will fail to qualify as causal.[4]

Put slightly differently, in order for X to cause Y there must exist some possible intervention, understood very roughly as an idealised experiment, either hypothetical or actual, on X that changes Y. If such a relationship of potential control and manipulation exists between X and Y then it will be true that the relationship between X and Y is in fact causal. Woodward captures this distinguishing feature of interventionism in the form of the slogan, "No causal difference without a difference in manipulability relations, and no difference in manipulability relations without a causal difference." (Woodward, 2003: 61)

This example also highlights the important practical focus of interventionism and its relationship to the notions of control and manipulation. Now, this practical focus follows naturally from the interventionist definition of causation, since knowing that intervening on X is a way of intervening on Y has a potential practical benefit in that it allows us to potentially use X as a means of *controlling* or *manipulating* Y. For example, knowing that intervening on

[4] It is worth noting that Woodward's main concern in his 2003 exposition of interventionism is with deterministic causation, for which these kinds of counterfactuals will be appropriate. Woodward does note that interventionism can also apply to indeterministic causation, but that the counterfactuals that are appropriate for assessing causation in this case will be different. This is an interesting issue, but is beyond the scope of this thesis.

attendance at private school is a way of changing scholastic achievement, allows us to use attendance as a means of controlling scholastic achievement (we might, for example, choose to send our children to a private school, given this causal knowledge). As I argue in Section 4.3, this potential practical payoff is missing on the SP concept of causation and as I explain in Chapter 5, this practical payoff plays an important role in the interventionist account of mental causation.

Moreover, as Woodward notes, because of this practical focus, this concept of causation is invoked in a wide variety of the sciences, including statistics, economics, computer science and molecular biology, as evidenced by the fact that the standard for proving causation in these theories implies something like an interventionist concept of causation.[5] Although the fact that interventionism is invoked by a broad range of the sciences does not guarantee that this theory provides a correct analysis of causation, I suggest that it is nonetheless a virtue of the theory that it seems to reflect the causal practices and judgements of a broad range of the sciences.

Given that interventionism understands the information that is relevant for determining[6] whether a relationship is causal or non-causal in modal, or

[5] For example, the 'gold standard' for determining whether X causes Y in many of these disciplines involves the notion of a randomised controlled experiment, whose features are relevantly similar to the technical notion of an intervention, to be spelled out below.

[6] Although much of the language that Woodward uses to define interventionism is somewhat epistemic in nature (e.g. Woodward often focuses on how we, as humans, can *determine* or *establish* whether X causes Y), this is not to say that interventionism is thereby *problematically* subjective or anthropocentric. I discuss this issue in detail in Section 4.5 below, but for now, it is worth noting the following point from Woodward (2003: 22). As Woodward rightly points out, although a theory of causation should distinguish between issues having to do with the content or meaning of causal claims and issues about how we *test* such claims, this is not to overlook the fact that a theory of causation should explain how these two issues fit together. As Woodward puts it, "In particular, our theory of the content of causal and explanatory claims should be accompanied by some epistemological story that makes it understandable how human beings can sometimes learn whether claims with that content are true or false from evidence that is actually available to them." (Ibid: 22) We should therefore expect a certain amount of epistemic language to feature in interventionism, although this should not lead the reader to conclude that interventionism is problematically subjective or anthropocentric for the reasons I outline below.

counterfactual terms (it states that we should consider what *would* happen to the effect under a suitable intervention of its purported cause) interventionism can be classed as a counterfactual theory of causation.[7] I highlight some of the consequences of this below.[8]

Firstly, according to interventionism, all causal claims and causal explanations provide counterfactual information about what *would* happen to the effect under a suitable manipulation of its purported cause, rather than only providing information about what actually *did* happen to the effect given the occurrence of its cause. Consequently, all causal explanations share the feature of being able to answer what Woodward calls 'what-if-things-had-been-different' questions, (or 'w-questions for short'). The practical implications of this constraint on causal explanation are made clear later in this chapter and its importance to the interventionist account of mental causation is made clear in Chapter 5.

Secondly, all causal claims and explanations have 'built into them' a contrastive structure or, as Woodward calls it, a 'contrastive focus'. This naturally follows from the interventionist account of causation, since as Woodward explains,

---

[7] There is on-going debate as to the *kind* of counterfactuals that are involved in causation. For example, can they be indicative counterfactuals, or are they limited to counterfactuals expressed in the subjunctive mood? These issues are not directly relevant to the argument in this thesis, but are discussed in detail in Hoerl, C. McCormack, S. Beck, S. R. (eds.) (2011).

[8] One may worry that since this account appeals to the truth of various counterfactuals to define causation, it needs to provide an account of what makes these counterfactuals true (i.e. what their truth-makers are), given that counterfactuals cannot be 'barely true'. It is usually thought that laws fulfil this role, but as we will see, interventionism rejects the idea that laws are required for causation and explanation and replaces the notion of a law with the notion of an invariant generalization. Although I do think that it is possible to provide an account of the 'truth-makers' of interventionist counterfactuals in terms of invariant generalizations, as Woodward points out, providing such an account will not be crucial, since "…what matters for the arguments that follow is whether causal claims and explanations are related to interventionist counterfactuals in the way that I have claimed—any account of the truth conditions for counterfactuals that is consistent with these relationships will be acceptable for the purposes of this essay." (Woodward, 2003: 10)

> "Any manipulation of a cause will involve a change from one state to some specific alternative, and how, if at all, a putative effect is changed under this manipulation will depend on the alternative state to which the cause is changed. Thus, if causal claims are to convey information about what will happen under hypothetical manipulations, they must convey the information that one or more specific changes in the cause will change the effect…This in turn means that all causal claims must be interpretable as having a contrastive structure, and it also has the implication…that to causally explain an outcome is always to explain why it rather than some alternative occurred." (Woodward, 2003: 145-146)

In other words, given that interventionism necessarily understands causation as involving a manipulation (hypothetical, or actual), which changes the cause variable from one state to another, causal claims and causal explanations will always have built into them a contrastive focus, which tells us that it is the fact that the cause variable took *this* value *rather than that* value, which caused the effect variable to take the value it did. As Woodward (2008a: 225) explains, the notion of contrastive focus thus captures the central interventionist idea that causes essentially 'make a difference' to their effects. The important practical implications of the notion of contrastive focus are explored in more detail below and in the next chapter, but it is important to recognise that contrastive focus is thus built into the interventionist concept of causation and is therefore a feature that *all* causal claims and explanations should exhibit.

4.2.1 Invariance

With this discussion as a background, I will now examine, in detail, the central interventionist notion of invariance. According to interventionism, it is invariance that is "the key feature a relationship must possess if it is to count as causal or explanatory" (Woodward, 2003: 239) and hence it is invariance that distinguishes genuinely causal from non-causal relationships. Moreover, as we shall see, invariance plays a central role in the argument against the SP concept of causation and in the interventionist account of mental causation outlined in the next chapter.

The basic idea of invariance is captured in the following passage:

> "…if a causal relationship between *C* and *E* holds at all, then it must be true that (and the relationship must correctly describe how) for some interventions and background circumstances, *E* will change under those interventions on *C*. This in turn implies that there must be some relationship between *C* and *E* and some interventions on *C* such that if these were to be carried out, that relationship between *C* and *E* would not break down but rather would continue to hold. When this is true, I say that the relationship is invariant under such interventions and background circumstances. Thus, according to a manipulationist account of causation, if a relationship is to qualify as causal, it must be invariant under some interventions." (Ibid: 69)

In other words, according to interventionism, if the relationship between two variables is genuinely causal, we should expect a certain degree of stability in the response of the effect to interventions on the purported cause variable. If no such stable response (to interventions) exists, then that relationship will fail to qualify as causal. For example, in order for it to be true that attendance at private school causes scholastic achievement, it must be true that the relationship between attendance and achievement is invariant, i.e. that it holds under at least some intervention on attendance. If this relationship fails to hold under any interventions on attendance, then the relationship will fail to be invariant and hence fail to qualify as causal.

Woodward provides the following precise definition of invariance:

> "A generalization *G* (relating, say, changes in the value of *X* to changes in the value of *Y*) is invariant if *G* would continue to hold under some intervention that changes the value of *X* in such a way that, according to *G*, the value of *Y* would change- 'continue to hold' in the sense that *G* correctly describes how the value of *Y* would change under this intervention." (Ibid: 15)

Now, as the passages above suggest, in order for a relationship or generalization to be invariant and hence causal according to interventionism, it is not necessary that that relationship is invariant under *all* changes and background conditions, but it is only necessary that it is invariant under a specific kind of change, namely an intervention.

The reason why it is invariance under interventions that takes a privileged role in determining whether X causes Y is simply because it is possible for mere correlations to remain invariant under some changes to background conditions.[9] As an illustration, consider the following example (originally due to Lewis, 1973b): the relationship between a barometer reading, B, and the occurrence of a storm, S, is invariant under certain changes, for example, changes to whether it is a Tuesday, or a Wednesday, whether the barometer is in London or Beijing and so on. However, despite being invariant under these changes, it is clear that B does not cause S, since both B and S are joint effects of a common cause, namely atmospheric pressure. It is for this reason that Woodward stipulates that it is only invariance under *interventions* (and more specifically, invariance under interventions on the variables that feature in the generalization or claim itself) that are necessary for determining whether the relationship between X and Y is causal.[10]

As the passages above also suggest, in order for some relationship or generalization to qualify as invariant and hence causal, it is not necessary that that relationship is invariant across *all* interventions, but it is sufficient that it is invariant under at least *some* intervention. In other words, there is a *threshold* of

[9] In Section 4.4 below, I discuss the interventionist notion of insensitivity, which *does* consider changes to background conditions as relevant for assessing the degree of insensitivity.

[10] In fact, it is a specific *kind* of intervention on those variables, namely a 'testing intervention' that is relevant for assessing invariance. Very roughly, the notion of a testing intervention captures the idea that interventions should test the *discriminating features* of a relationship, if they are to determine whether that relationship is causal. Consider the following example adapted from Woodward (2003: 248-249): imagine that a light is attached to a switch and consider the generalization that the light will remain off if the switch is in any position less than 57 degrees and will turn on if the switch is in any position greater than 57 degrees. In order to determine whether this relationship is causal, the intervention should change the discriminating feature of the switch, i.e. change the position of the switch from any position below 57 degrees to any position greater than 57 degrees. For the remainder of Woodward's discussion, he simply takes the term 'intervention' to refer to this specific kind of testing intervention and I follow Woodward in this usage.

invariance that a generalization or relationship must pass if it is to qualify as causal: those generalizations and relationships that are invariant under at least *some* intervention will pass the threshold of invariance and hence qualify as causal[11], whereas those generalizations and relationships that are not invariant under *any* interventions will fail to pass the threshold of invariance and hence fail to qualify as causal. This captures the intuitive idea that it is possible for X to cause Y even though it is not true that X causes Y in every situation and in all background conditions. Moreover, it also captures the idea that there is a *minimal* degree of invariance that a relationship or generalization must possess if it is to qualify as causal. These points will be especially relevant to our later discussion.

As well as having a threshold, a feature of invariance that is also relevant to the discussion in the next chapter is that invariance comes in varying *degrees*. Significantly, it is the contrast between highly invariant generalizations and relationships, on the one hand, and relatively unstable generalizations and relationships, on the other, that tracks the difference between highly explanatory generalizations and relationships and relatively explanatorily shallow generalizations and relationships.[12] The reason *why* highly invariant generalizations and relationships are also highly explanatory is fairly simple: by being invariant over a wide range of interventions, those generalizations and relationships will simply be able to answer a wider range of w-questions. Moreover, by being invariant over a wider range of changes, those relationships

[11] In saying this I do not mean that X can cause Y even if there is only one single intervention on X (that occurs just once, either hypothetically or actually and could never occur again) that changes Y, since it is built into the notion of invariance that if X causes Y, the invariant relationship between X and Y would be potentially reproducible in the sense that under *this specific* intervention, X would change Y.

[12] The other feature of interventionism that also tracks the difference between *better or worse* causal claims and explanations, and which will be extremely relevant to the argument in this thesis, is the notion of contrastive focus, which I will discuss in detail later in this chapter and in the next.

and generalization will also be more potentially exploitable for the purposes of control and manipulation, in the sense that they will continue to hold and hence continue to provide a potential means of control, over a wide range of interventions.

By contrast, those generalizations and relationships that are less invariant will qualify as less explanatory, since by being invariant over a much more limited range of interventions, they will be able to answer a much more limited range of w-questions. Moreover, those relationships that display a relatively low degree of invariance, whilst allowing some measure of control and manipulation, will be less potentially useful since they will break down outside a narrow range of interventions. By way of further contrast, note that those relationships that fail to be invariant under *any* interventions and hence fail to qualify as causal, will not be potentially useful for the purposes of control and manipulation whatsoever, in line with the manipulationist account of causation outlined thus far.

The notion of invariance thus explains how certain generalizations and relationships can fail to qualify as causal and explanatory (by failing to pass the threshold of invariance), but also explains how generalizations and relationships that do pass this threshold (and hence qualify as causal) can come in varying *degrees* and explains the relative explanatory depth of a generalization or relationship and its potential for control and manipulation in terms of its degree of invariance. As we will see in the next chapter, this feature of interventionism plays a central role in the interventionist account of mental causation, since it explains how mental properties can often be considered as *preferable* causes of

their effects in comparison to their physical realizers, given their relatively high degree of invariance.

As the discussion above should have made clear, in order for it to be true that X causes Y according to interventionism, there must exist some (at least minimally) invariant relationship between X and Y, which ensures that X causes Y, rather than being merely correlated with it. As Woodward (Ibid: 16) explains, we can therefore think of invariance as the feature, *in virtue of which* certain relationships and generalizations qualify as causal; a role that is usually assigned to laws of nature on other accounts of causation. What then is the relationship between laws and invariant generalizations? (Note that this issue is especially relevant to the argument in Section 4.3 below.)

It is immediately apparent that invariant generalizations do not meet one of the presumptive criteria for lawfulness, namely being exceptionless. Now, although some (Cartwright, 1980) argue that there are no truly exceptionless laws, even at the level of fundamental physics, it is usually thought that genuine laws hold without exception and that it is, at least in part, in virtue of being exceptionless that generalizations qualify as laws. For example, since the generalization 'all inertial bodies have no acceleration' is thought to be exceptionless and hence is thought to qualify as a genuine physical law, the status of this generalization as a law would be undermined by even one instance of an inert accelerating body (Carroll, Spring 2012).

By contrast, generalizations can qualify as invariant and genuinely causal and explanatory, even if there are some, if not many exceptions to those generalizations. This is because, as I explained above, it is only necessary for some generalization to qualify as invariant that there is *some* intervention on the

cause variable that changes the effect variable, allowing for the possibility that there are some (possibly many) exceptions to those generalizations.

For example, the generalization 'Smoking causes cancer' would be invariant and hence qualify as causal and explanatory to the extent that the variable 'cancer' occurs more frequently when the variable 'smoking' is introduced via interventions than when smoking is absent.[13] This remains true even though this relationship has exceptions (for example, some individuals may smoke and yet fail to develop lung cancer). Moreover, although this generalization may well be explanatorily shallow in comparison to a generalization which cites the biological mechanisms[14] involved in the relationship between smoking and cancer, it is important to emphasise that *both* kinds of generalizations can qualify as genuinely causal and explanatory according to interventionism, given that they both qualify as minimally invariant.

We can therefore see that whether a generalization qualifies as invariant and hence causal and explanatory is fairly independent of whether it meets one of the presumptive criteria for lawfulness, namely being exceptionless. For Woodward, this is significant because it means that interventionism is able to avoid a dilemma that other accounts of causation that appeal to a traditional account of laws inevitably face. As Woodward (2003: 239) explains, a dilemma arises because on the traditional account, it is assumed that laws (understood to be exceptionless) are required for causation and successful explanation. Then,

[13] Citing prior research, Woodward (2003: 312) explains that since this relationship does remain invariant across a range of circumstances, which control for confounding variables, such as gender, genetic background, variations in environment and diet and so on, it can be considered as a genuine causal generalization.

[14] This is not to imply that mechanistic causal explanations are *guaranteed* to provide preferable explanations of some effect in comparison to 'higher-level' (for example, sociological, psychological) explanations of some effect, since this depends on the *degree of invariance* that the relationships cited in those explanation possess, and/or on which explanation captures the correct contrastive focus.

given that special science generalizations do not appear to meet this criterion, it seems that one would be forced to conclude either that special science generalizations are not laws and hence are not genuinely causal or explanatory, or that they are laws, but that they need to be qualified, hence the many complex arguments for ceteris paribus laws. Interventionists are simply able to avoid this dilemma, since according to interventionism, special science generalizations can qualify as invariant and hence causal and explanatory, even if they do not meet the traditional criteria for lawfulness. Consequently, interventionism provides a useful and convincing account of the generalizations of the special sciences and most importantly for our purposes, of the generalizations of psychology.[15]

Does this mean that there are no such things as laws according to interventionism? Not necessarily. As Woodward notes, there may be examples of invariant generalizations, such as the gas laws in fundamental physics that do meet the traditional criteria for lawfulness and that may rightly be called laws, or even laws of nature. However, the crucial point to emphasise is that these laws of nature are not fundamentally different in *kind* to the 'loose generalizations' of the special sciences; laws of nature are simply generalizations that display a very high degree of invariance, whereas the generalizations of the special sciences will typically display a lower degree of invariance. As Woodward puts it, "rather than thinking of all invariant generalizations as laws, I urge instead that we think of laws as just one kind of invariant generalization." (Ibid: 267)

[15] Note also that if one does think that laws are required for causation and explanation, then the generalizations of the physical sciences, especially physics, will be considered as preferable, given that it is arguably only at this level that one is likely to find generalizations that possess the standard criteria for lawfulness. By contrast, according to interventionism, there would be no *automatic* preference for the generalizations of the physical sciences, given that invariant generalizations can exist at any level.

One final feature of interventionism that will be useful to highlight is the distinction between type and token causation (or as Woodward calls it 'actual causation' or AC). As should be clear from the discussion above, a type-level causal claim, such as 'Smoking causes cancer', implies that some token-level causal claim, such as 'Smith's smoking caused his cancer', *would* be true, but does not depend for its truth on the actual obtaining of any such particular occurrence. This is because all that matters for whether this type-level claim qualifies as causal is that there exists some intervention on smoking that changes the occurrence of cancer and this may be true even if the intervention is merely hypothetical, or even if it is not practically, physically, or even nomically possible (more on this below).

By contrast, token causation does imply the truth of some type-level generalization. This is because, in order for it to be true that X is a token, or actual cause of Y according to interventionism, there must exist some (type-level) invariant relationship between the variables. It is important to be clear on two things, the relevance of which will become clear in Section 4.3 below. Firstly, this is not to say that the associated type-level generalizations will always be highly invariant and hence law-like. For example, the type-level generalization associated with the token-level causal claim, 'Smith's smoking caused his cancer', (namely 'Smoking causes cancer') displays a relatively low degree of invariance and, as discussed above, does not meet one of the standard criteria for lawfulness, namely being exceptionless. Secondly, this is not to say that the user of the token causal claim will always be explicitly aware of the associated type-level generalization, or that it is only in virtue of this explicit knowledge that a subject can acquire causal understanding of token-level causal

claims. This is because according to interventionism, in order for X to qualify as an actual cause of Y, there must exist some intervention that changes the *actual* value of X to some other value that changes the *actual* value of Y to some other value and although this implies the truth of some invariant type-level generalization between X and Y, we can consider the truth of these interventionist counterfactuals independently of any explicit knowledge of this associated type-level generalization.

Now, this is obviously not to say that our causal understanding of token-level causal claims is *never* explicitly accompanied or supported by knowledge of some type-level generalization. For example, the token causal claim, 'Smith's smoking caused his cancer', may be accompanied and explained by the type-level generalization 'Smoking causes cancer'. Rather, the point is simply that it will often not be accompanied or supported by any such explicit knowledge.[16] To use Woodward's example, "I may know with confidence that a blow on the head caused Jones's death, even though I do not know any relevant nontrivial deterministic generalization about the circumstances under which blows on the head are followed by death." (Ibid: 75) The relevance of these points will become clear in Section 4.3 below.

[16] This issue of causal understanding and in particular, the relationship between singular causation and type-level generalizations is a complex one (these issues are discussed at length in, for example, Anscombe (1981), Strawson (1992), Hitchcock (1995) and more recently, in Roessler (2011)). Although an in depth discussion of these issues is beyond the scope of this thesis, it is sufficient for the purposes of my argument that the reader finds it at least plausible that a subject can acquire causal understanding of singular causal claims without at least any *explicit* knowledge of an associated type-level generalization. The reason why this is sufficient will become clear in Section 4.3 below.

4.2.2 Interventions

The definition (M) of interventionism outlined above states that it is a necessary and sufficient condition for X to cause Y that there exist a possible intervention on X that changes Y. The notion of an intervention is therefore central to interventionism and I will now examine this notion in detail.

As noted above, it is useful to think of an intervention as an idealised experiment (either hypothetical[17] or actual) that determines whether X causes Y. Woodward captures the basic idea of an intervention in the following passage:

> "…an intervention on some variable *X* with respect to some second variable *Y* is a causal process that changes the value of *X* in an appropriately exogenous way, so that if a change in the value of *Y* occurs, it occurs only in virtue of the change in the value of *X* and not through some other causal route." (Ibid: 94)

This passage raises several important questions: What are the precise criteria that an intervention must meet if it is to be considered as a suitable means for determining whether X causes Y? What is it for an intervention to be exogenous? In what sense must an intervention be possible?

Before answering the first question and outlining the specific criteria that a 'suitable' intervention must meet, it is important to point out that in specifying such criteria, I take it that Woodward's intention is not to provide conditions for

[17] Williamson (2007, see especially Chapter 5) argues that the *imagination* can be (and often is) successfully used to evaluate the truth of counterfactuals and consequently can be used to acquire genuine (and as interventionists suppose, *causal*) knowledge.

a 'perfect' experiment for determining whether X causes Y, but rather, to provide idealised conditions that specify what *must* be true in order for X to cause Y (i.e. specifying the truth conditions for interventionist causation). An intervention should therefore be understood as an 'idealised' experiment, which determines what *would* happen to an effect under a suitable intervention, rather than being thought of as a 'perfect experiment' that actually must take place in order for us to be able to make a causal judgement.[18]

What then are the conditions for such an idealised experiment? Woodward provides the following criteria for a 'suitable' intervention[19]:

> "**(IV)**
>
> I1. *I* causes *X*.
>
> I2. *I* acts as a switch for all the other variables that cause *X*. That is, certain values of *I* are such that when *I* attains those values, *X* ceases to depend on the values of other variables that cause *X* and instead depends only on the value taken by *I*.
>
> 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

[18] Note that interventionism does not claim that the only way to *learn* about causal relationships is through performing an actual or hypothetical intervention. As Roessler (2011) argues, we can acquire causal knowledge through passive observation or perceptual experience, for example. Nonetheless, interventionism does claim that the causal knowledge that one acquires in these kinds of cases is still interpretable in interventionist terms, i.e. they provide us (perhaps only implicitly) with information about what would happen *were* we to perform an intervention. This issue concerning the psychology of counterfactual reasoning is interesting, but is beyond the scope of this thesis. This issue is discussed in detail in Hoerl, C. McCormack, S. Beck, S. R. (eds.) (2011) and in Woodward (2011b).

[19] It is also worth noting that by appealing to these precise criteria to define the notion of an intervention and to determine whether some relationship is genuinely causal, interventionism does not need to appeal to a similarity metric between possible worlds, as Lewis' (1977a) account does. It is therefore able to avoid the problems that are associated with this account, such as problems concerning the apparent vagueness of similarity judgements (Fine 1975) and problems concerning the potential subjectivity of judgements of similarity.

> 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*." (Ibid: 98)[20]

To begin with the second criterion[21], this essentially expresses the idea that the intervention on X should act alone in manipulating X, as this ensures that any causal relation that is established between X and Y is known to be a result of intervention *I* alone, rather than a result of some other cause of X that could also cause Y. This essentially ensures that the idealised experiment is a controlled one in that it is only the influence of *I* on X that is under consideration at any one time.

The third criterion ensures that the intervention must go through the cause variable that is under consideration, rather than directly causing the effect itself, or causing some other variable that also causes the effect. For example, imagine that we are trying to establish whether some drug is effective in treating a particular disease. The third criterion rules out the possibility of any kind of 'placebo effect', in which the intervention of administering the drug directly causes recovery itself, potentially confounding any relationship between the drug

[20] Again, it is important to point out that although this definition is accurate, I present a slightly modified definition of these criteria in Chapter 6 in order to deal with the objection raised by Baumgartner (2009, 2010).

[21] I address a potential problem of circularity that arises as a result of the first criterion in Section 4.5 below.

and recovery. Or alternatively, imagine that we want to find out whether attendance at private school causes scholastic achievement, but that the experiment that divided the children into two groups (one of which attended private school and one of which attended a government funded school) failed to be suitably randomised, such that the group of children selected to attend private school also belonged to households with a higher socio-economic status. Since this intervention affects attendance *and* socio-economic status, which is a cause of achievement that is independent of the attendance-achievement relationship, it would not be a suitable intervention for establishing whether attendance causes achievement.

Similarly to the third criterion, the fourth criterion rules out the possibility that intervention *I* could directly cause some other variable Z that is also a cause of Y that does not go through X, again ruling out the possibility that Z could confound the relationship between X and Y.

The idea that an intervention should be 'exogenous' is also captured by the criteria above in the sense that intervention *I* must come from 'outside' the system under consideration in order to ensure that its influence is independent of, and breaks any ties with, any endogenous, confounding variables.

This leaves us with the following definition of an actual intervention:

> "**(IN)** *I*'s assuming some value $I = z_i$, is an intervention on *X* with respect to *Y* if and only if *I* is an intervention variable for *X* with respect to *Y* and $I = z_i$ is an actual cause of the value taken by *X*." (Ibid)

With the notion of an intervention outlined, we can now say that a relationship between X and Y is minimally invariant and hence causal, so long as there exists a possible intervention on X, which meets *these specific* criteria, and which brings about changes in Y.

The definition of an intervention outlined above makes reference to the existence of a *possible* intervention. How should we understand this notion of possibility? Firstly, as I have noted above, in order for X to cause Y, the intervention on X need not actually be carried out, but may instead take the form of a hypothetical intervention that considers what *would* happen to Y if we *were* to intervene on X. Secondly, it is important to note that interventionism operates with a fairly permissive notion of possibility. For example, interventionism *only* requires that these interventions (either hypothetical or actual) be logically, conceptually and metaphysically possible, rather than being practically, nomically (i.e. that they must conform to the laws of nature of this world) or even physically possible (Woodward, 2003: 127-133). For Woodward, the reason why it is not necessary that interventions be within the realm of practical possibility is that it is simply not relevant to the coherence of the interventionist counterfactuals (and to our assessment of the truth of the counterfactuals) that anyone should actually be able to perform the interventions. For example, it is possible to consider the causal claim 'the impact of an asteroid caused the extinction of the dinosaurs' in interventionist terms, given that this claim coveys information about what *would* have happened to the dinosaurs had there been no asteroid strike as a result of some intervention, even though it is obviously not practically possible for anyone to carry out this intervention. For the same reason, it is unnecessary that interventions must be physically, or even nomically

possible. It is also worth noting that it is for this reason that interventionism is able to avoid the problems concerning anthropocentricism discussed in Section 4.5 below, since the notion of an intervention is defined explicitly without reference to the action of any agent who actually does, or ever could perform an intervention.

4.2.3 Causal Explanation

The interventionist approach to causal explanation has been discussed briefly above, but it will be useful to now examine, in detail, how interventionism conceives of the relationship between causation and causal explanation, since this discussion will be especially relevant to the argument against the assumption of SP discussed below and to the interventionist account of mental causation outlined in the next chapter.

In line with the manipulationist account of causation outlined thus far, the distinguishing feature of all causal explanations, according to interventionism, is that they are explanations "that furnish information that is potentially relevant to manipulation and control: they tell us how, if we were able to change the value of one or more variables, we could change the value of other variables." (Ibid: 6). Non-causal explanations, by contrast, provide no such information relevant to potential control and manipulation. Again, this manipulationist conception of causal explanation reflects the practical focus of causation and causal explanation, in the sense that causal explanations should provide us with information with which we can potentially acquire greater control over our environment.

The fact that all causal explanations provide information that is potentially useful for control and manipulation is a natural consequence of the fact that, on this account, causal explanations cite genuine causes. It is thus information about causal relationships, which necessarily exhibit patterns of counterfactual dependence between cause and effect, which provides this information about potential control and manipulation. As I mentioned above, all causal explanations thus share the feature of being able to answer a range of w-questions, in the sense that they provide counterfactual information that allows us to consider what *would* happen to the effect given various changes to the cause.

For example, the explanation 'attendance at private school causes scholastic achievement', counts as a genuine *causal* explanation, since it provides counterfactual information about what *would* happen to the effect (scholastic achievement) if we were to manipulate the cause (attendance at private school). (For example, it tells us that if we were to intervene on attendance, we would bring about a change to scholastic achievement.) As a result, this explanation provides us with the information that intervening on attendance is a way of *controlling* and *manipulating* scholastic achievement. By contrast, consider Woodward's example of the explanation of why raven *a* is black:

All ravens are black.

a is a raven

a is black

According to interventionism, this explanation fails to qualify as a genuine *causal* explanation since it fails to exhibit any pattern of counterfactual dependence between the explanans and explanandum (according to

interventionism, this is explained by the fact that this explanation simply does not cite the *cause* of raven *a*'s colour). For example, it does not tell us about the conditions under which raven *a* would be a different colour, or how we might go about changing the colour of raven *a* or any other bird for that matter.[22]

Moreover, as I explained above, according to interventionism, causal explanations will be considered as better or worse to the extent that the generalizations that they refer to display either a relatively high or low degree of invariance and this is simply because highly invariant generalizations will, in general, be able to answer a wider range of w-questions. Moreover, by tracing the explanatory depth of a generalization to its ability to answer a wider range of w-questions, Woodward again emphasises the practical focus of interventionism: deeper or better causal explanations are simply those that provide us with more information that is potentially relevant to controlling and manipulating our environment.

Although the relationship between causation and causal explanation is a very close one according to interventionism (given that causal explanations cite genuine causes), one important difference between causation and causal explanation is that whereas causation is thought to be a natural relation that exists (or would exist) 'out there' in the world, independent of our epistemic awareness of it, causal explanation is essentially an epistemic activity that is carried out in order to acquire information about causal relationships. This difference is captured in the following passage:

[22] This example also emphasises the point made above, since it is the fact that there is no logically or conceptually possible intervention associated with this explanation that guides our intuition that it is non-causal. The notion of physical or nomic possibility does not appear to have the same effect. (Consider again the example of the explanation of the extinction of the dinosaurs, for which there is no physically possible intervention, but which nonetheless strikes us as causal.)

> "Causal relationships are features of the world: they are 'out there' in nature. By contrast, explanation is an activity carried out by humans and conceivably by some other animals, having to do with the discovery and provision of *information*, information about causal relationships." (Ibid: 23)

As we shall see, this epistemic constraint (i.e. the constraint that causal explanations should provide information that is *epistemically available* to us) plays a central role in my argument against the SP concept of causal explanation in Section 4.3 below.

I explained above that the claim 'attendance at private school causes scholastic achievement', qualifies as genuinely causal since there is some intervention on attendance at private school that changes scholastic achievement. Moreover, I explained that this explanation qualifies as a genuine *causal* explanation, in virtue of the fact that it conveys this counterfactual information about what would happen to scholastic achievement if we were to manipulate attendance at private school. However, it is not always this clear that the counterfactuals that are associated with some causal claims and explanations deliver consistent or intuitively correct causal judgments about those claims and explanations.

In order to illustrate this, consider Woodward's example of the token-level, or singular claim, 'The short circuit caused the fire'. Now, assuming that the short circuit does cause the fire and that there are no pre-emptive or overdetermining causes of the fire, according to interventionism, this singular

claim qualifies as causal since there is some intervention on the occurrence of the short circuit that changes the occurrence of the fire. Moreover, this explanation qualifies as genuinely causal in virtue of the fact that it conveys this counterfactual information about what would happen to the effect if we were to manipulate the cause. For example, the associated counterfactual, 'if the short circuit had not occurred, the fire would not have occurred', (which Woodward refers to as 5.8.4), tells us that intervening on the short circuit would be a way of intervening on the occurrence of the fire and hence tells us that the explanation is genuinely causal.

However, as Woodward points out in the following passage, there appear to be at least some counterfactual alternatives (i.e. some interventions on the non-occurrence of the short circuit) for which the counterfactual 'if the short circuit had not occurred, the fire would not have occurred' is not true and according to *these* counterfactual alternatives, it looks as though the explanation 'the short circuit caused the fire' does *not* qualify as causal.

> "How exactly should we understand a phrase like "if the short circuit had not occurred" in (5.8.4)? It seems that there are, so to speak, a variety of different possible ways in which "the" short circuit might have failed to occur and that (5.8.4) may be true under some of these possibilities but not under others. Suppose that the actual short circuit *s* occurred at a specific time *m* and reached a certain temperature *T*. Consider a short circuit *s** that occurs at a somewhat different time *m** and reaches a different temperature *T**. Will *s** be the same short circuit as the actual short circuit *s*? If, as seems arguable for some values of *m* and *T*, the

> answer to this question is no, then one of the ways the (actual) short circuit *s* might fail to occur is if the different short circuit *s** occurs instead. But, if the antecedent of (5.8.4) is understood to include this sort of possibility—that is, if the nonoccurrence of *s* in (5.8.4) is understood to encompass the occurrence of *s**—then it is far from obvious that the counterfactual (5.8.4) is true." (Ibid: 211-2)

How then can we avoid the problem that it looks as though this singular causal claim qualifies as causal and explanatory according to some counterfactual alternatives, but as non-causal according to others?

Firstly, it is important to emphasise that it is simply *not* true according to interventionism that the singular causal claim 'the short circuit caused the fire' qualifies as causal and explanatory according to some counterfactual considerations, but as non-causal according to others. Remember that according to interventionism, this explanation will qualify as causal so long as there is at least *some* intervention on the value of the short circuit that changes the value of the occurrence of the fire and I explained that this remains true even if there are some counterfactual alternatives for which this is not true.

Rather, the issue that has been highlighted points to a potential lack of clarity and consistency in our causal judgements about causal claims and explanations, given that the same causal claim can nonetheless *appear* to qualify as causal according to some counterfactual considerations, but *appear* to qualify as non-causal according to others. How then can we resolve the problem that the interventionist criteria for causation and explanation seem to generate somewhat inconsistent and potentially incorrect causal judgements? In order to address this

problem, Woodward appeals to the notion of contrastive focus, which was introduced above and which I will now examine in more detail.

Remember that according to interventionism, all causal claims have built into them a default contrastive focus, which tells us that it was the fact that the cause variable took *this* value, *rather than that* value that caused the effect variable to take the value it did. Consequently, a 'good' causal explanation is one that identifies exactly which changes to the cause variable are associated with changes to the effect variable. An explanation is deficient to the extent that it fails to do this (either because it omits vital information, or as is often the case, because it is overly specific).

In order to illustrate this, consider Stephen Yablo's (1992) example: a pigeon has been trained to peck on the presentation of any red object. On a particular occasion, the pigeon is presented with an object that happens to be a particular shade of scarlet and the pigeon proceeds to peck. Yablo asks to what we should attribute as the cause of the pigeon's behaviour: is it the redness of the object or the fact that it is scarlet? Yablo concludes that although the fact that the object is scarlet is sufficient for the behaviour (note that it also meets the interventionist requirements of causation, since there is *some* intervention on the property of scarlet, namely one that changes the colour from scarlet to any non-red shade, that changes the effect), it fails to capture what is relevant about the object that causes the pigeon to peck, namely the fact that it is red. In fact, Yablo claims that citing the fact that the object is scarlet as an explanation of the behaviour is actually misleading since it suggests that the pigeon would fail to peck in any case in which the object is not scarlet. In Yablo's terms, the explanation citing scarlet fails to be 'proportionate' to its effect, in the sense that

it fails to convey *all and only* such information about specific patterns of counterfactual dependence between cause and effect, (in this case, by both omitting relevant detail about such dependencies and including irrelevant detail). By contrast, the less specific explanation citing the property of red captures this information.[23]

Put into Woodward's terms, the explanation citing scarlet fails to capture the correct contrastive focus, since it fails to capture exactly which changes to the cause variable are associated with changes to the effect (namely a change from 'red' to 'not red') and in fact provides potentially misleading information about such patterns of counterfactual dependence. As a consequence, the explanation citing scarlet will be deficient in comparison to the one citing redness. Note also that by failing to capture the exact range of changes to the cause variable that are stably associated with changes to the effect (and in fact by providing potentially misleading information about such changes), the explanation citing the property of scarlet will provide information that is less useful for the purposes of control and manipulation. By contrast, given that it is specifically the contrast between whether the object is red/not red that is associated with whether the pigeon pecks/does not peck, the explanation citing the property of red will provide information with which we may *stably* and *systematically* control the effect. (These points are especially relevant to the interventionist account of mental causation outlined in the next chapter.)

Before demonstrating how this notion can help us with the problem identified above, it will be useful to highlight the following feature of contrastive

[23] Williamson (1998) argues somewhat similarly that 'good' explanations have an appropriate generality built into them and that explanations do not necessarily get 'better' by being more specific.

focus, since it will be relevant to our later discussion: whether some claim or explanation captures the correct contrastive focus can depend on the context of the situation and on the somewhat subjective consideration of our goal as enquirers. For example, the context dependence of contrastive focus can be seen in Woodward's (2008a: 236) variant of the Yablo example in which the pigeon is trained to peck specifically at scarlet objects. In this case, it is now the contrast between whether the object is scarlet, rather than not scarlet that is associated with changes to whether the pigeon pecks, or fails to peck, rather than the contrast between whether the object is red or not red. Hence the explanation citing the property of scarlet now captures the correct contrastive focus and provides a preferable causal explanation of the behaviour.

As an illustration of the way in which contrastive focus can depend on the goal of the enquirer, consider Woodward's (Ibid: 227) example of the platform that will collapse if a weight greater than 1000kg is placed onto it. Suppose that a weight that happens to weigh 1600kg is placed onto the platform and the platform collapses. Now consider causal claim (1), which states that it is the fact that the weight is greater than 1000kg that caused the platform to collapse, compared with causal claim (2) which states that it is the fact that the weight is 1600kg that caused the platform to collapse. Now, according to interventionism both claims qualify as causal, since there is an intervention on both whether the weight is 1000kg and whether it is 1600kg that changes the effect. However, since it is the contrast between whether the weight is greater than 1000kg, or less than 1000kg that specifically changes whether the platform collapses or fails to collapse, causal claim (1) captures the correct contrastive focus and will be considered as preferable in comparison to causal claim (2), which by being

overly specific, omits this information and provides potentially misleading information about the conditions under which the platform will collapse.

Now imagine that we are interested not just in why the platform collapsed, but in why it collapsed at such and such a velocity. I suggest that given our new explanatory goal, it is now causal claim (2) that would capture the correct contrastive focus and which would be considered as preferable, in comparison to causal claim (1), since it is now changes to whether the weight is specifically 1600kg, or some other specific weight that would be associated with changes to the specific velocity of the collapsing platform. The relevance of these points will become clear later in this chapter and in the next.

How does the notion of contrastive focus help us with the problem highlighted above? As Woodward explains, when dealing with complex causal claims, for which it is not immediately clear exactly which counterfactuals we should appeal to in order to assess the truth of the causal claim or explanation, we should appeal to the notion of contrastive focus to help us to determine this.[24] For example, since in the case of the short circuit, it is the contrast between the occurrence of the short circuit and a situation in which no short circuit occurs *at all* that explains the contrast between the situation in which the fire occurs and a situation in which no fire occurs, we should consider counterfactuals relating to this contrast as most relevant for assessing the truth of the causal claim. By appealing to those counterfactuals that are associated with the contrastive focus of the claim, it is possible to avoid the problem outlined above that it appears that

[24] Woodward notes that this is especially useful when the values of the variables under consideration are not simply 'present' or 'not present', but instead take many different values.

the same causal claim comes out as true under some counterfactual considerations, but not others.

Once again, it is important to be clear that this is not to suggest that the notion of contrastive focus should be used to constrain the counterfactual truth conditions for interventionist causation.[25] Remember that according to interventionism, the causal claim involving the short circuit will come out as true so long as there is *some* intervention on the actual value of the short circuit that changes the actual value of the occurrence of the fire, even if there are some counterfactual alternatives for which this is not true *and* even if the counterfactuals associated with the correct contrastive focus of the claim come out as false (the relevance of this last point will become especially clear in the next chapter). Rather, what has been suggested is that when it is not clear exactly which counterfactuals to appeal to when assessing the truth of some causal claim or explanation, or when different counterfactuals deliver different causal judgments, we can appeal to the notion of contrastive focus to help us to determine which counterfactuals are *most* relevant to consider.

4.3 Interventionism versus the SP Concept of Causation: Problems for the SP Concept

In this section I draw attention to some problems that the SP concept of causation that was introduced in the previous chapter faces and contrast this account with the interventionist account of causation and causal explanation outlined thus far. By highlighting the problems that the SP concept faces and

[25] If the counterfactual truth conditions for interventionist causation were constrained in this way, then contrastive focus or proportionality would, in effect, become a necessary condition for interventionist causation. I discuss a problem with this approach in Chapter 5 in which I examine two alternative interventionist accounts of mental causation that adopt this approach.

presenting interventionism as a coherent alternative theory of causation that avoids these problems, I aim to undermine the assumption that causation is identical to sufficient production and (if my argument in the previous chapter is correct) demonstrate that the non-reductive physicalist need not accept Kim's a priori exclusion problem.

In order to do this, I argue, in Section 4.3.1 that the SP concept of causation does not provide *necessary* conditions for causation. Then, in Section 4.3.2, I argue that providing nomologically sufficient conditions for the occurrence of some effect is not *sufficient* for causal explanation and demonstrate that interventionism is able to avoid these problems. I claim that these arguments undermine the assumption that causation is identical to sufficient production and undermine the explanatory counterpart of the assumption of SP, which states that providing a causal explanation of some effect is simply a matter of identifying nomologically sufficient conditions for its occurrence. Lastly, in Section 4.4, I address the worry that despite the problems that the SP concept faces, interventionism fails to provide a viable alternative to this theory and hence fails to undermine the assumption of SP, since it faces serious problems of its own, which the SP concept seems to avoid. I demonstrate that not only can interventionism overcome these problems, but it is able to deal with many of these problems in a more satisfying way than the SP concept. I conclude that interventionism does, after all, provide a coherent alternative theory of causation, which undermines the assumption of SP and hence demonstrates that the non-reductive physicalist need not accept Kim's a priori exclusion problem.

4.3.1 Is Sufficient Production Necessary for Causation?

As I introduced it in the previous chapter, the SP concept of causation states that it is a necessary and sufficient condition for X to cause Y that X *produces*, *generates* or *determines* Y's occurrence and that X is a *sufficient* cause of Y, where 'cause' is understood in this productive/generative sense. In this section, I argue that sufficient production does not appear to provide *necessary* conditions for causation, undermining the assumption that causation is identical to sufficient production. The section is divided into two parts.

The first set of examples and arguments that I present argue that being sufficient, or more accurately, being *nomologically* sufficient is not necessary for causation. (Recall that in Chapter 3 I noted that we should understand the idea that causes are sufficient for their effects as entailing the view that causes are *nomologically* sufficient for their effects). More accurately, they demonstrate that it is not part of the *meaning* or *concept* of causation that underlying laws exist. I argue that this is sufficient for the purposes of my argument, since it nonetheless undermines the assumption that causation is *identical* to (nomological) sufficient production.

The second set of examples and arguments that I present suggest that production, generation and determination are not necessary for causation. Since the arguments in this section cast doubt on the idea that sufficient production is *necessary* for causation and since I demonstrate that interventionism is simply able to avoid these problems, I conclude that they undermine the assumption that causation is identical to sufficient production.

To begin, it will be helpful to appeal to a series of arguments originally proposed by Woodward against the Deductive-Nomological, or DN model of explanation proposed by Carl Hempel (1969).[26] As I will demonstrate, these arguments work equally well against the SP concept of causation, since the DN model *also* assumes that laws are necessary for causation.

To begin, consider the following singular causal claim:

> "(2.4.1) The impact of my knee on the desk caused the tipping over of the inkwell." (Woodward, 2010b)[27]

Now, according to the SP concept of causation, in order for 2.4.1 to express a genuine causal claim, it must be true that some law is instantiated, which guarantees that the cause is sufficient for the occurrence of the effect, given certain initial conditions. However, as Woodward points out, 2.4.1 does not, at least explicitly, appear to instantiate any law, for example, one that would lawfully link the impact of knees on desks with the tipping over of inkwells. Nonetheless, 2.4.1 does intuitively appear to express a genuine causal claim. Thus, it appears, at least at first glance, that (nomological) sufficient production is not necessary for causation, since 2.4.1 appears to express a genuine causal claim, while failing to meet the nomological requirement of the SP concept of causation.

There are two responses noted by Woodward that Hempel makes to preserve the necessary status of laws on the DN model. I outline these responses

[26] Although Hempel's theory is specifically concerned with the nature of explanation, rather than causation, since Hempel himself accepts that at least some explanans cite causes, we can assume that a nomological sufficient conception of causation features in the DN model of explanation.
[27] Example originally due to Scriven (1962).

below and agree with Woodward that neither response deals adequately with this problem. I demonstrate that these arguments count equally well against the SP concept of causation.

Hempel's first response is to claim that 2.4.1 is after all backed by the instantiation of a law, but that the law is 'implicit' in 2.4.1, rather than explicit. As Woodward explains, this implicit law could take the following form:

> "(2.4.2) Whenever knees impact tables on which an inkwell sits and further conditions *K* are met (where *K* specifies that the impact is sufficiently forceful, etc.), the inkwell will tip over. (Reference to *K* is necessary since the impact of knees on table with inkwells does not always result in tipping.)" (Ibid)

Hempel's second response is to add that while it may be true that the entire complex content of 2.4.2 is not implicit in 2.4.1, we should understand 2.4.2 as an 'ideal' explanation of the effect, in contrast with 2.4.1, which provides only a 'partial explanation'.

According to Hempel then, on either response we can assume that some law implicitly 'underlies' and grounds the causal status of 2.4.1, whether this underlying complex law is implied completely by 2.4.1, or whether it is only partly implied by 2.4.1. Is it true then that despite initial appearances, laws are necessary for causation, in accordance with the SP concept of causation? Woodward offers several convincing objections to both of Hempel's responses and consequently against the SP concept of causation, which I outline below.

In response to Hempel's first point, Woodward questions exactly what *kind* of law Hempel takes to implicitly underlie 2.4.1, since it is not clear which, if any, of the various kinds of laws are supposed to fulfil this role. Firstly, as Woodward explains, it could be something like 2.4.2, however, this assumes that condition K can be specified non-trivially. Presumably, Woodward's worry is that we could just add conditions to 2.4.2 to ensure that the effect occurs given the instantiation of this 'law', but that it would be difficult to provide such conditions non-trivially and non ad-hoc.

This point brings out a related problem, which is the general problem of defining exactly what a law is. As Woodward points out, there is little consensus in the literature as to what a law actually is and although this problem does not prove that the DN *or* SP concept of causation do not provide necessary conditions for causation, it does suggest that these theses require clarification. This is because without a clear understanding of what a law is, it is not clear that the claim that causation requires the backing of laws is coherent

Secondly, Woodward explains that the underlying law could be one specified at the level of classic physics, for example one that referred to the behaviour of liquids when not confined to a container. However, as Woodward points out, the problem with this approach is that it is highly unlikely that those laws could be known by an ordinary user of 2.4.1 and hence it is highly unlikely that our causal *understanding* of 2.4.1 is acquired in virtue of knowledge of those laws. Put slightly differently, although the fact that an ordinary user of 2.4.1 is unlikely to be aware of these laws does not imply that the laws of physics do not in fact underlie 2.4.1, it does make the idea that we understand or recognise 2.4.1

as causal *in virtue* of knowledge of these underlying laws look highly dubious. I return to this epistemological issue below.

In response to Hempel's second point, Woodward questions the sense in which the more specific, 'lawful' explanation provided by 2.4.2 should be considered as explanatorily 'ideal', in comparison to the seemingly straightforward explanatory claim offered by 2.4.1, given that it is at best only implicit, or as Woodward puts it, given that it is 'epistemically hidden' in 2.4.1 and therefore unlikely to be known by an ordinary subject.

Now, I suggest that these points count against the assumption of SP *and* the explanatory counterpart of the assumption of SP, which, as I described it in Chapter 3, states that providing a causal explanation of some effect is simply a matter of identifying nomologically sufficient conditions for its occurrence. This is because once one recognises that the nomological aspect of causation and explanation will often be either implicit in the causal claims and explanations or epistemically unavailable to an ordinary subject, it undermines the idea that causation is *identical* with (nomological) sufficient production and undermines the idea that causal explanation is simply a matter of providing such (nomological) sufficient conditions.

This point can be brought out further by appealing to the following example of P.F Strawson's (1992), (originally due to Mill): a man falls down a flight of steps; the fact that the steps were slippery and the fact that the man's mind was elsewhere is offered as a sufficient explanation of the event. However, as Strawson correctly observes, there are no general regularities or universal laws linking fallings, slipperyness of steps and absent-mindedness that could ground our causal understanding of this singular causal claim. Strawson does note that in

such cases of singular causation, there may in fact be *some* general, mechanistic law (presumably found at the level of physics), which underlies that singular causal claim, but he argues that such general and mechanistic laws would be quite abstract and removed from any causal understanding we might have of the singular causal claim. Strawson suggests that Mill's account is therefore quite "wide of the mark" (Ibid: 127) in so far as ordinary causal explanation is concerned and concludes that this points to a "great gap" (Ibid) between ordinary causal explanation and strict law.

To summarise, what both Strawson's and Woodward's examples suggest is that our causal *understanding* of causal claims and explanations is not acquired in virtue of knowledge of laws, given that any knowledge of those laws is likely to be either implicit in the causal claim or epistemically unavailable to an ordinary subject. This point is further supported by the fact that we seem to have an intuitive grasp of the causal status of singular causal claims, such as 2.4.1, in the absence of any knowledge of underlying laws.

Now, one could reply that the problems that I have identified merely prove that knowledge of laws is not necessary for causal *understanding*, but that this is very different to proving that laws are not in fact necessary for causation and that the latter is required to undermine the assumption of SP. This is an important point and does require further discussion, however, I suggest that this epistemological issue *does* have an important bearing on the plausibility of the SP concept of causation and consequently on the assumption of SP.

This is because what Woodward's and Strawson's examples *do* suggest is that although it may be true that some law necessarily underlies all causal relationships, our understanding of those relationships and explanations as causal

is not acquired *in virtue* of knowledge of those laws. This epistemological point is sufficient to cast doubt on the idea that it is part of the very meaning or concept of causation that some underlying law exists (a view that Woodward calls the 'meaning thesis') and this is sufficient to undermine the assumption that causation should be defined in terms of, or *identified* with (nomological) sufficient production.[28]

In fact, as Woodward explains, although the idea that some law necessarily underlies all causal claims (which Woodward calls the 'underlying thesis'), may in fact be true, if it is true, it is simply an empirical fact that those laws exist, rather than something that should form part of the meaning or concept of causation. As Woodward explains in the following passage, at the most basic level, the truth of the underlying thesis may simply commit us to the truth of physicalism[29]:

[28] Again, one could reply here that although it may be true that it is not part of the meaning of causation that some underlying law exists, this does not undermine the idea that causation just *is*, in reality, a relationship of nomological sufficient production. By way of analogy, one could argue for example, that although it is not part of the meaning of 'water' that it is H_2O, water just *is*, nonetheless, H_2O. I cannot discuss this issue in detail here, but it seems that this objection would only be decisive if one favoured an empirical analysis of causation, which takes as its primary focus the discovery of what causation is in reality, as opposed to a conceptual analysis of causation, which seeks to understand the concept of causation as it is used in ordinary language. This is because empirical analyses of causation are not undermined by the fact that their resultant theories of causation do not accord with ordinary usage or form part of the meaning of causation, as it is understood in ordinary language. The theories produced via conceptual analysis will, by contrast, be constrained by ordinary usage and be undermined by the fact that the theory does not form part of the meaning of causation, as it is understood in ordinary language. Thus, so long as one finds the conceptual analysis approach more plausible than the empirical analysis approach, then this objection will not be very forceful. This is not to say that interventionism fits straightforwardly into the category of conceptual analysis. Note, for example, that a central feature of interventionism is that it makes normative claims about how causation *should* be understood, rather than merely analysing how the concept is used (Woodward, 2003: 7). Nonetheless, since interventionism does seek to describe both ordinary and scientific usage of the concept of causation and relates the concept to practical notions such as control and manipulation, it can be broadly considered as providing a conceptual analysis of causation.

[29] This accords with the *a posteriori* definition of causal closure outlined in Chapter 2.

> "In this sense, it seems unlikely that we will ever find cases in which the underlying thesis is clearly false. What seems much more likely is that there are many causal claims for which it is unclear what the backing laws are or even what the backing relation amounts to…In such cases, it may be harmless to say that an explanation in terms of underlying laws must be "there" even if we do not know what it is, but we should also realize that we do not have much purchase on what "underlying" means. In such cases, the underlying thesis may express little more than a commitment to physicalism and to the idea that physical phenomena are law-governed." (Woodward, 2003: 174)

In summary, although the examples and arguments presented in this section do not strictly prove that laws are not necessary for causation, they do suggest that laws are not a necessary part of the meaning or concept of causation and this is sufficient to undermine the assumption that causation is *identical* to (nomological) sufficient production.

I suggest that this conclusion can be supported further once we recognise that interventionism avoids many of the problems identified above, whilst being able to successfully distinguish between genuinely causal and non-causal generalizations and relationships. The reason why interventionism is able to avoid these problems is because, as I explained in Section 4.2.1 above, interventionism replaces the notion of a law with the notion of an invariant generalization and, as will become clear, invariant generalizations simply avoid many of the problems that traditional laws face.

Firstly, unlike laws, remember that it is not necessary that invariant generalizations are exceptionless. For this reason, interventionism is able to avoid the problem, noted by Woodward in his first response to Hempel, that the underlying laws that are thought to be necessary for causation will either have to be strict and exceptionless (raising the problem that it is unlikely that they could contribute to causal understanding) or be qualified with ceteris paribus clauses (raising the problem that those generalizations appear somewhat ad hoc and may not even qualify as laws).

Secondly, interventionism avoids the problem noted by Woodward in his second response to Hempel, which was that it is difficult to accept that our causal understanding of singular causal claims and explanations is acquired in virtue of conveying information about underlying laws, when those laws are likely to be either implicit or epistemically unavailable to an ordinary subject. Interventionism avoids this problem since remember that it states that our causal understanding is grounded in the fact that genuine causal claims and explanations convey information about the outcome of interventionist counterfactuals and this information is *explicit* in those causal claims and explanations.

Lastly, interventionism is able to avoid the problem identified above with the 'meaning thesis', which explained that given that our causal understanding of causal claims and explanations is not acquired *in virtue* of knowledge of laws, it undermines the idea that it is part of the very meaning or concept of causation that underlying laws exist. Firstly, interventionism is able to avoid this problem, since remember that the invariant generalizations associated with singular causal claims and explanations need not be strict and exceptionless and are therefore more likely to be known by an ordinary subject and contribute to causal

understanding. (For example, as in the case of the singular causal claim, 'Smith's smoking caused his cancer', which is supported by the invariant generalization, 'Smoking causes cancer'.) Secondly, it does not undermine interventionism that the associated invariant generalizations are often only implicit in the causal claims and explanations and are therefore unlikely (at least explicitly) to contribute to causal understanding. This is because, interventionism does not understand the meaning or concept of causation or causal explanation in terms of these invariant generalizations, but as I explained above, instead understands the meaning of causation and causal explanation in terms of the outcome of interventionist counterfactuals and this information is explicit in those causal claims and explanations.[30]

In the next section I present some examples and arguments, which suggest that *production*, *determination* and *generation* are not necessary for causation. More specifically, these examples suggest that spatiotemporally continuous processes that produce, generate or determine the occurrence of their effects (which I explained in Chapter 3 are entailed by the SP concept of causation) are not necessary for causation. Once again, the arguments in this section will not strictly *prove* that production, generation and determination are not necessary for causation, but will nonetheless cast doubt on this idea and this is sufficient to undermine the assumption that causation is *identical* to sufficient production.

[30] As a result, interventionism is also able to avoid the worries raised by Anscombe (1981), since although interventionism does claim that all singular causal claims imply the existence of some associated type-level generalization, it is not necessary that those generalizations be exceptionless (and hence standardly 'law-like'), or be explicitly known by an ordinary subject in order for that subject to be able to grasp the causal status of some particular instance of causation.

To begin, note that there are examples in physics, which seem to support the idea that spatiotemporally continuous productive processes, which involve the transfer of some conserved physical quantity, are not necessary for causation. As an illustration, consider Woodward's (2003: 148) example of the inverse square law. Now, this law seems to allow for the possibility of 'action at a distance', in which two physical objects stand in what appears to be a genuine causal relationship, even though there is no spatiotemporally continuous process that connects the two objects. According to the SP concept, these relationships could not possibly count as causal in virtue of the fact that they fail to instantiate a spatiotemporally continuous productive process of some kind.

From an interventionist perspective, by contrast, the relationships described by the inverse square law can count as genuinely causal, "as long as it is true that manipulating the mass or position of the first (second) body will change the gravitational force exerted by the second (first)…" (Ibid: 148) and this can be true even if there is no spatiotemporally continuous process that connects the two bodies.

The notion of action at a distance is not, of course, uncontroversial, but as Woodward notes, although it may be true (putting aside issues having to do with quantum indeterminacy) that given that conserved quantities that are conserved in some interaction are conserved *locally*, and hence given that it will be true that causal interactions that involve the transfer of some conserved quantity will involve a spatiotemporally continuous process, it is not true that *all* causal interactions involve the transfer of some conserved quantity and so it is not true that *all* causal interactions must involve spatiotemporally continuous processes. Rather, Woodward suggests that we think of this fact as an empirical, a posteriori

fact about physical causation, rather than as a necessary condition for causation.[31] Most importantly for the argument in the next chapter, even when we are presented with examples of physical causation, in which it does seem necessary that some spatiotemporally continuous process is involved, interventionism denies that the causal and explanatory status of those relationships is acquired *in virtue* of the fact that they involve spatiotemporally continuous processes, but claims instead that it is acquired solely in virtue of the fact that they meet the interventionist requirements of causation.

In any case, we do not have to look to physics to find intuitive examples of causation, which do not seem to involve spatiotemporally continuous processes. Take, for example, cases of so called causation by omissions, which do not involve any spatiotemporally continuous process, but which nonetheless appear to be genuinely causal. Examples include, 'the lack of oxygen in the chamber caused X to die', 'the Titanic hit the iceberg because there were no binoculars on the lookout deck', 'the inattentiveness of the driver caused the car crash', and so on. Again, this is not to say that the issue of causation by omissions is uncontroversial (I address a problem concerning causation by omissions in Section 4.4.2 below), but from an interventionist perspective, each of these intuitive claims and explanations can count as genuinely causal, since there are interventions on each of the purported causes which are associated with changes to the effects. This does appear to be a virtue of interventionism that is crucially lacking on the SP concept of causation.

[31] I made a similar point in Chapter 3, in which I explained that although causal closure does imply that every physical effect has a sufficient physical cause (that presumably produces this effect via a spatiotemporally continuous process), this only generates the a priori exclusion problem if one assumes that this kind of sufficient production is identical to causation.

In summary, what these general examples suggest is that spatiotemporally continuous processes, which produce, generate or determine the occurrence of their effects, are not necessary for causation and that interventionism is simply able to avoid the problems highlighted for the SP concept. Once again, it is important to emphasise that just as with the first set of examples, although these examples do not strictly *prove* that spatiotemporally continuous productive processes are not necessary for causation, they do at least cast doubt on this idea and this is sufficient for the purposes of my argument, since it nonetheless undermines the assumption that causation is *identical* to sufficient production.

4.3.2 Is Sufficient Production Sufficient for Causation?

In this next section I present some examples which suggest that sufficient production is not *sufficient* for causation, or more accurately, that providing sufficient conditions for the occurrence of some effect is not sufficient for causal *explanation*. In order to illustrate this, I provide some examples of explanations, which seem to meet the requirements of the SP concept of causal explanation, but which either seem to lack certain features that we expect from successful causal explanation, or which seem to provide deficient causal explanations, in comparison to those explanations that meet the requirements of interventionism. Once again, although these examples will not strictly *prove* that providing sufficient conditions for the occurrence of some effect is not sufficient for causal explanation, they do cast doubt on this idea and this is sufficient for the purposes of my argument, since it nonetheless undermines the explanatory counterpart of the assumption of SP.

As a first illustration, recall Yablo's example introduced above of the pigeon trained to peck specifically at red objects. In this example we saw that although the fact that the object is scarlet is *sufficient* to produce, or determine the pecking behaviour, it fails to capture what is causally relevant about the object that causes the pigeon to peck, namely the fact that it is red. By being overly specific, it both omits vital information (this being that the pigeon would fail to peck in any case in which the object is not red) and includes irrelevant and potentially misleading information (this being that the pigeon would fail to peck in any case in which the object is not scarlet). This example therefore suggests that providing sufficient conditions for the occurrence of some effect is not *sufficient* for causal explanation, since the explanation citing the property of scarlet clearly meets the SP conditions for causal explanation and yet appears to be explanatorily deficient in an important respect.

This conclusion is supported further when we see that from an interventionist perspective, although both explanations do technically qualify as causal, the explanation citing the property of scarlet comes out as explanatorily deficient in comparison to the explanation citing redness (in line with our intuition), given that it captures the wrong contrastive focus and given that it is therefore less useful for the purposes of control and manipulation. Note that from the point of view of the SP concept of causation, this potentially useful distinction *amongst* causal explanations is simply lost.

It is also worth pointing out that on the SP concept of causal explanation, this potential practical benefit of causal explanation is completely lost. This is because, as Yablo's example illustrates, it is simply not true that by identifying sufficient conditions for the occurrence of some effect we thereby acquire greater

practical information about how we might go about manipulating or controlling the effect. In fact, what Yablo's example illustrates, is that often, by identifying *more* precise sufficient conditions for the occurrence of some effect, we actually acquire information that is *less* useful for the purposes of control and manipulation. Although this does not prove that providing sufficient conditions for the occurrence of some effect is not sufficient for causal explanation, it does suggest that something (namely the potential practical benefit of causal explanation) is missing according to this assumption about causal explanation and this goes some way to undermine the explanatory counterpart of the assumption of SP.

One final point that I suggest also undermines the explanatory counterpart of the assumption of SP can be seen by drawing attention to the fairly unintuitive consequences of this assumption. Remember that according to the explanatory counterpart of the assumption of SP, genuine causal explanations cite 'full and sufficient conditions' for the occurrence of their effects. As a consequence, it is simply not possible for some effect to have more than one causal explanation without running into the problem of overdetermination. It was for this reason that Kim claimed that when presented with a case in which some explanation did not appear to be sufficient for its effect, we should either think of that explanation as a 'part-cause' or part-explanation of the effect that somehow adds together with other part-causes to provide a sufficient explanation of the effect, or we should think of the various explanations as somehow competing with one another.[32] Recall Kim's example of the highway crash introduced in Chapter 3:

[32] *Or* we can assume that the explanations are related via some dependency relation, such as supervenience. As I explained in Chapter 3, this had serious consequences for the prospects of psychological explanation, since Kim argued from the fact that physical explanations are, by

> "Thus a car accident is explained by a highway designer as having been caused by the incorrect camber of the highway curve, and by a police officer as caused by the inattentive driving of an inexperienced driver. But in a case like this we naturally think of the offered causes as partial causes; they together help make up a full and sufficient cause of the accident." (Kim, 1998a: 66)

Now, according to Kim, in a case like this, we will have failed to explain the car crash until we have identified *all* of the sufficient conditions that together 'add up' to provide what we may call 'the' cause of the crash, which is alone sufficient for the occurrence of the effect. However, this does not seem to be in line with the way that we ordinarily think about causal explanation. For example, it seems perfectly natural to appeal to the inattentiveness of the driver as 'the' cause of the car crash, even though it is presumably not true that this fact alone was sufficient to produce the effect. Moreover, it seems equally unintuitive and unnatural to think that if the various causes did not somehow 'add up' to provide sufficient conditions for the occurrence of the effect, they would thereby compete with each other for explanatory status.

Interventionism is able to avoid these fairly unintuitive consequences, since from an interventionist perspective, each of the explanations noted in the example can count as genuinely causal given that they convey information about the outcome of interventionist counterfactuals. From an interventionist

---

definition, sufficient to explain their effects and from the fact that mental properties supervene on physical properties and hence cannot overdetermine their effects by providing sufficient explanations of their own, that they must inherit all of their explanatory power from their subvenient physical realizers (c.f. Kim's 'Causal Inheritance Principle' discussed in Chapter 3).

perspective, it simply does not make sense to think that causal explanations should either 'add up' to provide a sufficient explanation of some effect, or compete for causal and explanatory status. Once again, although this does not of course *prove* that providing sufficient conditions for the occurrence of some effect is not sufficient for causal explanation, it does at least suggest that there is something wrong with this idea and this is sufficient to undermine the explanatory counterpart of the assumption of SP.

In this section, I have presented some examples and arguments which undermine the assumption that causation is identical to sufficient production and which undermine the explanatory counterpart of the assumption of SP, which assumes that causal *explanation* is simply a matter of providing such sufficient conditions for the occurrence of some effect. Given that in Chapter 3, I argued that Kim crucially depends on this assumption and its explanatory counterpart to generate the exclusion problem, this discussion should have therefore demonstrated that the non-reductive physicalist need not accept Kim's a priori exclusion problem.

4.4 Interventionism versus the SP Concept of Causation: Problems for Interventionism

In response, however, one could argue that despite the problems that the SP concept faces, interventionism fails to provide a viable alternative to this theory and hence fails to undermine the assumption of SP, since it faces serious problems of its own, which the SP concept seems to avoid.

In this final section, I argue that interventionism is able to avoid many of the standard problems that counterfactual theories of causation face, such as

problems that arise from cases of overdetermination, non-paradigmatic causation and causation by omissions, which are often thought of as cases that a production based concept of causation, such as the SP concept, can easily deal with. I argue that as well as being able to overcome these problems, interventionism is actually able to deal with many of these problems in a more satisfying way than the SP concept. I therefore conclude that interventionism *does*, after all, provide a coherent and viable alternative theory of causation, which does undermine the assumption of SP and hence does demonstrate that the non-reductive physicalist need not accept Kim's a priori exclusion problem. (It will be equally important to demonstrate that interventionism is able to deal with these standard objections, if interventionism is to provide a coherent account of mental causation and satisfactory solution to the exclusion problem.)

4.4.1 Overdetermination

I demonstrated above that interventionism can be applied to both type and token level causal claims. However, there are a significant class of token-level claims, for which the associated interventionist counterfactuals deliver highly un-intuitive causal judgements. These are cases involving overdetermination. In fact, the problem of accommodating cases of overdetermination within counterfactual theories of causation, such as interventionism, is often considered to be one of the most significant problems that these theories face. Moreover, the fact that production based theories of causation, such as the SP concept, seem to be able to easily deal with these cases has led some to argue that we must retain the SP concept of causation, despite its apparent shortcomings. In fact, it has led some, such as Hall (2004), to argue for a two-concept theory of causation, incorporating

both counterfactual dependence and production. However, I demonstrate that unlike other counterfactual theories of causation, interventionism is able to deal with cases of overdetermination, such that it is not necessary to retain the SP concept of causation, or posit a two-concept theory of causation.

In order to see this, let us first recall the interventionist account of token, or 'actual' causation and see how cases of overdetermination apparently generate a problem for this account. Remember that in order for it to be true that X is a token, or actual cause of Y, there must exist some intervention on X that changes the *actual* value of X (e.g. from x to x') that changes the *actual* value of Y (e.g. from y to y'). Woodward provides the following, more precise definition of token or actual causation (AC):

> "**(AC)** (AC1) The actual value of $X = x$ and the actual value of $Y = y$.
>
> (AC2) There is at least one route R from X to Y for which an intervention on X will change the value of Y, given that other direct causes Z_i of Y that are not on this route have been fixed at their actual values. (It is assumed that all direct causes of Y that are not on any route from X to Y remain at their actual values under the intervention on X.) Then $X = x$ is an actual cause of $Y = y$ if and only if both conditions (AC1) and (AC2) are satisfied." (Woodward, 2003: 77)

In order to see how overdetermination causes trouble for this definition of token causation, consider the following paradigmatic example of overdetermination cited by Woodward: two campers each throw a lighted cigarette into a forest (represented by the variables c_1 and c_2) and each cigarette

on its own is sufficient to bring about the occurrence of the forest fire (represented by variable e). As an illustration of how AC applies (or more accurately, misapplies) to this case, consider the following passage from Woodward:

> "…let *A*=1 or 0 according to whether c_1 occurs, *B*=1 or 0 according to whether c_2 occurs, and *C*=1 or 0 according to whether *e* occurs…Fixing *A* at its actual value = 1 in accord with (AC), we see that changing the value of *B* from its actual value (*B* = 1) does not change the value of *C*. So, according to AC, c_2 (*B* = 1) is not an actual cause of *e* (*C* = 1). By parity of reasoning, c_1 is also not a cause of *e*." (Ibid: 82)

In other words, given that in this case there is no intervention on the actual values of c_1 or c_2 that changes the value of e, while holding fixed c_1 or c_2 at their actual values (in accordance with AC), it turns out, according to interventionism, that neither c_1 nor c_2 qualify as actual causes of the forest fire, despite the strong intuition that at least one of the events caused the fire. Given this strong intuition, it is argued that cases of overdetermination actually prove that counterfactual dependence is not necessary for causation, since these relationships strike us as genuinely causal even though it appears that there is no counterfactual dependence between cause and effect. Moreover, cases of overdetermination clearly do not present this problem for the SP concept, since that account provides a straightforward explanation of how c_1 or c_2 (or both) can qualify as causes of the fire: both c_1 and c_2 are each sufficient to produce the effect and hence qualify as overdetermining causes of the forest fire.

How then can we resolve the fact that AC seems to deliver the problematic conclusion that counterfactual dependence is not actually necessary for causation and that we have to turn to a production based theory of causation, such as the SP concept to deal with these cases?

To resolve this issue, we can appeal to the following solution offered by Woodward, which draws on the idea of what Christopher Hitchcock has called the 'redundancy range' of variables within a system. The notion of a redundancy range is explained in the following passage:

> "Consider a particular directed path *P* from *X* to *Y* and those variables $V_1...V_n$ that are not on *P*. Consider next a set of values $v_1...v_n$, one for each of the variables V_i. The values $v_1...v_n$ are in what Hitchcock calls the *redundancy range* for the variables V_i with respect to the path *P* if, given the actual value of *X*, there is no intervention that in setting the values of V_i to $v_1...v_n$, will change the (actual) value of *Y*. The actual values of the variables V_i are, of course, in the redundancy range with respect to *P* but nonactual values of the variables V_i will also be in the redundancy range if, given the actual value of *X*, we can set the variables V_i to those values without disturbing the actual value of *Y*." (Ibid: 83)

As the passage above explains, certain values will be within the redundancy range of variables iff given the actual values of X and Y, setting the variable to those values does not change the value of Y.[33] Woodward suggests that this

[33] This allows one to avoid the problem that by modifying the values of other direct causes in this way, we could disrupt, or confound the causal relationship between X and Y. This is because, as

provides a potential solution to the problem of overdetermination, since he suggests, contrary to AC, that it is *not* always appropriate to hold fixed the other direct causes of Y that are not on the path from X to Y at their *actual* values, but that we may fix the other direct causes of Y to *non-actual* values if those values are within the redundancy range of values for those variables. Woodward suggests modifying AC to incorporate the notion of a redundancy range in the following way:

> "**(AC*)**: (AC*1) The actual value of $X = x$ and the actual value of $Y = y$. (AC*2) For each directed path P from X to Y, fix by interventions all direct causes Z_i of Y that do not lie along P at some combination of values within their redundancy range. Then determine whether, for each path from X to Y and for each possible combination of values for the direct causes Z_i of Y that are not on this route and that are in the redundancy range of Z_i, whether there is an intervention on X that will change the value of Y. (AC*2) is satisfied if the answer to this question is "yes" for at least one route and possible combination of values within the redundancy range of the Z_i. $X = x$ will be an actual cause of $Y = y$ if and only if (AC*1) and (AC*2) are satisfied." (Ibid: 84)

How then does this modification help in the case of symmetrical overdetermination above? Note that in the case of the forest fire, the value B = 0 is within the redundancy range for the variable B because "…given the actual

---

Woodward points out, by being within the redundancy range, those values do not, by definition, influence the relationship between X and Y.

value of A, A = 1, the value of C, C = 1 would be unchanged if B = 0." (Ibid: 83) Then, fixing B to its non-actual value, B = 0, changing the value of A from its actual value, A= 1 to A= 0, *does* change the value of C; hence according to AC*, c_1 qualifies as a cause of e (the same goes for c_2).

By introducing the notion of a redundancy range of values it is therefore possible for both c_1 and c_2 to qualify as causes of e, avoiding the problem that interventionism seems to deliver the counter-intuitive judgement that neither event caused e to occur and hence avoiding the problem that counterfactual dependence does not appear to be necessary for causation. By fixing the values of the variables within the redundancy range to non-actual values, the counterfactual dependencies (and hence the causal relationships) between the variables becomes apparent. Moreover, by modifying AC in this way, we are not forced to appeal to a production based concept of causation, such as the SP concept, in order to deal with these cases.

4.4.2 Non-Paradigmatic Causation and Causation by Omissions (and insensitivity as a solution to these problems)

In this section I discuss another set of cases, which apparently cause trouble for counterfactual theories of causation. These are cases of non-paradigmatic causation and causation by omissions. As I will demonstrate, these cases are thought to be problematic for counterfactual theories, since they appear to show that it is possible to have counterfactual dependence without causation. In other words, they appear to illustrate that counterfactual dependence is not *sufficient* for causation. Moreover, it looks as though production based concepts of causation, such as the SP concept, are simply able to avoid this problem.

In this section I argue that by appealing to the interventionist concept of *insensitivity*[34], not only can interventionism overcome the problems associated with these cases[35], but it is actually able to deal with these problems in a more satisfying way than the SP concept. This supports the conclusion that interventionism *does*, after all, provide a coherent and viable alternative theory of causation, which does undermine the assumption of SP and hence does demonstrate that the non-reductive physicalist need not accept Kim's a priori exclusion problem. (I begin by outlining and examining the interventionist notion of insensitivity and will then demonstrate how this feature can be used to avoid the problems associated with these cases.)

What exactly is insensitivity? In order to explain this concept, it is useful to compare it to the central interventionist notion of invariance. As the discussion above should have made clear, we can think of invariance as a *necessary* feature that a generalization or relationship must possess if it is to qualify as genuinely causal. I suggest that we can then think of insensitivity as a *further* condition, which considers whether that relationship (that is at least minimally invariant and hence causal) would *continue to hold* (or alternatively, whether the counterfactuals associated with that claim would continue to hold) over a range of changes and varying background conditions. If that relationship does continue to hold under these changes, then we may regard that causal relationship as insensitive. If it does not, then that causal relationship will be considered as sensitive. Woodward defines the notion of insensitivity as follows:

[34] Woodward (2006) discusses the notion of insensitivity in detail.
[35] This feature also allows interventionism to avoid the problem of 'double-prevention', noted by Hall (2004), which cannot be discussed here. See Hall (2004) for further details.

> "Broadly speaking, a causal claim is sensitive if it holds in the actual circumstances but would not continue to hold in circumstances that depart in various ways from the actual circumstances. A causal claim is insensitive to the extent to which it would continue to hold under various sorts of changes in the actual circumstances. The sensitivity of counterfactuals is understood similarly." (Woodward, 2006: 2)

The idea of insensitivity can be elucidated further with the following example of Woodward's:

> "Suzy stands in front of a fragile glass bottle with a large rock in her hand. No other possible causes of the bottle's breaking — no backup or preemptive throwers, no earthquakes and so on — are waiting in the wings. Suzy throws; the rock strikes the bottle squarely, and it shatters. The impact of the rock caused the bottle to shatter." (Ibid: 1)

Now, according to most theories of causation, it was Suzy's throwing of the rock that caused the bottle to shatter. Interventionism supports this conclusion, since there is an intervention on Suzy's throw that changes whether the bottle shatters (i.e. the relationship between the throwing of the rock and the bottle shattering is minimally invariant). This conclusion is supported by the truth of the following two counterfactuals:

(1.1) If Suzy throws the rock, the bottle will shatter.

(1.2) If Suzy does not throw the rock, the bottle will not shatter.

Now, as Woodward explains, in order to assess the insensitivity of the causal claim relating Suzy's throwing of the rock to the bottle shattering we should consider the insensitivity of the associated counterfactuals (1.1) and (1.2). As Woodward goes on to explain, in order to do this, we should consider whether, for example, counterfactual (1.1) "would continue to hold under changes that do not depart too much from the actual state of affairs or that do not seem too far-fetched or that are not judged to be unimportant or irrelevant for subject-matter-specific reasons" (Ibid: 11) (I explore these ideas in detail below). This idea can be expressed in the form of the following counterfactual:

> (1.1.2) "If the rock thrown by Suzy were to strike the bottle in circumstances B_i different from the actual circumstances, the bottle would (still) shatter." (Ibid: 5)

Since it appears that (1.1.2) is true, (for example, if Suzy were to throw the rock at a slightly later time, or with a slightly different degree of force, the bottle would still shatter), we should consider counterfactual (1.1) to be fairly insensitive and consequently also consider the causal claim relating Suzy's throwing of the rock to the bottle shattering as insensitive. If, on the other hand (1.1.2) turned out false (if, for example, this relationship failed to hold across a range of such changes), we should instead consider counterfactual (1.1) *and* the causal claim to be fairly sensitive.

I noted above that a counterfactual, or causal claim will be judged to be insensitive if it holds over a range of changes to background conditions that do

not depart much from actuality or that seem important or relevant. Can we be more precise about the kinds of changes that are relevant for assessing insensitivity?

As Woodward notes, certain changes will seem irrelevant for assessing the insensitivity of a counterfactual or causal claim. These include, for example, changes to the colour of Suzy's blouse, or the sneezing of a man in Chicago. (One explanation of why these kinds of changes are not considered as relevant is that they fail to change the relevant features of the counterfactual, for example, the throwing of the rock and the smashing of the bottle.[36]) In other kinds of cases, (to be discussed below), certain changes will seem irrelevant because of the particular context of the counterfactual or causal claim.

Within those changes that we *do* judge to be relevant for assessing insensitivity, according to Woodward, we can be more precise about the specific nature of the changes by appealing to a similarity metric, along the lines of David Lewis' notion of the closeness of possible worlds. Essentially, the possible worlds account would consider, as relevant for assessing the insensitivity of some counterfactual or causal claim, changes which are as close to actuality as possible. For example, when considering whether counterfactual (1.1.2) is true, it states that we should consider situations which do not depart much from the actual situation, such as a situation in which Suzy throws the rock at a slightly later time or with a different degree of force, rather than, for example, considering a situation in which Suzy throws the rock, which happens to be fitted with a navigation device which ensures that the rock reaches its target.

[36] This is closely related to the idea of a 'testing intervention', described above.

Now, one important aspect of insensitivity that will be especially relevant to the discussion below is that insensitivity provides an *explanation* of our causal judgements and in particular, can explain how certain claims can qualify as causal according to the interventionist criteria for causation, while nonetheless striking us as non-causal. For example, take the case of Suzy: according to Woodward, it is the truth of counterfactuals (1.1) and (1.2), *as well as* the insensitivity of these counterfactuals that informs our judgement that Suzy's throw is causally relevant to the bottle shattering. By way of contrast, consider Woodward's variant of the Suzy case, which is extremely sensitive to slight changes to the actual situation: Suzy scratches her nose and the bottle shatters. As Woodward explains, although there may be counterfactual situations in which it is true that the shattering is caused by the scratching of Suzy's nose, if for example, Billy had promised to shatter the bottle if Suzy scratches her nose, the reason why we would not ordinarily judge that Suzy's action is the cause of the shattering is that it is extremely sensitive to slight changes to the actual circumstances. For example, Billy may renege on his promise or fail to be present. For Woodward, it is the relative sensitivity of the second claim that informs our judgement that although Suzy's action may technically cause the shattering of the bottle (since there is some intervention on whether Suzy scratches her nose that changes whether the bottle shatters), that causal claim nonetheless strikes us as un-paradigmatically causal, since it is extremely sensitive to small changes to background conditions.

Why does insensitivity affect our causal judgements in this way? Note that just as with the notion of invariance, by being stable or invariant over a wide range of changes (in this case, it is specifically stability under changes to

background conditions and circumstances, rather than just stability under interventions to the variables under consideration), causal claims and explanations that are insensitive will be able to answer a wider range of w-questions, since they tell us what *would* happen to the effect under these wide range of changes. Moreover, those causal claims and explanations will be more potentially useful for the purposes of control and manipulation, since those relationships will continue to hold and hence continue to provide a means of control over these wide range of changes. As Woodward (Ibid: 7) explains, the idea of insensitivity thus captures the intuitive idea that causal claims and explanations should possess a certain degree of 'generalizability' and 'context independence'.

Before moving on to discuss how the notion of insensitivity helps interventionism to overcome problems concerning cases of non-paradigmatic causation and causation by omissions, it is important to address the question of whether the conditions for determining the relevance of changes to background conditions for the assessment of insensitivity are problematically subjective. It is important to address this question if the notion of insensitivity is to be used to avoid the problems noted above. Moreover, as we will see, the issues that I discuss concerning subjectivity are of crucial importance to subsequent discussions in this chapter and in the next.

Now, a problem concerning subjectivity seems to arise since, as I explained above, the relevance of certain changes for assessing insensitivity can depend on context, pragmatics, the expectations of the subject and so on. For example, as Woodward notes, many generalizations in economics will be extremely sensitive to changes to the neurological processes of economic agents,

but it would seem strange to consider these changes as relevant for assessing the insensitivity of economic causal generalizations. Or, alternatively, it is possible for some changes to strike one individual as extremely close to actuality, but for the same changes to strike another individual as fairly far-fetched, given, for example, a difference in social expectations between the subjects. In other words, a potential problem arises since one could argue that by employing a similarity metric along the lines of Lewis' notion of closeness of possible worlds, this account is simply open to the same problems that Lewis faces, which concern the apparent vagueness of the notions of closeness and similarity (Fine 1975).

However, as Woodward explains in the following passage, the concept of insensitivity is not *problematically* subjective:

> "First, as emphasized above, one of my primary interests in the role of sensitivity is in using this notion to describe actual practices of causal judgment. It is an empirical question to what extent people's judgments of sensitivity depend on the factors I have described; to the extent that they do, it is not an objection to the account that some of these features strike us as "subjective." Second, it is also of course an empirical question to what extent there is intersubjective agreement in people's judgments of sensitivity; it may be that we are largely able to agree on such judgments despite their highly contextual and highly multifaceted character. Finally, one obvious response to worries about subjectivity and context dependence is to relativize judgments of sensitivity to particular sets of changes in backgrounds. Even if you and I disagree about whether such and such a departure *B** from actuality is large or far-fetched, it may

> be an "objective" matter (or at least a matter about which we may expect far more agreement) whether some counterfactual or causal claim would hold under *B**. Thus, even if we disagree about whether the introduction of a solid steel barrier between Suzy and the bottle represents a large departure from the actual state of affairs, we can presumably agree that if such a barrier were introduced, it would no longer be true that if Suzy were to throw, the bottle would shatter." (Ibid: 15)

There are several points captured in this passage that are relevant to our discussion. Firstly, what this passage suggests is that judgements of insensitivity may not actually differ to as great an extent as one might initially think, but to the extent that they do differ, this is not fatal to the notion of insensitivity, since this notion is essentially a practical one that concerns the nature of *our* causal judgements and as such we should expect a certain degree of subjectivity to enter into these judgements. Secondly, this passage suggests that to the extent that there are differences in judgements about which changes are more or less relevant for assessing insensitivity, we can assume that whether the counterfactuals actually hold under those changes is not a subjective matter. Although this point is not captured in the passage above, most importantly, I suggest that this kind of subjectivity does not introduce a *problematic* kind of subjectivity into interventionism, since as I explained above (and as I explain in further detail in Section 4.5 below), whether or not some relationship or generalization qualifies as causal depends *solely* on whether that relationship or generalization is invariant under interventions and I suggest below that we have good reason to think that *this* question is entirely objective.

With the notion of insensitivity outlined, I will now demonstrate how this notion helps interventionism to overcome two standard problems that less developed counterfactual theories of causation face. These are the problems that arise from cases of non-paradigmatic causation and causation by omissions.

Firstly, consider the problem that it is possible, according to interventionism, for some relationship to qualify as genuinely causal in virtue of the fact that there is counterfactual dependence of the right kind, namely invariance under interventions, even though those relationships strike us as un-paradigmatically causal (or even non-causal). One might reach this conclusion, for example, in the case of the claim relating the scratching of Suzy's nose to the bottle shattering. One could argue that these kinds of cases actually illustrate that counterfactual dependence is not sufficient for causation, since they appear to illustrate that it is possible for there to exist counterfactual dependence of the right kind without causation.

The notion of insensitivity helps interventionism to avoid this problem, since it explains how it is possible for some relationship to qualify as genuinely causal, even though it may be *judged* as un-paradigmatically causal, or even non-causal, given that it is fairly sensitive.[37] When faced with examples of this kind there is therefore no reason to conclude that counterfactual dependence is not sufficient for causation, since our causal judgements that these cases are somewhat problematic and non-paradigmatic are explained by the relative sensitivity of those claims.[38]

---

[37] Note that this also explains how certain background conditions (such as the presence of oxygen in the environment) can strictly qualify as causal according to interventionism, whilst striking us as somehow unparadigmatically, or problematically causal.

[38] There are of course those who will simply dig their heels in and argue that these cases are so un-paradigmatic that they should not qualify as causal at all. However, the point of this

A similar problem arises for interventionism as a result of cases of causation by omissions.[39] Remember that according to interventionism, omissions can qualify as genuine causes, so long as there is counterfactual dependence of the right kind between the variables. Moreover, I argued above that the fact that interventionism can accommodate cases of negative causation, while the SP concept cannot, goes some way to undermine the assumption of SP. However, a problem seems to arise since it appears that counterfactual theories of causation, including interventionism, deliver the result that most, if not all negative events and states can qualify as genuine causes, which in many cases seems highly un-intuitive. In order to illustrate how the notion of insensitivity helps to overcome this problem, consider the following example of Woodward's:

> "First, consider
>
> (5.1) My writing of this very essay was caused by my not being hit by a large meteor
>
> and the associated counterfactuals,
>
> (5.2) If I were not struck by a large meteor, I would have written this very essay
>
> and
>
> (5.3) If I were struck by a large meteor, I would not have written this very essay." (Ibid: 24)

---

discussion is not to convince those who are deeply sceptical about counterfactual theories of causation that they are wrong, but rather to illustrate that this problem can be dealt with within an interventionist framework, by appealing to the notion of insensitivity.

[39] The problems associated with causation by omissions (in particular for counterfactual accounts of causation) are captured concisely by Beebee (2004). Although I cannot discuss Beebee's arguments here, I believe that the notion of insensitivity addresses the problems that Beebee highlights in her paper.

Now, according to interventionism, there is a sense in which (5.1) is true, since it is true that intervening on whether Woodward is struck by a meteor is a way of intervening on the writing of the essay. However, as Woodward points out, it is also true that counterfactual (5.2) is highly sensitive[40], in the sense that there are a range of relatively small variations to the situation, which would result in Woodward not writing the essay. For example, he may fail to have had the conversation with his colleague that gave him the idea for the essay, or he may simply have not had the time to write the essay. For Woodward, it is the sensitivity of counterfactual (5.2) that explains why, although (5.1) may technically qualify as causal according to interventionism, it nonetheless strikes us as an un-paradigmatic case of causation. Just as with the case above, we therefore need not conclude from such examples that counterfactual dependence is not sufficient for causation, since the notion of insensitivity explains how these kinds of cases can qualify as strictly causal, but *appear* un-paradigmatically causal or non-causal nonetheless.

Moreover, this problem does not undermine the argument that I made above, which was that it is a virtue of interventionism that it accommodates cases of negative causation, while the SP concept does not. This is because while the meteor case seems to strike us as un-paradigmatically causal, there are nonetheless cases of causation by omissions, which do strike us as genuinely causal. Consider the following examples of Woodward's: 'The absence (of

[40] According to Woodward, the insensitivity of a positive counterfactual (i.e. a counterfactual relating the *occurrence* of the supposed cause to the effect) carries more weight than the insensitivity of a negative counterfactual (i.e. a counterfactual relating the *absence* of the supposed cause to the effect). (Also note that in the meteor case above, the positive counterfactual actually concerns the *non-o*ccurrence of the meteor strike.)

access) to oxygen caused N's death', and 'Many German civilians were caused to die from starvation (that is, from absence of food) by the British naval blockade of 1919'. Now, it does seem that we would be more willing to attribute causality to these negative claims than we would for (5.1) and for Woodward, this is simply because these claims are *less* sensitive than (5.1). My suggestion is that it is a virtue of interventionism that these intuitive claims can qualify as genuinely causal, while the same cannot be said for the SP concept.[41]

The benefit of introducing the notion of insensitivity into interventionism is clear: it remains possible to regard certain cases of causation by omissions as genuinely causal, while explaining the fact that there is something nonetheless un-paradigmatic about many cases of negative causation. By contrast, note that according to the SP concept, *none* of these claims (either intuitive or unintuitive) can count as causal.

In summary, in this section I hope to have demonstrated that interventionism is able to avoid many of the standard objections that counterfactual theories of causation face, such as the problems that arise from cases of overdetermination, non-paradigmatic causation and causation by omissions. Given that interventionism is able to avoid these standard objections and given that I have demonstrated that in many cases, interventionism is able to

[41] As Godfrey-Smith (2007: 13) points out, the issue of negative causation also arises in the context of responsibility. Consider Godfrey-Smith's example: suppose that you walk past a child who has fallen into a pond, who then drowns. As Godfrey-Smith points out, although we *might* hold you responsible for the child's death even though you didn't actually cause it (we can imagine, for example, that someone sympathetic to the SP concept of causation might argue along these lines on the grounds that there is no spatiotemporally continuous physical process that connects your 'inaction' with the child's death), it is less problematic (and, I suggest, more natural) to hold you responsible (morally and perhaps legally) if we treat your inaction as a *cause*. I suggest that these kinds of examples provide further support for the idea that some cases of negative causation do strike us as genuinely causal and that it is a virtue of interventionism that these intuitive claims can qualify as genuinely causal, while the same cannot be said for the SP concept.

deal with these problems in a more satisfying way than the SP concept, it is possible to conclude that interventionism *does*, after all, provide a viable alternative theory of causation to the SP concept, which does undermine the assumption of SP and hence does demonstrate that the non-reductive physicalist need not accept Kim's a priori exclusion problem. *Moreover*, given that interventionism is able to deal with these standard objections, it is possible to conclude that interventionism can be used to provide a coherent account of mental causation and satisfactory solution to the exclusion problem.

4.5 Remaining Problems

There are, however, some remaining problems that I will discuss in this final section, concerning the potentially *anthropocentric* (Section 4.5.1), *anti-realist* (Section 4.5.2) and *circular* (Section 4.5.3) nature of interventionism. Although these problems do not directly influence the arguments that I made above against the assumption of SP, it is also important to address these general problems if the interventionist account of mental causation that I outline in the next chapter is to be considered as providing a coherent theory of mental causation and satisfactory solution to Kim's exclusion problem.

4.5.1 The Problem of Anthropocentricism

The definition of interventionism outlined in this chapter appeals to the notions of control and manipulation and to the notion of an intervention to characterise causation. One could therefore argue that interventionism is problematically anthropocentric in the sense that whether X causes Y seems to

depend on whether we, as humans, find certain relationships useful for the purposes of control and manipulation and on whether some agent actually performs an intervention.

However, while interventionism does place a great deal of importance on the practical focus of causation and does appeal to human centred concepts, such as control and manipulation, interventionism is not *problematically* anthropocentric. In order to see this, remember firstly that according to interventionism, in order for X to cause Y, it is not necessary that some agent actually *does* perform, or *ever could* perform an intervention on X. For example, remember that the intervention on X need not actually be carried out, but may instead take the form of a hypothetical intervention that considers what *would* happen to Y if we *were* to intervene on X. Furthermore, for the reasons that I outlined above, interventionism *only* requires that these interventions be logically, conceptually and metaphysically possible, rather than being practically, nomically or even physically possible. Thus interventionism is not straightforwardly anthropocentric in the sense that whether X causes Y depends on the possibility that some agent actually does, or ever could perform an intervention on X.[42]

Secondly, in order for X to cause Y, it is not necessary that the intervention on X has an *actual* practical benefit, or 'payoff'. Once again, this is

[42] Unlike, for example, the theory of Menzies and Price (1993), which does make essential reference to the notion of human agency to define causation. Although Menzies and Price do address the issue of anthropocentricism, their solution involves developing fairly complex arguments that appeal to the notion of intrinsic similarity between cases in which some intervention is humanly possible and those cases in which it is not. (Roughly, the idea is that the latter kinds of cases can be considered as causal even though there is no humanly possible intervention associated with those cases, in virtue of the fact that they share intrinsic features with those cases in which some intervention is humanly possible.) By contrast, Woodward's interventionism avoids the problem of anthropocentricism all together, since his notion of an intervention does not make essential reference to the notion of human agency.

because according to interventionism, X causes Y if there is some intervention on X that changes Y, even if that intervention is merely hypothetical, or is nomically, physically or practically impossible. I.e. even if the intervention *is* never, and *could* never be carried out and have an actual practical benefit. Rather, for Woodward, the *potential* practical benefit of interventionist causation actually explains *why* we, as humans, have an interest in causation and in discovering genuinely causal relationships over merely correlative ones. For example, as Woodward (2003: 28-33) explains, if the difference between genuinely causal relationships and merely correlative ones is that the former provide a potential means of control and manipulation over our environment, while the latter do not, then it is no wonder that we place such a great importance on understanding causation and on discovering genuinely causal relationships over merely correlative ones.[43]

4.5.2 The Problem of Realism

There is, however, a remaining concern regarding the potentially anti-realist nature of interventionist causation. Before addressing the issue of realism and interventionism, let us first consider why the issue of realism is so important. Putting aside, for the moment, Kim's general worries with the metaphysical credentials of counterfactual accounts of causation that I addressed briefly in Chapter 3 and will address below and again in Chapter 5[44], there is a strong intuition that our concept of causation and our concept of mental causation

[43] This issue of how interventionism provides an explanation of our *motivation* for understanding causation is extremely interesting, but cannot be discussed further here. See Woodward, 2003, especially Chapters 2 and 3 for further details.

[44] These worries concerned the question of whether counterfactual dependence can sustain the kinds of causal relationships that are involved in human agency, or whether productive causation is required to account for these causal relationships.

should not be subjective and anti-realist in the sense that whether X causes Y depends on facts about us. For example, it should not depend on whether we find certain relationships useful for the purposes of control and manipulation, or on whether the counterfactuals associated with the claims strike us as insensitive. In short, there is a strong and reasonable intuition that we want realism about causation and realism about mental causation in the sense that whether X causes Y *does not* depend on facts about us, but rather, that causation exists objectively 'out there' in reality independently of us.[45]

Moreover, regardless of whether one is sympathetic to the points that I made in response to Kim in Chapter 3 (and to the points that I go on to make in Chapter 5) about counterfactual dependence being sufficient to sustain the kinds of relationships that are involved in mental causation, if it turns out that interventionism is straightforwardly anti-realist, Kim would be justified in claiming that interventionism could not provide a satisfactory account of mental causation and solution to the exclusion problem.

One worry is that the notions of contrastive focus and insensitivity, which are somewhat subjective, introduce a problematic kind of subjectivity into interventionism and generate an anti-realist conception of causation. However, I have argued at length that despite having somewhat subjective features, interventionism is not *problematically* subjective, or anti-realist.

For example, I explained that although the notion of insensitivity is somewhat subjective, in the sense that our assessments of insensitivity can be

[45] This is of course not to endorse an empirical analysis of causation, which takes as its primary focus the discovery of what causation is in reality, as opposed to a conceptual analysis of causation, which seeks to understand the concept of causation as it is used in ordinary language. As I explained above, (see footnote 28), interventionism can be broadly understood as providing a conceptual analysis of causation.

influenced by context, pragmatics, the expectations of the subject and so on, it does not follow that interventionism is thereby problematically subjective, or anti-realist. This is because, considerations of insensitivity *do not* determine whether X causes Y, but merely explain our causal *judgements* and as such, we should actually expect a certain degree of subjectivity to feature in these considerations. As I explained above, whether X causes Y depends *solely* on whether there is counterfactual dependence of the right kind between the variables, i.e. invariance under interventions and whether such counterfactual dependence exists is independent of whether we consider those counterfactuals to be insensitive.

Similarly for the notion of contrastive focus, although I have suggested that this notion is somewhat subjective in the sense that whether some claim or explanation captures the correct contrastive focus can depend on the context of the situation and on our goal as enquirers, this does not introduce a problematic kind of subjectivity or anti-realism into interventionism. This is because the notion of contrastive focus is not introduced to distinguish between *causal and non-causal* claims and explanations, but rather is introduced to distinguish between *better or worse* causal claims and explanations and as such, it will inevitably feature a certain degree of subjectivity. Again, whether X causes Y, or whether X counts as a causal explanation of Y, depends *solely* on whether there is counterfactual dependence of the right kind between the variables and whether such counterfactual dependence exists is independent of whether we consider those counterfactuals to capture the correct contrastive focus.

So, although interventionism has a subjectivist element in the sense that what we *accept* or *judge* to be causal or explanatory depends on somewhat

subjective notions such as sensitivity and contrastive focus, this does not introduce a *problematic* kind of subjectivity or anti-realism into the theory. This is because, as I hope to have made clear, whether X causes Y or whether X counts as a causal explanation of Y depends *solely* on whether there is counterfactual dependence of the right kind between the variables, (i.e. invariance under interventions) and whether such counterfactual dependence exists is completely independent of any subjective considerations.[46] As Woodward puts it, "the patterns of counterfactual dependence are, as it were, the "objective core" that lies behind our particular causal judgments, and it is such patterns that are the real objects of scientific and practical interest." (Ibid: 85)[47]

In fact, as Woodward explains, it is actually built into the interventionist understanding of causation that causation is genuinely mind-independent. To use one of Woodward's examples, suppose that an agent wishes to bring about some effect, Y, and wonders whether she can use X as a means of doing so. As Woodward explains, although it is up to the agent whether she discovers that this relationship is causal or non-causal, or actually uses X as a means of controlling Y, it is built into the very idea that the agent can discover whether this relationship is genuinely causal via interventions that whether X causes Y is not also 'up to her'. As Woodward explains,

---

[46] It is not clear that Woodward would agree that subjective considerations *never* influence whether some relationship qualifies as causal. For example, in his discussion (2003, especially pp. 87-89) of the notion of 'serious possibility' (which, for our purposes, is relevantly similar to the notion of insensitivity), it is not clear whether Woodward actually thinks that serious possibility can be used to explain how it is possible for a relationship to meet the interventionist criteria for causation and yet fail to be causal, given that the associated counterfactuals are not serious possibilities. In any case, I believe that it is only if subjective considerations, such as serious possibility, never influence whether some relationship qualifies as causal that interventionism is able to avoid problems concerning realism that I discuss below and again in Chapter 5.

[47] Woodward actually makes this point in relation to cases of overdetermination.

> "…it is a presupposition of her deliberation that if it is possible to change *Y* by intervening on *X*, then there must be an independently existing, invariant relationship between *X* and *Y* that the agent makes use of when she changes *X* and, in doing so, changes *Y*—a relationship that would exist and have whatever characteristics it has even if the agent were unable to manipulate *X* or chose not to manipulate *X* or did not exist. In other words, it is built into the whole notion of a manipulation that the agent's activities, manipulative or otherwise, don't somehow create or influence or constitute whether there is a relationship between *X* and *Y* that allows us to manipulate *Y* by manipulating *X*." (Ibid: 119)

Now, one point that will be especially relevant to our later discussion and which is a point that Woodward himself acknowledges, is that this interventionist notion of realism is "metaphysically modest" (Ibid: 121), for example, in comparison to a conception of causation that posits the transfer of some conserved physical quantity as a necessary condition for causation (such as the SP concept). For example, according to interventionism, in order for X to cause Y (i.e. in order for the relationship between X and Y to be minimally invariant and causal), it is not necessary that X and Y are connected via any kind of spatiotemporally continuous physical process, or that X transfers some conserved physical quantity to Y. Moreover, it is also not true that relationships that display a relatively high degree of invariance thereby contribute something more 'metaphysically' than relationships that are less invariant.[48] But rather,

[48] Of course, in interventionist terms, relationships that are highly invariant can nonetheless be considered as more causally relevant than relationships that are less invariant, given the important

interventionist realism only implies "that there be facts of the matter, independent of facts about human abilities and psychology" (Ibid), namely facts about counterfactual dependencies.

Now, as I make clear in the next chapter, this "metaphysically modest" (Ibid: 121) account of causation is actually the *only* account of mental causation and solution to the exclusion problem that we can give as serious *physicalists*, given that it is only by being "metaphysically modest" (Ibid) that this account is able to uphold all of the minimal commitments of non-reductive physicalism. Nevertheless, an important theme of the next chapter will be proving that this "metaphysically modest" (Ibid) account does nonetheless provide a *satisfactory* solution to the exclusion problem. By demonstrating that interventionism is not *straightforwardly* anti-realist in the sense that whether X causes Y depends on facts about us and by demonstrating that it could *never* therefore provide a satisfactory account of mental causation and solution to the exclusion problem, the discussion in this chapter should have gone some way to prove this.

4.5.3 The Problem of Circularity

The final problem that I will discuss concerns the potential circularity of interventionism. This problem arises since the notion of an intervention, which is central to interventionism, is itself a causal notion. In order to see this, note that the criteria for a suitable intervention, (IV1-4), outlined above include the requirements that intervention *I* should *cause* X and that it should not ca*use* Y directly. This raises the question of whether this account of causation is

---

practical and explanatory benefits that are acquired in these cases. This point is especially important to the argument for mental causation outlined in the next chapter.

problematically, or viciously circular in the sense that it employs a causal notion to analyse causation and so cannot explain what causation is.

Now, although the notion of an intervention is itself a causal notion and although this means that interventionism cannot be used to provide a reductive analysis of causation, (more on this below), interventionism is not viciously circular. We can see this by considering the following two points. Firstly, consider Woodward's (Ibid: 104-105) point that the causal information that is required to establish whether intervention *I* is a suitable intervention on X, (for example, that it is a cause of X and that it does not cause Y directly and so on), is not the *same* causal information that is used to establish whether X causes Y. In other words, we have an understanding of what an intervention *I* on X would consist of and although this involves causal notions, it is also true that this is independent of (i.e. it does not presuppose) whether X causes Y. Indeed, the purpose of considering whether an intervention *I* on X changes Y is to establish whether X and Y are causally related. Woodward is right that it would only be if the (causal) notion of an intervention presupposed a causal relationship between X and Y that interventionism would be viciously circular and this is simply not true.

Secondly, although the fact that interventionism appeals to causal notions to define causation means that it cannot provide a reductive analysis of causation, this would only be fatal to interventionism if it were true that only reductive analyses of causation could provide a genuine and non-viciously circular understanding of what causation is. However, this is simply not true. For example, what I hope the discussion in this chapter has demonstrated is that although interventionism is a non-reductive theory of causation it certainly does

provide an understanding of what causation is: it successfully distinguishes between causation and correlation, it explains the potential practical payoff of causation in terms of control and manipulation, it provides a plausible account of causal explanation, it provides an account of how omissions and overdetermining causes can all count as genuine causes, while explaining why our causal intuitions and judgements about these cases nonetheless vary. Moreover, as Woodward (Ibid: 149) points out, since interventionism is inconsistent with many other theories of causation, (including the SP concept), it undermines these alternative theories, all the while being a non-reductive theory.

What the discussion in this last section should have demonstrated is that interventionism does not generate a problematically anthropocentric, anti-realist, or circular conception of causation and that it can therefore provide a coherent account of mental causation and satisfactory solution to the exclusion problem.

4.6 Conclusion

In this chapter, I began by outlining the central features of interventionism and in particular, examined those features of the theory that I appeal to in the next chapter, in which I present Woodward's interventionist account of mental causation. I then highlighted some of the problems that the SP concept faces and presented interventionism as a viable alternative theory of causation that avoids these problems, undermining the assumption of SP and thereby demonstrating that the non-reductive physicalist need not accept Kim's a priori exclusion problem. I then addressed the worry that despite the problems that the SP concept faces, interventionism fails to provide a viable alternative to this theory and so fails to undermine the assumption of SP, since it faces serious

problems of its own. I argued that not only can interventionism avoid these problems, but it can actually deal with many of these problems in a more satisfying way than the SP concept. I concluded that interventionism *does*, after all, provide a viable alternative theory of causation to the SP concept and does undermine the assumption of SP, hence demonstrating that the non-reductive physicalist need not accept Kim's a priori exclusion problem. In the final section, I addressed some general problems concerning the potentially anthropocentric, anti-realist and circular nature of interventionist causation. I demonstrated that interventionism avoids these problems and that it can therefore provide a coherent account of mental causation and satisfactory solution to the exclusion problem.

5. Interventionism and Mental Causation

5.1 Introduction

My aim in this chapter is to present a positive account of mental causation and to demonstrate how this account avoids the exclusion problem, whilst upholding all of the minimal commitments of non-reductive physicalism, thereby providing a successful non-reductive *physicalist* solution to the exclusion problem. Before doing this, it will be useful to recap the argument thus far. In Chapter 2, I demonstrated that the exclusion problem appears to follow a priori from five apparently inconsistent theses of non-reductive physicalism. I argued that these theses are all in fact minimal commitments that cannot be rejected in order to overcome the exclusion problem and that they must therefore be upheld by any physicalist account of mental causation and solution to the exclusion problem. In Chapter 3, I argued that despite its apparent inevitability, the exclusion problem only follows a priori from these minimal commitments when they are combined with an assumption regarding causation, this being the assumption that causation is identical to sufficient production. In the previous chapter, I outlined and examined the interventionist theory of causation and highlighted some problems that the SP concept faces. By highlighting these problems and by demonstrating that interventionism provides a viable alternative theory of causation that avoids these problems, it undermined the assumption that

causation is identical to sufficient production and proved that the non-reductive physicalist need not accept Kim's a priori exclusion problem. In this chapter, I outline Woodward's interventionist account of mental causation and demonstrate how this account avoids the exclusion problem, whilst upholding all of the minimal commitments of non-reductive physicalism, thereby providing a viable and successful non-reductive *physicalist* account of mental causation and solution to the exclusion problem.

The chapter is organised as follows: in Section 5.2, I outline the interventionist account of mental causation. I demonstrate that not only does interventionism provide an account of causation by which both mental and physical properties can qualify as causes of the same effects, but that when causation is understood in interventionist terms, mental properties can actually be considered to provide better causal explanations of their effects in comparison to those offered by their physical realizers. In order to demonstrate this, I appeal to the central features of interventionism that I outlined in the previous chapter, namely invariance (Section 5.2.1) and contrastive focus (Section 5.2.2). Most importantly, I demonstrate that when causation is understood in interventionist terms, the question of mental causation becomes an entirely *a posteriori*, not *a priori* question. In Section 5.3, I make explicit how this account of mental causation avoids Kim's a priori exclusion problem and in Section 5.3.1, demonstrate that this account upholds all of the minimal commitments of non-reductive physicalism and does therefore provide a viable non-reductive *physicalist* solution to the exclusion problem. Then, in Section 5.3.2, I demonstrate that this account also provides a *satisfactory* account of mental causation and solution to the exclusion problem. Finally, in Section 5.4, I

examine two alternative manipulationist accounts of causation, which I argue fail to provide satisfactory accounts of mental causation and solutions to the exclusion problem, given that they generate anti-realist conceptions of causation. I conclude that Woodward's interventionist account of mental causation therefore provides the *only* satisfactory non-reductive physicalist account of mental causation and solution to the exclusion problem.

5.2 An Interventionist Account of Mental Causation

It appears, at least at first glance, that interventionism provides a straightforward account of the causal relevance of mental properties. We do, for example, routinely observe that intervening on our own or others' mental (or psychological) states is a way of bringing about both physical and psychological effects. For example, I may tell you that there is no more milk in the fridge (intervening on your belief[1]), causing you to go to the supermarket. Or I may manipulate your belief that there are biscuits in the cupboard, so that you do not eat them.

[1] This does depend on the acceptance of the commonsensical idea that it is possible to intervene on others' mental states through verbal communication, etc. Campbell (2007) questions whether these 'ordinary', or folk-psychological (i.e. non-idealised) interventions should be required to be 'surgical' in the specific sense captured by condition IV-1 (which, remember, requires that intervention *I* should 'break ties' with any endogenous causes of X, to rule out the possibility of confounding) in order to be able to determine whether X causes Y. The reason that Campbell gives is that this would implausibly require that the surgical intervention 'suspend' the subject's rationality, e.g. break ties with the subject's usual reasons for possessing mental property X, which are also causes of X. I will not discuss this issue in great detail, but it will be helpful to make the following points. Firstly, (if I have understood Campbell correctly), I do not agree that we should modify the criteria for a suitable intervention in the case of psychological causation *in general*, since these criteria still provide *idealised* conditions (in hypothetical or actual experimental situations) to determine whether X causes Y. However, I do think that it is plausible that we can (and do) routinely intervene on our own and others' psychological states and agree with Campbell that it is unlikely (and somewhat implausible) that these interventions are strictly surgical, and *moreover* agree that we are justified in making causal inferences on the basis of these interventions, so long as we recognise that these interventions simply take the form of *non-ideal* interventions for determining whether X causes Y.

These kinds of examples of mental causation are commonplace in ordinary life, but the idea that mental states can be causes of both physical and psychological effects is also widely accepted in scientific practice, especially in psychology and other social sciences. For example, experiments across the sciences attempt to control for the so-called placebo effect, in which the mere belief that a subject will receive treatment can bring about a physical change to their recovery. Or consider Woodward's example from psychiatry that since positive thinking is associated with changes in depression, some claim that positive thinking can be considered as part of an effective treatment for depression. Since it appears that there are some interventions on these mental properties that change these physical and psychological effects, they can qualify as genuine causes according to interventionism. However, as we will see, there are examples of seemingly intuitive cases of mental causation, which fail to qualify as causal according to interventionism.[2] I will now therefore examine the interventionist account of mental causation in detail.

Woodward (2008a) cites recent research into the neural coding of intentions to reach for specific objects carried out by Richard Andersen and colleagues at Caltech.[3] The research involved recording neural signals in the PRR (parietal reach region) of the brain in Macaque monkeys, which is thought to encode for 'intentions to reach for specific targets'. Woodward notes that the researchers were able to relate variations in 'aggregate features' of the neural signals to variations in intentions to reach for specific goals (as evidenced by the reaching behaviour of the monkeys) and that they were able to accurately

[2] As I will shortly explain, this is because those mental properties fail to stand in the particular relationship to those effects that is required for mental causation (or more accurately, for all supervenient causation), namely a 'realization independent dependency relation', or RIDR.
[3] This reference is listed as Musallam et al (2004) in the bibliography.

'forecast' specific reaching behaviour from these neural aggregates. What was also apparent was that although it was possible to relate an intention to a specific aggregate pattern of neurons, the same intention might be realized by a variety of neural patterns, so that each intention is multiply realized at the neural level.

Consider a particular instance of this experiment:

> "Suppose then that on some specific occasion *t* a monkey forms an intention I_1 to reach for a particular goal—call this action R_1. Suppose N_{11} is the particular (token) pattern of firing in the relevant set of neurons that realizes or encodes the intention I_1 on this particular occasion. Assume also that there are other token patterns of neural firing, N_{12}, N_{13} that realize the same intention I_1 on other occasions, so that I_1 is multiply realized by N_{11}, N_{12}, etc." (Ibid: 239)

Now, according to interventionism, mental property I_1 qualifies as a cause of physical effect R_1 if there is some intervention on I_1 that changes R_1, i.e. if the relationship between I_1 and R_1 is at least minimally invariant. For example, if in the experiment, an intervention sets the value of the intention from I_1 to I_2, and thereby makes it the case that the monkey exhibits reaching behaviour R_1 rather than reaching behaviour, R_2, I_1 will qualify as a cause of R_1.

Moreover, it will *also* be true (as guaranteed by causal closure, more on this below) that physical realizer N_{11} qualifies as a cause of R_1, since there is *some* intervention on N_{11} that changes R_1, i.e. the relationship between N_{11} and R_1 is minimally invariant. For example, if we imagine altering the monkey's

neural firing pattern from N_{11} to N_{15}, this may result in different reaching behaviour, R_5 being performed, rather than R_1.[4]

So, by appealing to the theory of interventionism, it is possible to provide a fairly straightforward account of how *both* supervenient mental property, I_1, and its physical realizer, N_{11}, can qualify as causes of the same physical effect. In fact, although it is true that both I_1 and N_{11} qualify as causes of R_1, there is a sense in which the causal claim and explanation citing I_1 can actually be considered as *better* than the causal claim and explanation citing N_{11}. In what sense can the causal claim and explanation citing I_1 be considered as 'better' and what explains this difference?

5.2.1 Invariance and Realization Independent Dependency Relations (RIDR)

According to Woodward, whenever *any* supervenient property, such as a mental property, stands in a particular relation to some effect, namely a 'realization independent dependency relation', or RIDR (to be discussed below), it will usually be the case that that supervenient property will provide a preferable causal claim and explanation in comparison to its subvenient realizer. On the other hand, when supervenient properties fail to stand in this specific relationship with those effects, it will instead be the case that the subvenient property will provide the preferable causal claim and explanation (as I shortly

[4] As I explain in Chapter 6, given supervenience, whenever a mental property qualifies as a cause of some effect, it is actually the *same* intervention that secures the causal status of the physical realizer of that mental property. Very roughly, this is because supervenience requires that any change at the mental level requires a change at the physical level. Consequently, any intervention that changes the mental property from one value to another (for example, from I_1 to I_2) will also change the value of the physical realizer of that property (for example, from N_{11} to N_{14}) and hence also secures the causal status of that physical realizer. I explore this issue in more detail in Chapter 6 and address a potential problem that arises regarding the non-reductive physicalist's commitment to non-identity.

explain, in this kind of case, the supervenient property will actually fail to qualify as a cause of the effect).

So, what exactly is RIDR? As Woodward explains,

> "…what is required [for RIDR] is the existence of a relationship that both involves a dependency between the upper level variables (different values of M_1, produced by interventions map into different values of M_2) and that is realization independent in the sense that it continues to stably hold for a range of different realizers of these values of M_1 and M_2. It is the presence of this sort of *realization independent dependency relationship* (hereafter *RIDR*) that ensures that interventions that change M_1 are stably associated with changes in M_2—hence that M_1 causes M_2." (Ibid: 241)

In other words, whenever supervenient mental properties stand in this realization independent dependency relation to other properties, those relationships will display at least a minimal degree of invariance and will qualify as causal. I suggest that we can therefore think of RIDR as what *makes it possible* for supervenient mental properties to stand in invariant and hence causal relationships with other properties.

Moreover, we can also see that when the relationships at the supervenient level display a high degree of realization independence, those relationships will display a higher degree of invariance than the relationships at the subvenient level. This is simply because, given that the supervenient properties are realized by a variety of physical properties, each of which lead to the same effect, there will simply be a wider range of interventions on those supervenient properties

that change the effect, than there are for those physical properties. For example, since I_1 is realized by a variety of neural properties, each of which lead to the same reaching behaviour, R_1, there will simply be a wider range of interventions on I_1 (for example, interventions that change the intention from I_1 to I_2, I_3, I_4 and so on) that change R_1, than there are for N_{11}. (Remember that changing the value of the neural realizer from N_{11} to one of the other realizers of I_1, for example, to N_{12}, or N_{13} would fail to bring about any change to R_1.)

Going back to the discussion in Chapter 4, we can see that this high degree of invariance has two important benefits. Firstly, by being invariant over a wider range of interventions, the explanation citing I_1 will be able to answer a wider range of w-questions, since it tells us what *would* happen to the effect under a wide range of interventions on I_1. For example, it tells us what *would* happen to physical effect, R_1, if we were to change the subject's intention from I_1 to I_2, I_3, I_4, and so on. By contrast, given that the relationship between N_{11} and R_1 is invariant over a much more limited range of interventions, the explanation citing N_{11} will be able to answer a much more limited range of w-questions. For example, it will fail to tell us what would happen to the physical effect if we changed N_{11} to one of the alternative realizers of I_1, such as N_{12}, or N_{13}.

Secondly, given that the relationship between I_1 and R_1 holds over this wide range of interventions, it will be more exploitable for the purposes of control and manipulation, since this relationship will continue to hold and hence continue to provide a means of control, over this wide range of interventions. By contrast, the relationship between N_{11} and R_1 will be less useful for the purposes of control and manipulation, since it will break down outside a narrow range of interventions.

We can illustrate the relevance of this level of control and manipulation in the case of mental causation by considering what the eventual goal of the research of Andersen et al is. As Woodward notes, the researchers hope to use this information to provide paralysed subjects with control over prosthetic limbs. Now, although it is true that in order to produce the movement in the prosthetic limb, the prosthesis must be 'wired up' to the subject's neural system, it will be 'wired up' in such a way that on the mere formation of an *intention* by the subject, for example, on the formation of intention I_1, any one of the specific realizers of I_1 could be instantiated and would bring about a change to the effect. (By contrast, imagine that the prosthesis was 'wired up' directly to one single neural property, for example N_{11}, in which case given that certain interventions on this specific neural property would *not* bring about a change to the effect, this level of control would simply be lost[5].) In other words, it is by forming an *intention* that the subject will acquire the desired level of control over the prosthetic limb, the relevance of which for the paralysed subject goes without saying.

I suggest that this demonstrates further that although interventionism provides an account of causation by which both mental properties and their physical realizers can qualify as causes of the same effects, there is a real and important sense in which the relationships and explanations at the mental level can be considered as better than those at the physical level and this can simply be explained in terms of the fact that those relationships are more stable, or invariant over a wider range of interventions. Moreover, it highlights the fact that from an

[5] Note that from the perspective of the SP concept of causation, this practical benefit is completely lost. This is because by being sufficient to produce the behaviour, physical property N_{11}would automatically be considered as preferable over mental property I_1, even though it is the latter, not the former that is most useful for the purposes of control and manipulation.

interventionist perspective, there is nothing 'special' or privileged about the explanation citing N_{11} that follows from the fact that it is a *physical* explanation. (I explore these issues in further detail below.)

As well as explaining how supervenient mental causation is possible, the notion of RIDR therefore also explains why supervenient mental properties will often provide better causal claims and explanations than their physical realizers: by standing in highly realization independent dependency relationships with their effects, those supervenient relationships will simply be more invariant over a wider range of interventions and hence will be able to answer a wider range of w-questions and will be more potentially useful for the purposes of control and manipulation.[6]

Now, one could argue that I have merely demonstrated that mental properties often provide better *explanations* of their effects in comparison to their physical realizers and that they merely play an 'instrumental' role in explaining behaviour, but that this is very different from proving that mental properties are genuine *causes* of physical and psychological effects.

In response, it is important to remember that according to interventionism, although explanation and causation are distinct notions in the sense that explanation is essentially an epistemic activity, concerned with the provision of causal information, whereas causation is an objective relation existing independently of any epistemic awareness of it, explanations nonetheless cite genuine causes. So, the worry that mental explanations are merely

[6]As I explain in further detail in Section 5.3.2 below, the notion of RIDR also explains how mental properties can qualify as *causally* distinct (i.e. as causes that cannot be identified or reduced) from the physical properties on which they supervene, allowing the non-reductive physicalist to avoid the threat of reduction. (Very roughly, this is because it allows mental and physical properties to exhibit *distinct* levels of invariance in relation to their effects.)

instrumental in explaining behaviour, rather than describing genuine causal relationships, simply does not arise. Secondly, remember that according to interventionism, *all* that is required for some property to qualify as a cause of some effect is that there is some intervention on that property that changes the effect and according to this definition, mental property I_1 qualifies as a bona fide cause of physical effect R_1.

Before I provide some examples that will help to clarify the notion of RIDR, it is worth pointing out that it is not simply the fact that mental properties are *multiply realized* that ensures that they can qualify as causes, but it is the multiple realizability of mental properties *and* the specific notion of RIDR that makes this possible. As we will see, it is possible for a supervenient property to be multiply realized at the physical level, but for the relationship between that mental property and the effect to fail to be realization independent and hence fail to exhibit any degree of invariance and hence qualify as causal.

As an illustration, consider the following example of Woodward's (Ibid: 242): an ordinary roulette wheel is spun by a croupier C, who has a varied set of hand movements B_i that he can use to spin the wheel. Although it may be true that C is able to distinguish B_i in a fairly fine grained way, so that C has maximal control over B_i, each B_i will be multiply realized at the micro level given all of the different variations of the starting positions and momenta of the wheel. Suppose then that on some particular occasion, C employs movement B_k to spin the wheel and the ball falls into a red slot. Now, according to interventionism, although B_k is multiply realized, B_k does not cause the ball to fall into the red slot, since there is no intervention on B_k that changes whether the ball falls into

the red slot.[7] This is because whether or not the ball lands in the red slot depends on the specific realization of B_k, which varies on every occasion that C employs B_k. Since the relationship between B_k and the ball falling into the red slot is not realization independent the relationship fails to be even minimally invariant under interventions to B_k and hence fails to qualify as causal, even though B_k is multiply realized.

In order to elucidate the notion of RIDR and the argument for supervenient mental causation further, it will be helpful to appeal to some examples. Firstly, let us consider the case of mental property I_1: as I explained above, the relationship between I_1 and R_1 is highly realization independent in the sense that it would continue to hold for all of the different physical realizations of I_1. I illustrate this in Figure 5.1 below.

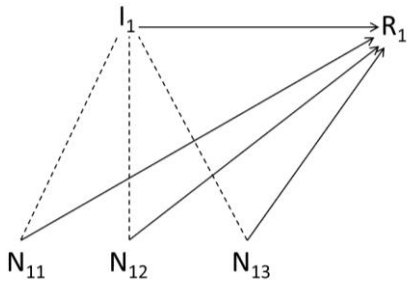

Figure 5.1: RIDR/Intention I_1

Solid arrows represent genuine causal relationships. Broken lines represent supervenient relationships.

[7] Remember that although it may be true that there is one single intervention on B_k that changes whether the ball falls into the red slot, this does not guarantee that B_k qualifies as a cause of the effect, since remember that according to interventionism, for X to cause Y, even if it is only discovered by one single intervention, that intervention must in theory be repeatable, to rule out the possibility that the intervention on X is only associated with the change in Y by chance.

We can see that by standing in this realization independent dependency relationship to R_1, interventions on I_1 will be stably associated with changes to R_1; hence I_1 will qualify as a cause of R_1. *Moreover*, by being highly realization independent and hence by displaying a higher degree of invariance than the relationship between N_{11} and R_1, the causal claim and explanation citing I_1 will be better than the one citing N_{11}. As I explained above, this is simply because by being stable or invariant over a wider range of changes, the explanation citing I_1 will answer a wider range of w-questions and will be more useful for the purposes of control and manipulation, which I hope to have shown in the case of mental causation, is especially important.

On the other hand, let us consider those supervenient relationships that fail to be even minimally realization independent and hence fail to be minimally invariant and causal. According to Woodward, these kinds of non-RIDR cases are usually cases of 'causal heterogeneity', in which a supervenient property, such as a single mental property is multiply realized by a variety of physical properties and in which case each different realization of that mental property leads to a different effect.

As an illustration, consider the following example of Woodward's (Ibid: 260-261): to the extent that the general concept 'fear' is realized by a number of more specific 'fear systems' that are causally heterogeneous in the sense that each of the different physical realizations is associated with a different effect, it is likely that any generalization linking the general concept 'fear' to a particular behavioural effect will be completely unstable and hence will fail to qualify as minimally invariant and causal. Put slightly differently, given that any intervention on the property 'fear' will instantiate any one of the specific

realizers of 'fear', each of which lead to a different effect, there is no intervention on the general property 'fear' that would lead to a change to the effect under consideration.[8] Instead, it is only by intervening on one of the specific physical realizers of 'fear' that we will find any invariant and hence causal relationship between cause and effect. I illustrate this in Figure 5.2 below. In line with the interventionist account of causation and explanation outlined thus far, these non-RIDR and hence non-invariant and non-causal cases would not qualify as even minimally explanatory or potentially useful for the purposes of control and manipulation.

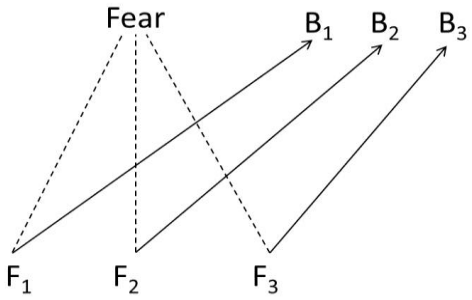

Figure 5.2: Non-RIDR/Fear

Solid arrows represent genuine causal relationships. Broken lines represent supervenient relationships. Fear systems F_1, F_2 and F_3 represent the different realizers of the supervenient concept 'fear'. B_1, B_2 and B_3 represent different behavioural effects.

[8] As I mentioned in footnote 7 in relation to the roulette wheel example, although it may be true that there is one single intervention on 'fear' that changes, for example, behavioural effect B_1, this does not guarantee that 'fear' qualifies as a cause of B_1, since given the lack of realization independence of the relationship between 'fear' and B_1 (which can in turn be explained by the causally heterogeneous nature of the realizers of 'fear') there is no intervention on 'fear' that would change B_1 that would be repeatable in the sense required to rule out the possibility that the intervention on 'fear' is associated with the change in B_1 by chance.

Further still, I suggest that in other kinds of cases, in which the relationships at the supervenient level possess a fairly low degree of realization independence and hence display a lower degree of invariance than the relationships at the physical level, we may prefer the causal claims and explanations offered by the subvenient realizers of those supervenient properties. As an illustration, let us again consider the psychological concept of positive thinking: it is possible for the relationship between positive thinking and depression to be somewhat realization independent, in the sense that there are a number of realizations of positive thinking that lead to a change in depression. As a result, there will be *some* intervention on positive thinking that changes depression, hence this relationship will display a minimal degree of invariance and will qualify as causal. However, it may also be true that there are some realizations of positive thinking that are not associated with changes to depression, but which may instead be associated with changes to the immune system, or scholastic achievement, for example, and the relationship between positive thinking and depression will break down under any intervention on positive thinking that instantiates one of these realizers. I illustrate this in Figure 5.3 below. As a consequence, the relationship between positive thinking and depression will not be as invariant as the relationships at the physical level between the specific realizers of positive thinking and the various effects. As a result, the explanations offered at the supervenient level will not answer as wide a range of w-questions, or provide information that is as useful for the purposes of control and manipulation as the explanations offered at the physical level.

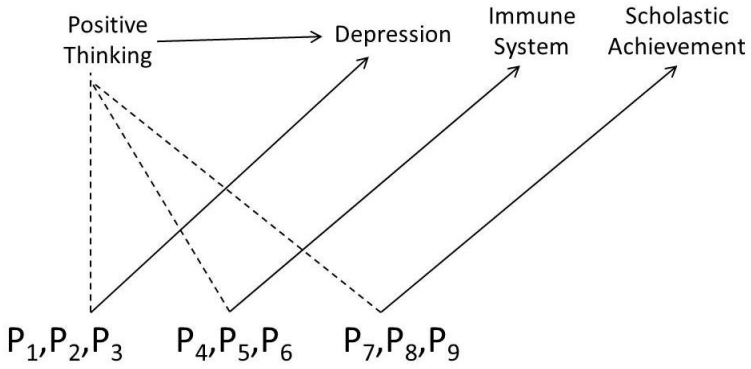

Figure 5.3: Minimal-RIDR/Positive thinking

Solid arrows represent genuine causal relationships. Broken lines represent supervenient relationships. P_1, P_2 and P_3 represent the group of physical realizers of positive thinking that cause changes in depression. P_4, P_5 and P_6 represent the group of physical realizers of positive thinking that cause changes in the immune system. P_7, P_8 and P_9 represent the group of physical realizers of positive thinking that cause changes in scholastic achievement.

We can elucidate these ideas further by referring back to the idea of 'degrees of invariance' that I introduced in the previous chapter. Remember that according to interventionism, there is both a threshold of invariance that a relationship or generalization must pass if it is to qualify as causal *and* varying degrees of invariance that a relationship or generalization can possess. In a recent paper, Woodward (2008b) importantly asks where psychological relationships and generalizations are likely to lie on the scale of invariance. Woodward's suggestion is that they will typically lie around the middle of the scale, in the sense that they fail to be as highly invariant as the claims and generalizations of physics, for example, but that many will pass the threshold for invariance and will qualify as genuinely causal.

How does the notion of RIDR fit into this scale? I suggest that we can think of the scale of RIDR as 'mapping onto' the scale of invariance (I illustrate this in Figure 5.4 below). In general, we can think of psychological causal claims and generalizations that are highly realization independent as lying at the upper end of the mid-range of the scale, (for example, as in the case of intention $I_{1)}$, whereas I suggest that we can think of mental properties that do not display a great deal of realization independence as lying at the latter end of the mid-range, (for example, as in the case of the psychological variable 'positive thinking'). In other kinds of cases, in which the mental properties are not even minimally realization independent and hence are not minimally invariant or causal, (for example, as in the case of the general concept 'fear'), we can see that those relationships will fail to pass the threshold of invariance and will fail to qualify as genuine causes.

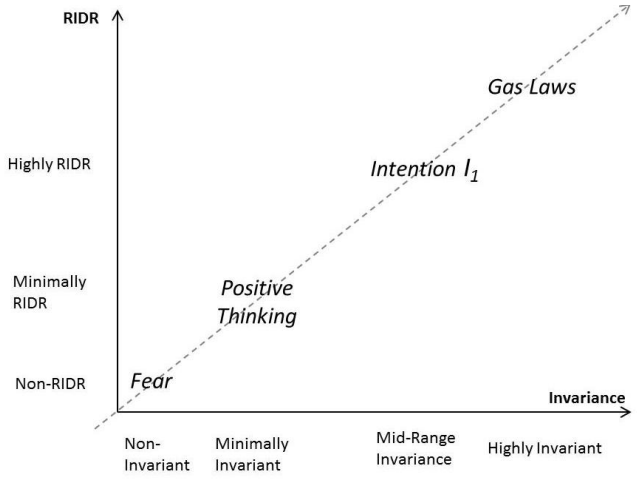

Figure 5.4: The Scale of Invariance and RIDR[9]

[9] It is interesting to note that it is not just mental properties that can stand in realization independent relationships with their effects and hence exhibit distinct levels of invariance in comparison to their more specific physical realizers, but as the figure above illustrates, macro*physical* properties, such as those invoked in the gas laws can also display a degree of realization independence and hence a distinct level of invariance in comparison to their specific *micro*physical realizers. Woodward (2008a: 233) provides a detailed example.

What I hope to have shown in this discussion is that when the relationships at the supervenient level are highly realization independent and hence display a higher degree of invariance than the relationships at the physical level, supervenient mental properties will often provide better causal claims and explanations than their physical realizers, given that those supervenient relationships will be able to answer a wider range of w-questions and will be more potentially useful for the purposes of control and manipulation (the implications of this level of control in the case of mental causation were made clear above). In other cases, in which the relationships at the supervenient level are less realization independent and hence less invariant than the relationships at the physical level, we may prefer the causal claims and explanations offered by the subvenient realizers of those supervenient properties, since those explanations will now answer a wider range of w-questions and will provide information that is more potentially useful for the purposes of control and manipulation. Further still, there are cases in which the supervenient relationships in question are non-RIDR and non-invariant and hence non-causal, leaving only the physical realizers of those supervenient properties to qualify as causes of the effects.

Most importantly for the purposes of my argument, according to this account, whether and to what extent a supervenient relationship is RIDR and invariant and hence whether a supervenient mental property qualifies as a cause of some effect, and further still, whether it qualifies as a *preferable* cause of that effect over its physical realizer, become entirely *a posteriori* questions, dependent on the specific details of the case at hand, rather than something that can be settled a priori, as Kim suggests. Woodward is correct when he suggests

that this a posteriori question is the *only* relevant question left regarding the causal status of mental properties:

> "Whether and to what extent such stability is present is an empirical question that depends both on the upper level relationship and the nature of their realizers and the generalizations governing them. I want to conclude this essay by suggesting that to the extent there are issues about the reality and extent of mental causation, these have to do with such empirical consideration, rather than with the very general arguments for the causal inertness of the mental…" (Woodward, 2008a: 259-260)

5.2.2 Contrastive Focus

I demonstrated above that the notions of RIDR and invariance explain how it is possible for mental properties to qualify as causes of their effects *and* explain how it is possible, in some cases, for mental properties to provide better causal claims and explanations than their physical realizers. There is another feature of interventionism that I introduced in the previous chapter, which also distinguishes between better or worse causal claims and explanations and which is therefore relevant to our current discussion. This is the notion of contrastive focus.

In this next section I demonstrate how the notion of contrastive focus explains how causal explanations that cite mental properties can often be considered as preferable in comparison to causal explanations that cite the physical realizers of those mental properties. In order to illustrate this, let us recall Yablo's example of the trained pigeon introduced in the previous chapter.

In this example we saw that although the fact that the object is scarlet qualifies as a cause of the pecking behaviour (since there is some intervention on the property of scarlet that changes the behaviour), the more specific explanation citing the property of scarlet is deficient in comparison to the explanation citing the property of red. This is because by being overly specific, it fails to capture exactly which changes to the cause variable (namely a change from 'red' to 'not red') are associated with changes to the effect variable (namely a change from 'pigeon pecks' to 'pigeon does not peck'). Moreover, the explanation citing the property of scarlet is potentially misleading, since it suggests that the pigeon would fail to peck in any case in which the object is not scarlet. In Yablo's terms, the property of scarlet fails to be 'proportionate' to its effect, in the sense that the causal explanation citing the property of scarlet fails to convey *all and only* such information about specific patterns of counterfactual dependence between cause and effect, (in this case, by both omitting relevant detail about such dependencies and by including irrelevant detail).

Moreover, remember that by failing to capture the exact range of changes to the cause variable that are associated with changes to the effect (and in fact by providing potentially misleading information about such changes), the explanation citing the property of scarlet will provide information that is less useful for the purposes of control and manipulation. By contrast, given that it is specifically the contrast between whether the object is 'red' or 'not red' that is associated with changes to whether the pigeon 'pecks' or 'does not peck', the explanation citing the property of red will provide information with which we can *stably* and *systematically* control the effect.

As Woodward points out in the following passage, the same seems to be true for the example of the research of Andersen et al:

> "Just as with [Yablo's example], the causal claim/causal explanation that appeals to N_{11} to explain R_1 seems overly specific. It fails to convey a relevant pattern of dependence: that there are some alternatives to N_{11} (namely, N_{12} and N_{13}) that would have led to the same reaching behavior R_1 and other alternatives (those that realize some different intention I_2, associated with reaching for a different goal) that would not have led to R_1. Put slightly differently, Andersen's concern in this example is in finding the cause of variations in reach toward different goal objects—why the monkey exhibits reaching behavior R_1 rather than different reaching behavior R_2. According to the interventionist account, to do this, he needs to identify states or conditions, variations in which, when produced by interventions, would be correlated with changes from R_1 to R_2. Ex hypothesi, merely citing N_{11} does not accomplish this, since it tells us nothing about the conditions under which alternatives to R_1 would be realized. By way of contrast, appealing to the fact that the monkey's intention is I_1 rather than some alternative intention I_2 does accomplish this, assuming (as we have been all along) that there is a stable relationship between the occurrence of I_1 (however realized) and R_1 and that under I_2 some alternative to R_1 (reaching toward a different goal) would have occurred." (Ibid: 239)

In other words, by being overly specific and hence by failing to capture the exact range of changes to the cause variable that lead to stable changes to the behavioural effect, namely changes to $I_{1,}$ (and in fact, by providing potentially misleading information about such changes), the explanation citing N_{11} fails to capture the correct contrastive focus and consequently provides a deficient explanation of the effect in comparison to the one citing I_1. Moreover, by failing to capture the exact range of changes to the cause variable that are *stably* and *systematically* associated with changes to the effect (and by providing potentially misleading information about such changes), the explanation citing N_{11} will also provide information that is less useful for the purposes of control and manipulation, in comparison to the explanation citing I_1. (For the paralysed subject who would presumably wish to acquire this kind of stable and systematic control over the prosthetic limb, the relevance of this information goes without saying.)

One final aspect of contrastive focus that will be useful to highlight at this stage (which I briefly drew attention to in Chapter 4) is that the correct contrastive focus of some explanation or causal claim can vary depending on the context of the situation.[10] I illustrated this with Woodward's (2008a: 236) variant of the Yablo example in which the pigeon is trained to peck specifically on the presentation of scarlet objects. In this example, we saw that the explanation citing scarlet now captures the correct contrastive focus and will now be considered as preferable in comparison to the explanation citing the property of red. This is because it is now the contrast between whether the object is scarlet,

[10] I also explained that the correct contrastive focus of some explanation or causal claim can vary depending on the goal of the enquirer.

rather than not scarlet that is specifically associated with changes to whether the pigeon pecks, or fails to peck, rather than the contrast between whether the object is red or not red. Moreover, it is now *this* explanation that provides information that is most useful for the purposes of control and manipulation.

As Woodward points out, what this suggests is that there is nothing privileged about the explanation citing I_1 that follows from the fact that it is a *mental* explanation. By the same token, I suggest that there is nothing privileged about the explanation citing N_{11} that follows from the fact that it is a *physical* explanation, as I have argued Kim's argument suggests. Rather, depending on the details of the case, the correct contrastive focus might be captured by an explanation that cites some mental property, or by an explanation that cites the physical realizer of that mental property. Most importantly for my argument, what this again emphasises is that according to interventionism, the question of whether mental properties provide 'better' causal explanations of their effects in comparison to their physical realizers is not a matter that can be settled a priori, as Kim's argument suggests, but is instead an a posteriori question, which will depend on the specific details of the case.

In summary, I have demonstrated that interventionism provides a straightforward account of how both mental properties and their physical realizers can qualify as causes of the same effects; if interventions on mental or physical properties are associated with changes to the effects under consideration, they will qualify as bona fide causes according to interventionism. Moreover, I have demonstrated that whenever mental properties stand in highly realization independent relationships to their effects and hence display a relatively high degree of invariance *and* capture the correct contrastive focus,

they can actually be considered to provide better causal explanations than those offered by their physical realizers.

Most importantly, what this discussion should have shown is that whether and to what extent a mental level relationship or explanation is RIDR, invariant, or provides the correct contrastive focus and hence whether some mental property qualifies as a cause of some effect *and* qualifies as a preferable cause in comparison to its physical realizer, are *all* a posteriori questions that are to be determined depending on the details of the case at hand. What this means is that there is nothing on the interventionist account of causation that *a priori* excludes mental properties from qualifying as genuine causes, but rather, the question of mental causation and the question of whether it is at the mental or physical level that we will find preferable causal claims and explanations become entirely a posteriori questions.

5.3 A Solution to the Exclusion Problem

Thus far I have demonstrated that interventionism provides a fairly straightforward account of how *both* mental properties and their physical realizers can qualify as causes of the same effects and that according to this account, the question of mental causation becomes an entirely a posteriori, *not* a priori one. Although I argued in Chapter 3 that Kim's exclusion problem only follows a priori for the non-reductive physicalist when combined with the assumption of SP, which I went on to undermine and although interventionism demonstrates that the question of mental causation is an entirely a posteriori, not a priori one, it is worth making clear exactly how this interventionist account of

mental causation avoids Kim's a priori exclusion problem. In order to do this, I will refer back to the arguments that I made in Chapter 3.

The first way that I argued that the assumption of SP motivates the exclusion problem was in the sense that if one assumes that causation is *identical* to sufficient production it suggests that by being sufficient to produce its effect, that property simply exhausts all there is to cause and explain regarding that effect and implies that there would literally be 'nothing left' for any additional property to causally contribute. Given the minimal commitments of non-reductive physicalism, it seemed that the exclusion problem was inevitable: supervenience guarantees that any supposed mental cause of a physical effect necessarily supervenes on a physical cause, which causal closure states is *sufficient* for that effect. Then, given that this cannot be a case of overdetermination, whereby both the mental and the physical property could qualify as metaphysically distinct, sufficient causes of the effect, it seems that there really would be 'nothing left' for the mental property to causally contribute. On the assumption that causation is identical to sufficient production, physical causation, by its very definition would capture all there is to cause and explain regarding the occurrence of an effect and would seem to make mental causation 'dispensable'.

However, we can see that when causation is understood in interventionist terms, the idea that physical causes leave 'nothing left' for mental properties to contribute, thereby rendering them 'dispensable' simply does not make sense. As an illustration, consider again the example of the research of Andersen et al: according to interventionism, mental property I_1 qualifies as a bona fide cause of physical effect R_1, since there is an intervention on I_1 that changes R_1. I_1's causal

status is not somehow undermined, or made 'dispensable' by the fact that I_1's physical realizer, N_{11}, also qualifies as a cause of R_1, since so long as interventions on I_1 are associated with changes to R_1, I_1 qualifies as a cause of R_1, regardless of whether R_1 has any additional causes. In fact, as the discussion above illustrated, according to interventionism, far from being rendered dispensable, mental properties can often be considered as providing preferable causal claims and explanations in comparison to those offered by their physical realizers. It is of course true that causal closure guarantees that physical causes, such as N_{11}, are sufficient to produce the occurrence of their effects, but, as I argued in Chapter 3, without the assumption that sufficient production is identical to causation, there is no reason to conclude that physical causation somehow renders mental causation as dispensable. Rather, I suggested that we think of the fact that physical causes are, by definition, sufficient to produce their effects merely as an empirical fact about physical causation.

The second way that I argued that the assumption of SP motivates the exclusion problem was in the following sense: given that cases of mental causation cannot be cases of overdetermination, whereby mental properties could cause their effects via metaphysically distinct, sufficient productive chains or processes and given the assumption that this kind of sufficient production is identical to causation, it implied that mental properties must derive their causal status from the only productive (and hence only causal) processes and chains, these being the *physical* processes and chains. This led Kim to conclude that mental properties are reducible to the physical properties on which they supervene, since the causal powers of those mental properties are apparently

acquired in virtue of and are identical and hence reducible to the causal powers of those physical properties.[11]

However, once again, we can see that in interventionist terms, the idea that mental properties must derive their causal status from the physical properties on which they supervene does not make sense. Consider again the research of Andersen et al: according to interventionism, mental property I_1 qualifies as a *distinct* and *irreducible* cause of R_1, in addition to R_1's physical cause, N_{11}, since there is some intervention on I_1 that changes R_1 *and* since the relationship between I_1 and R_1 displays a distinct level of invariance in comparison to the relationship between N_{11} and R_1 (I discuss this issue of causal distinctness below and again in Chapter 6). As I argued in Chapter 3, one would only reach the conclusion that I_1 must derive its causal status from N_{11} if one assumed that causation is identical to sufficient production, since it is true that I_1 could have no 'new causal powers', independent of N_{11} in *this specific* sense.

In summary, what I hope to have made clear is that interventionism not only provides an account of causation by which both mental properties and their physical realizers can qualify as causes of the same effects, but that when causation is understood in interventionist terms, rather than in terms of sufficient production, Kim's exclusion argument simply does not go through.

5.3.1 A *Physicalist* Solution?

In Chapter 2 I argued that any successful non-reductive physicalist account of mental causation must uphold all of the minimal commitments of non-reductive physicalism, namely mental causation, non-identity, supervenience,

[11] See Kim's (2003b: 208) 'Causal Inheritance Principle'.

causal closure and non-overdetermination, if that account is to provide a viable non-reductive *physicalist* solution to the exclusion problem. In fact, in Chapter 3 I suggested that the *real* challenge that faces the non-reductive physicalist regarding mental causation is providing an account of mental causation that explains how mental properties can have genuinely distinct causal roles (thus avoiding the threat of reduction), whilst being ontologically identical with and metaphysically inseparable from their subvenient physical realizers, which are sufficient to produce their effects.

In this section I address each of the commitments of non-reductive physicalism, in order to demonstrate that the interventionist account of mental causation outlined in this chapter does uphold all of these commitments and does therefore provide a viable non-reductive *physicalist* solution to the exclusion problem.

Let us begin with the thesis of mental causation, which is usually understood as the thesis that distinctly mental *properties*, such as intentions, beliefs and desires, have physical effects. Now, it seems clear that the interventionist account of mental causation does uphold this thesis. For example, in the case of the research of Andersen et al, it is the fact that the monkey instantiates *mental property* I_1 on this occasion that causes physical effect R_1 to be instantiated. As I explained in Section 5.2.1, this kind of mental causation is possible because the relationship between I_1 and R_1 is realization independent and hence invariant. I also explained that according to interventionism, whenever *any* mental property stands in this particular relationship to some effect, that property will not only qualify as a cause of the effect in question, but may

actually qualify as a preferable cause of that effect, in comparison to its physical realizer.

What about the non-reductive physicalist's commitment to the thesis of non-identity? (Note that the points that I discuss in this section will be especially relevant to my argument in the next chapter.)

Now, as I briefly mentioned above, according to interventionism, supervenient mental properties (that qualify as causes of effects), although not *ontologically* or *metaphysically* distinct from the physical properties on which they supervene, can nonetheless qualify as *causally distinct* from those physical properties (i.e. as causes that cannot be identified with or reduced to those physical properties). That this is possible is implied in the discussion in Chapter 4, but since this issue is especially important to the argument in this section and to the argument in Chapter 6, it is worth making this issue of causal distinctness explicit.

According to the interventionist criteria for causation, in order for some property X to qualify as *causally distinct* from some property Z in relation to effect Y, X must exhibit a *distinct* level of invariance and hence *distinct* manipulability and causal relations in relation Y, compared with Z. (Recall the interventionist maxim introduced in Chapter 4, "No causal difference without a difference in manipulability relations, and no difference in manipulability relations without a causal difference." (Woodward, 2003: 61))

Going back to the discussion on RIDR, we can see that it is actually the realization independence of the supervenient relationships that mental properties stand in with their effects that makes this possible for mental properties. This is simply because it is the realization independence of those supervenient

relationships that makes it possible for mental properties to exhibit *distinct* levels of invariance and hence *distinct* manipulability and causal relations in relation to their effects in comparison to their physical realizers. As I explained above, depending on the nature of the dependency relationship, those properties may exhibit either more or less invariance than their physical realizers. However, it is important to be clear that in either case I take it that this varying degree of invariance is sufficient to distinguish the causal roles of those properties. Moreover, as the example of mental causation illustrates, nothing on this account of causal distinctness requires that the properties under consideration are *metaphysically* distinct.

Now, one implication of this account of causal distinctness, the relevance of which will become clear in the next chapter, is that it *does* matter, according to interventionism, how we 'pick out' the variables that are under consideration. This is because, to the extent that two properties enter into exactly the same invariant relationship with some effect and hence enter into exactly the same manipulability relations, it is appropriate, in interventionist terms, to consider them as the *same* cause.[12]

As an illustration, consider the following example of Woodward's (2008a: 239), which again refers to the research of Andersen et al: consider property A_1, which appeals to physically characterized facts about the aggregate pattern of the firing rates, which correspond to intention I_1. As Woodward explains, "…insofar as this aggregate profile A_1 corresponds to the different ways N_{11}, N_{12}, N_{13} of realizing I_1, and A_1 leads to R_1 and A_1 contrasts with whatever aggregate profile of neural activity A_2 corresponds to the different intention I_2, it

[12] I made this point briefly in Chapter 1, Section 1.1.2.1.

will be equally appropriate to cite A_1 as causing or figuring in the causal explanation for the monkey's exhibiting R_1." (Ibid) In other words, according to interventionism, A_1 also qualifies as a cause of R_1. However, as Woodward goes on to explain, "insofar as A_1 and I_1 enter into exactly the same manipulability or dependency relationships with respect to R_1, it is natural (from an interventionist point of view) to think of them as involving the same rather than competing causal claims with respect to R_1." (Ibid: 239-240) So, although A_1 also qualifies as a cause of R_1, since there is some intervention on A_1 that changes R_1, they are the *same* interventions that bring about the *same* changes to R_1 that are involved in the causal relationship between I_1 and R_1 and so it is appropriate to consider A_1 and I_1 as the same cause.[13] I return to this issue in the next chapter.

What about the thesis of supervenience? The account of mental causation that I have presented certainly seems consistent with the fact that mental properties supervene on physical properties with metaphysical necessity. In fact, as I have explained, the interventionist account of mental causation (which appeals to the notion of RIDR to explain mental causation) applies more generally to *all* supervenient causation and therefore explains how any supervenient property can qualify as a cause of some effect, in addition to its subvenient realizer.[14]

---

[13] Papineau (2013) argues that mental properties can qualify as causes of physical effects and can 'outcompete' their specific physical realizers for causal status (based on a proportionality requirement for causation), but only when those mental properties are type-identical (and hence reducible) to *some* physical property (for example, physical aggregate A_1). An in depth discussion of Papineau's argument is beyond the scope of this chapter, but it is important to recognise that for the purposes of defending non-reductive physicalism against Kim's exclusion problem, what matters is that we demonstrate that mental properties are not identical to their specific physical realizers. This is because it is *these specific* physical properties that feature in the exclusion argument and which Kim therefore takes to pre-empt and exclude mental properties.

[14] There are, however, some remaining issues regarding the thesis of supervenience, raised by Michael Baumgartner (2009, 2010), which will be addressed in the next chapter.

In Chapter 3 I outlined an argument that proved that overdetermination is not possible in the case of mental causation, given a supervenience relation between mental and physical properties. Nevertheless, it is worth noting that the interventionist account of mental causation is consistent with the thesis of non-overdetermination, since it provides an account of mental causation by which mental properties qualify as genuine causes of their effects, without being *metaphysically distinct*, sufficient productive causes of those effects, which would be required for genuine overdetermination to occur.

What about the thesis of causal closure, which states that every physical effect has a sufficient physical cause? I have already illustrated one way in which the interventionist account of mental causation does not violate causal closure, since I demonstrated that it provides an account of mental causation by which mental properties qualify as genuine causes of their effects, without exerting any additional force or energy into the physical domain to cause their effects. But, is it still true that every physical effect has a sufficient physical cause when causation is understood in interventionist terms?

In order to see why the answer to this question is 'yes', note that as I described it in Chapter 2, the thesis of causal closure entails that every physical effect is sufficiently determined by some prior physical state. I suggested that this follows from the acceptance of the conservation laws of physics, in addition to relatively recent discoveries in science. Now, this level of physical determinism essentially guarantees that every physical effect has an interventionist physical cause, given that it guarantees that intervening on a prior physical state will *always* be a way of intervening on a physical effect. (If every physical effect is sufficiently determined by some prior physical state, it will

always be possible to bring about a change to that physical effect by intervening on that prior physical state.) Put slightly differently, my claim is that the physical determinism that is implied by causal closure ensures that the relationships at the physical level display at least a minimal degree of invariance and hence ensures that every physical effect has a physical cause, even when 'cause' is understood in interventionist terms.[15] What I hope to have made clear by now is that according to interventionism, it is merely an empirical fact about the physical world that those physical causes are sufficient to produce their effects, rather than something that constitutes their causal status and I have argued that it is only when one makes this latter assumption that the exclusion problem becomes inevitable for the non-reductive physicalist.

In this section I have argued that not only does interventionism provide a coherent account of mental causation that avoids the exclusion problem, but it also provides a viable non-reductive *physicalist* account of mental causation and solution to the exclusion problem, since it upholds all of the minimal commitments of non-reductive physicalism.

[15] Remember that this is not true for the case of mental properties, as the example of 'fear' above illustrated. Although this is not an issue that Woodward discusses, the fact that physical causation is *guaranteed* under interventionism seems to give physical causation a kind of primacy over other kinds of causation. Now, this is not to undermine the fact that physical properties will often fail to provide *preferable* causal claims and explanations of their effects in comparison to mental properties; I have demonstrated that according to interventionism, there is nothing 'privileged' about physical causation in this sense. Nevertheless, it is an interesting feature of interventionism that fits well with our physicalist intuitions, that physical causation is guaranteed under interventionism (which, remember, is explained by the empirical facts about our world, rather than following from some a priori notion of the physical), while the same cannot be said for mental causation.

5.3.2 A *Satisfactory* Solution?

One final question that I will now address, which I have alluded to in previous chapters, is whether this interventionist account of mental causation provides a *satisfactory* account of mental causation and solution to the exclusion problem. This question arises, since, as I briefly discussed in Chapter 3, Kim (2010a) argues that attempts to overcome the exclusion problem that appeal to counterfactual theories of causation (of which interventionism is an example) fail to provide satisfactory accounts of the causal relevance of mental properties and hence fail to provide satisfactory solutions to the exclusion problem. According to Kim, this is because 'mere' counterfactual dependence cannot sustain the kinds of causal relationships that are involved in human agency, i.e. in the idea that human agents can perform physical actions, such as the movements of limbs and bring about physical effects, such as picking up the morning paper. According to Kim, what is required to sustain these kinds of causal relationships is the metaphysically richer notion of causation as production/generation. In fact, Kim goes as far as to claim that "without productive causation, which respects the locality/contiguity condition, such causal processes are not possible." (Ibid: 236)

Now, as I explained in Chapter 4, it is true that interventionism operates with a "metaphysically modest" (Woodward, 2003: 121) conception of causation, in comparison, for example, to the SP concept. For example, in order for X to cause Y, it is not necessary that X and Y are connected via any spatiotemporal process, or that X and Y exchange any conserved quantity, such as energy or momentum. Instead, it is only necessary that X and Y are connected

via counterfactual dependence (of the interventionist kind) and as I explained in Chapter 4, this only commits one to the idea "that there be facts of the matter, independent of facts about human abilities and psychology" (Ibid), namely facts about counterfactual dependencies.

What this means is that when we give an account of the causal relevance of a mental property in interventionist terms, it is not necessary that that mental property instantiates any kind of productive process in order to bring about its effect, or contributes any kind of energy or momentum to the production of that effect. Is Kim right to argue that this "metaphysically modest" (Ibid) conception of causation does not provide a satisfactory account of the causal relationships that are involved in human agency? In order to see why the "metaphysically modest" (Ibid) interventionist account of mental causation does provide a satisfactory account of mental causation and solution to the exclusion problem, firstly consider the argument that I made in Chapter 3.

As I explained in Chapter 3, the non-reductive physicalist who endorses an interventionist account of mental causation would not be committed to denying that the physical effects of mental causes are also caused by the subvenient physical realizers of those mental properties, which are sufficient to produce, or determine those effects (presumably via a continuous productive process of some kind), but in fact, given the non-reductive physicalist's commitment to causal closure and supervenience, she would be minimally committed to this idea. Thus, it is simply not true that when mental causation is understood in interventionist terms, there would be no physical effects produced as a result of human agency, since so long as the non-reductive physicalist is committed to causal closure and supervenience, the physical effects of mental

causes will continue to be produced, or determined by the subvenient physical realizers of those mental causes. The interventionist simply denies that this kind of sufficient production is identical to causation and instead claims that the causal relevance of both mental properties and their physical realizers can be understood in interventionist terms (i.e. in terms of the fact that there is an intervention on both the mental and the physical property that changes the effect in question).[16]

Moreover, I argued that as non-reductive *physicalists* we should not actually be surprised to discover that mental properties can only produce their effects, or be considered as sufficient causes of those effects, in virtue of the fact that they supervene on physical properties, since it was because of our commitment to causal closure (which implies that mental properties cannot exert any force or energy into the physical domain to produce or determine physical effects) and our commitment to the idea that the widespread overdetermination of physical effects by two metaphysically distinct sufficient causes would be implausible, that we accepted that the mental must supervene on the physical and hence that we should be physicalists in the first place.

What I hope this discussion therefore also demonstrates is that it is only by being "metaphysically modest" (Ibid) that interventionism is able to provide a

[16] In this sense, interventionism differs from the account of mental causation offered by Frank Jackson and Philip Pettit (1990a, 1990b). Jackson and Pettit make a distinction between so called 'causal relevance' and 'causal efficacy' (where the latter is thought to involve production/generation and the former is thought to involve something like counterfactual dependence) and argue that the causal role of mental properties can be understood in terms of relevance, rather than in terms of causal efficacy. However, by making a distinction between causal relevance and causal efficacy and by acknowledging that mental properties can only be considered to have causal relevance, rather than efficacy, this leaves them open to critique from Kim, who claims that 'full blown causal efficacy' is required to vindicate mental causation. By contrast, interventionists can avoid the charge from Kim that efficacy is required to vindicate mental causation, since the interventionist simply denies that there is a distinct concept of causation as efficacy and instead claims that the causal role of both mental and physical properties can be understood solely in interventionist terms.

viable non-reductive physicalist account of mental causation and solution to the exclusion problem, since it is only in this way that interventionism is able to uphold all of the minimal commitments of non-reductive physicalism. For example, I have demonstrated that it provides an account of mental causation by which *supervenient* mental properties can count as genuine causes of physical effects, in addition to their physical realizers. It respects the theses of *causal closure* and *non-overdetermination* by guaranteeing that mental properties cannot contribute to or interact with the sufficient physical causes of physical effects, or qualify as metaphysically distinct sufficient productive causes of those effects. Moreover, I demonstrated that this account also upholds causal closure in the sense that it remains true that every physical effect has a sufficient physical cause, even when causation is understood in interventionist terms. Lastly, I demonstrated that this account nonetheless upholds the theses of *non-identity* and *mental causation*, since it assigns genuinely distinct causal roles to mental properties, such as intentions, beliefs and desires.

Consequently, I suggest that if the non-reductive physicalist is looking for a metaphysically richer account of mental causation, involving the transfer of some conserved physical quantity, or spatiotemporally continuous physical process, for example, they will inevitably fail (since this would directly violate causal closure and non-overdetermination). However, what I hope the discussion in this thesis has shown is that although this account of mental causation may not be satisfactory for some (and although it is not the kind of causation that is usually discussed in the causal exclusion debate), it does nonetheless provide a satisfactory account of mental causation and solution to the exclusion problem

and is in fact the *only* viable account of mental causation and solution to the exclusion problem that we can give as serious *physicalists*.

Secondly, there are important practical reasons for thinking that this conception of mental causation is satisfactory. In order to see this, consider the following passage from Woodward:

> "Consider again a paralysed subject who is able to move a prosthetic limb (or a cursor on a screen) merely by thinking or by forming the right intention. Would most lay people and scientists think that this sort of "instrumental efficacy" is insufficient for true mental causation, with something metaphysically richer being required in addition? I suspect not. Certainly if we ask why we should care about whether there is mental causation, this looks very much like an issue about instrumental effectiveness: the concern is that we are deluded in our common sense belief that our intentions, desires, beliefs play a role in controlling our mental life and behavior, that we can change our behavior by changing these, that we can manipulate the mental states and behavior of others by changing other mental states of theirs and so on. This concern is adequately addressed by showing that mental states are causes in the sense captured by the interventionist account. We are thus left with the possibility that the only people who think that vindicating the claim that mental states are causes requires showing that they are causes in a richer, more metaphysical sense are certain philosophers of mind." (Woodward, 2008a: 248-9)

What this passage emphasises is that in interventionist terms, the mere formation of an intention by a paralysed subject would qualify as a full blown cause of the movement in the prosthetic limb (even though it is true that the production of this movement is due entirely to the fact that some physical property, namely the physical realizer of the intention on this occasion, is instantiated). Woodward is right to ask why we should insist on a metaphysically richer notion of mental causation, when this conception adequately captures the intuitive causal roles that we attribute to our mental states *and* captures the important practical implications of mental causation. (Remember that in Chapter 4 I argued that these important practical implications are in fact lost on the SP concept of causation.) When faced with such examples, I suggest that the onus is on Kim to explain why we should insist on a metaphysically richer notion of mental causation than interventionism offers.

5.4 Alternative Manipulationist Accounts of Mental Causation and the Problem of Realism

In this thesis I have argued that those features of interventionism that are somewhat subjective, for example, judgements of insensitivity and contrastive focus, do not introduce a problematic kind of subjectivity into interventionism and generate an anti-realist conception of causation because according to interventionism, whether X causes Y depends *solely* on whether there is counterfactual dependence of the right kind between the variables, (i.e. invariance under interventions) and I have argued that we have good reason to believe that this is an entirely objective matter.

Before moving on to discuss two alternative manipulationist accounts of causation, which I argue *do* generate anti-realist conceptions of causation, it is worth reminding ourselves why the issue of realism is so important. As I explained in the previous chapter, there is a strong and reasonable intuition that our concept of causation and our concept of mental causation should not be subjective and anti-realist in the sense that whether X causes Y depends on facts about us. For example, it should not depend on whether we find certain relationships useful for the purposes of control and manipulation, or depend on whether a certain claim captures the correct contrastive focus. Rather, we expect our concept of causation and our concept of mental causation to be realist in the sense that whether X causes Y *does not* depend on facts about us, but rather, that causation exists objectively 'out there' in reality, independently of us.

Moreover, despite my argument above that the "metaphysically modest" (Woodward, 2003: 121) account of mental causation that interventionism provides is satisfactory and despite my claim that it is in fact the *only* kind of account that we can give as serious non-reductive *physicalists*, if it turns out that interventionism is straightforwardly anti-realist, Kim would be justified in claiming that interventionism could not provide a satisfactory account of mental causation and solution to the exclusion problem. In other words, although I have argued that while the interventionist account of mental causation is "metaphysically modest" (Ibid) it does nonetheless provide a satisfactory account of mental causation and solution to the exclusion problem, the same could not be said if that "metaphysically modest" (Ibid) account turned out to be anti-realist.

While Woodward's interventionist account of causation and mental causation is not anti-realist (for the reasons that I have outlined thus far), I argue

that the same cannot be said for other accounts of mental causation that are also interventionist in spirit. In this section, I examine two alternative manipulationist accounts of mental causation, put forward by Christian List and Peter Menzies (2009) and John Campbell (2007, 2008a, 2008b, 2010). I argue that despite the individual merits and potential benefits of these theories, they each introduce a problematic kind of subjectivity into their theories and generate anti-realist conceptions of mental causation and so fail to provide satisfactory solutions to the exclusion problem. I demonstrate that the reason *why* these alternative accounts generate anti-realist conceptions of mental causation is because they each incorporate the notion of *contrastive focus*, or *proportionality*[17], which are somewhat subjective notions, into the necessary conditions for causation, while the same is not true for Woodward's account. More specifically, while Woodward appeals to the notion of proportionality to distinguish between *better or worse* causal claims and explanations, both List and Menzies and Campbell take proportionality to be a necessary condition for causation that distinguishes between *causal and non-causal* relationships, which inevitably generates an anti-realist concept of mental causation. I conclude that Woodward's interventionist account of mental causation therefore provides the *only* satisfactory non-reductive physicalist account of mental causation and solution to the exclusion problem.

[17] For the most part, I use the term 'proportionality' in this section, rather than the term 'contrastive focus', since this is the term that List and Menzies use in their paper and it is therefore more useful for the purposes of this argument. Although proportionality is the specific term that Yablo (1992) introduces, it is relevantly similar to Woodward's notion of contrastive focus and so it will be harmless to use these terms interchangeably in this section.

5.4.1 List and Menzies and the Problem of Realism

In a recent paper, List and Menzies (2009) argue that by adopting a 'difference making', or 'DM' approach to causation (which they claim is common to a number of theories of causation, including Woodward's interventionism), Kim's a priori exclusion problem turns out to be false.[18] Moreover, they argue that when the exclusion problem is reformulated in DM terms, it becomes a contingent, rather than a priori matter whether or not mental properties can have physical effects. An in depth discussion of List and Menzies' detailed argument is beyond the scope of this chapter, so I limit myself to discussing those features of their argument that are most relevant to my argument.[19] (I direct the reader to the footnotes for more specific details of List and Menzies' argument.)

What exactly does the DM conception of causation entail? List and Menzies offer the following account of the truth conditions for DM causation (List and Menzies appeal to a standard possible worlds analysis of counterfactuals)[20]:

> "The presence of F makes a difference to the presence of G in the actual world if and only if [it] is true in the actual world that (i) F is present □→ G is present; and (ii) F is absent □→ G is absent." (Ibid: 6)

[18] A similar argument is put forward by Raatikainen (2010) and by Menzies (2008).
[19] Shapiro (2011) also argues that List and Menzies' argument fails to provide a solution to Kim's exclusion problem, but argues instead that their argument fails specifically because of their use of the notion of 'realization-insensitivity'.
[20] This account is explained in detail in List and Menzies (2009: 6-8).

With the counterfactual truth conditions for DM causation outlined, List and Menzies argue that Kim's a priori exclusion problem (which they refer to specifically as the principle that "If a property F is causally sufficient for a property G, then no distinct property F* that supervenes on F causes G" (Ibid: 3)), turns out to be false. This is because they claim to prove that in DM terms, it *is* possible for some property F to be sufficient for some effect, G, and for a distinct property F*, that supervenes on F to cause G. In order to illustrate this, they appeal to Yablo's example of the trained pigeon and to the example of the research of Andersen et al. already mentioned above.

As a first illustration, note that in the Yablo example, the property of red qualifies as a DM cause of the pecking behaviour, given the truth of the two counterfactuals, 'target is red □→ pigeon pecks' and 'target is not red □→ pigeon does not peck', even though it supervenes on the property of scarlet, which is sufficient for the behaviour. By contrast, note that the property of scarlet does not qualify as a DM cause of the behaviour, given that one of the counterfactuals, namely, 'target is not scarlet □→ pigeon does not peck' is not true. According to List and Menzies, this can be explained in terms of a similarity relation between possible worlds, since presumably the closest world in which scarlet is not present is a world in which another shade of red is instantiated, in which case the pigeon still pecks.

Similarly for the example of Andersen et al, mental property I_1 qualifies as a DM cause of physical effect R_1[21], given the truth of the two counterfactuals 'monkey has intention I_1 □→ monkey performs R_1' and 'monkey does not have

[21] List and Menzies actually refer to this property as A_1, but for the sake of consistency, I refer to it as R_1.

intention I_1 □→ monkey does not perform R_1', even though it supervenes on neural property N_{11}, which is sufficient for R_1. By contrast, neural property N_{11} does not qualify as a DM cause of R_1, given that the negative counterfactual, 'monkey does not have neural property N_{11} □→ monkey does not perform R_1' is not true. Again, this is based on the assumption that the closest world in which N_{11} does not occur is one in which an alternative realizer of I_1 is instantiated, in which case R_1 still occurs. Thus, List and Menzies conclude that when causation is understood in DM terms, Kim's a priori exclusion principle turns out to be false.[22]

In the next stage of their argument, List and Menzies reformulate Kim's exclusion principle in DM terms, such that under certain conditions, no effect can have more than one DM cause. They also crucially extend this revised exclusion principle to incorporate both an 'upwards' and a 'downwards' formulation:

> "*Revised exclusion principle (upwards formulation)*: If a property F causes a property G, then no distinct property F* that supervenes on F causes G.
>
> *Revised exclusion principle (downwards formulation)*: If a property F causes a property G, then no distinct property F* that subvenes or realizes F causes G." (Ibid: 11)

[22] List and Menzies (Ibid: 10) acknowledge that Kim would not find this argument against his exclusion principle convincing, since he is concerned with vindicating mental causation as 'production', rather than in terms of counterfactual dependence. Although it cannot be discussed here, List and Menzies (Ibid) do put forward a convincing argument against Kim's objections and defend their DM approach to mental causation.

According to List and Menzies, which version of this revised exclusion principle applies, if any, is an entirely a posteriori, not a priori matter, since it depends solely on the details of the case at hand. This is because whether or not the exclusion principle applies (and whether it applies in its upwards or downwards formulation) depends on whether the two properties under consideration meet the requirements of DM causation (i.e. whether each of the properties meets the two counterfactual truth conditions outlined above). When each of the properties under consideration meet these requirements and hence qualify as DM causes of some effect, which List and Menzies label as cases meeting the 'compatibility requirement', the exclusion principle, on either formulation, will be false.[23]

According to List and Menzies, there are, however, a number of cases for which the exclusion principle will hold. These are cases in which one of the properties meets the requirements of DM causation, but in which case the other property fails to meet these requirements, resulting in either upwards, or downwards exclusion.[24]

---

[23] More precisely, they claim that this is possible when the following conditions are all met (where B represents some behavioural effect, M represents some mental property and N represents the physical realizer of M): "(i) B is present in all closest M-worlds; (ii) B is absent in all closest ~M- worlds; and (iii) B is absent in all closest ~N-worlds that are M-worlds." (Ibid: 12) As List and Menzies go on to explain in detail, this is possible when the relationships between supervenient properties and their effects are realization *sensitive*, in that small changes to how the supervenient property is realized leads to the absence of the effect. (Note that this guarantees that the negative counterfactual '~F □→ ~G' is met and that the subvenient property therefore also qualifies as a cause of G.)

[24] List and Menzies provide the following, more precise criteria for upwards and downwards exclusion. (Once again, B represents some behavioural effect, M represents some mental property and N represents the physical realizer of M): "*Necessary and sufficient conditions for upwards exclusion*: An instance of upwards exclusion occurs if and only if N is a difference-making cause of B and either (i) B is absent in some closest M-worlds that are ~N-worlds or (ii) B is present in some closest ~M-worlds outside the smallest ~N-permitting sphere." (Ibid: 13-14) and "*Necessary and sufficient conditions for downwards exclusion*: An instance of downwards exclusion occurs if and only if M is a difference-making cause of B and B is present in some closest ~N-worlds that are M-worlds." (Ibid: 15)

Upwards exclusion occurs when a subvenient property meets both of the counterfactual truth conditions for DM causation and hence qualifies as a DM cause of the effect, while the property that supervenes on it does not meet one of the counterfactual truth conditions and hence fails to qualify as a DM cause of the effect. For example, this occurs in the variant of the Yablo example in which Yablo's pigeon is trained to peck specifically at scarlet objects. This is because in this case, the property of scarlet now meets both of the counterfactual truth conditions and hence qualifies as a DM cause of the effect, while the property of red fails to meet one of the counterfactual truth conditions, namely, 'target is red $\Box\rightarrow$ pigeon pecks' and hence fails to qualify as a DM cause, supposedly resulting in a case of upwards exclusion.

By contrast, downwards exclusion occurs when a supervenient property meets both of the counterfactual truth conditions for DM causation and hence qualifies as a DM cause of the effect, while the subvenient property that realizes it does not meet one of the counterfactual truth conditions and hence fails to qualify as a DM cause of the effect. For example, this occurs in the case of the research of Andersen et al, since supervenient mental property I_1 meets both of the counterfactual truth conditions, while physical property N_{11} fails to meet the negative counterfactual, 'monkey does not have neural property N_{11} $\Box\rightarrow$ monkey does not perform R_1', and hence fails to qualify as a DM cause of the effect, supposedly resulting in a case of downwards exclusion.

Now, it is worth noting that according to List and Menzies, downwards exclusion is possible when the relationship between a supervenient property and its effect is realization *insensitive*, i.e. it holds under small changes to the

realization of the supervenient property.[25] Very roughly, this is because this insensitivity essentially guarantees that the supervenient property meets both of the counterfactual truth conditions, while the physical realizer of that supervenient property fails to meet the negative counterfactual truth condition. (This is because the insensitivity of the supervenient relationship ensures that the effect will still occur even if the physical realizer is changed to any one of the alternative realizers of the supervenient property.) In fact, because of multiple realization and insensitivity, it looks as though physical properties will often fail to meet the negative counterfactual truth condition for DM causation and hence fail to qualify as DM causes of the physical effects of the properties that supervene on them.[26]

List and Menzies conclude that by understanding causation in DM terms it is not only possible to prove that Kim's exclusion principle is false, but it is also possible to reformulate the exclusion principle in DM terms, such that exclusion becomes an entirely a posteriori matter that can actually support cases of mental causation, rather than providing a priori grounds for the causal exclusion of the mental.

[25] Although this notion of insensitivity appears to be similar to Woodward's notion of realization independence, this is not technically true. List and Menzies actually liken their notion of insensitivity to Woodward's notion of sensitive/insensitive causation. However, as I have explained, for Woodward, it is the fact that supervenient relationships are specifically realization independent *not* insensitive that guarantees that they qualify as causal, as this guarantees that there is at least a minimal degree of invariance at that level. For Woodward, the notion of insensitivity simply explains our causal judgements and helps to distinguish *between* better or worse causal claims and explanations, rather than distinguishing between causal and non-causal relationships.

[26] This formulation of interventionism (and Campbell's) therefore leads to the somewhat unintuitive conclusion that most, if not all, subvenient physical properties will fail to qualify as causes of the effects of the properties that supervene on them.

So, what exactly is wrong with List and Menzies' argument and why does it fail to provide a satisfactory account of mental causation and solution to the exclusion problem?

It is possible to diagnose the problem by looking more closely at the DM conditions for causation, since these conditions essentially rule out the possibility that a cause could fail to be *proportionate* to its effect.[27] For example, condition (i) rules out the possibility that causes could be too *general* for their effects, while condition (ii) rules out the possibility that causes could be too *specific* for their effects. This can be illustrated more clearly by appealing to the following example of List and Menzies':

> "Suppose, for example, there is a drug that causes patients to recover from an illness. The effect variable is a binary variable whose values are recovery or non-recovery. But the cause variable is a many-valued variable that can take the values 0mg, 50mg, 100mg, 150mg, and 200mg. Suppose that any regular dose at or above 150mg cures a patient, but any lower dose does not. Suppose a patient has taken a regular dose of 150mg and has recovered from the illness. What made the difference to the patient's recovery? According to the truth conditions above, the answer is "Giving the patient a dose of at least 150mg". It satisfies both conditions (i) and (ii): all relevantly similar patients who take a regular dose at or above 150mg recover and all those who take a lower dose don't. Other answers are either too specific, or not specific enough. For example, the cause cannot be "Giving the patient a dose above 50mg" because that

[27] This is something that List and Menzies (Ibid: 6) acknowledge.

> does not meet condition (i): some relevantly similar patients who are given a dose above 50mg, say 100mg, do not recover. Similarly, it cannot be "Giving the patient a dose of exactly 150mg" because that does not meet condition (ii): some relevantly similar patients who are not given a dose of exactly 150mg, say they are given 200mg, nonetheless recover. In this way, condition (i) rules out causes that are not specific enough to account for the change in the effect variable, while condition (ii) rules out causes that are too specific to account for it." (Ibid: 6)

Similarly, we can see that by being overly specific (i.e. by failing to be proportionate to the effect), physical property N_{11} fails to meet condition (ii), (the negative counterfactual truth condition) and hence fails to qualify as a proportionate DM cause of the effect. (As I explained above, this is based on the assumption that the closest world in which N_{11} does not occur is one in which an alternative realizer of I_1 is instantiated, in which case R_1 still occurs.) As I mentioned above, because of multiple realization and insensitivity, it looks as though physical properties will often fail to meet this negative counterfactual truth condition and hence will fail to qualify as DM causes of the physical effects of the properties that supervene on them. By contrast, by being proportionate to its effect, mental property I_1 meets both of these specific counterfactual truth conditions and hence qualifies as a proportionate DM cause of the effect.

Now, I argue that this version of interventionism, unlike Woodward's, does generate an anti-realist conception of mental causation. This is because whether X causes Y clearly depends on whether X is *proportionate* in relation to

Y[28] and given that the notion of proportionality is a *subjective* notion, for the reasons that I will outline below, this inevitably generates an anti-realist conception of mental causation.

Firstly, recall that in Chapter 4 I demonstrated that whether or not some property qualifies as a proportionate cause of some effect can vary depending on the context of the situation and most importantly for our present purposes, on the somewhat subjective consideration of our goal as enquirers. I illustrated this point in Chapter 4 with Woodward's example of the platform, but I suggest that this can also be illustrated by considering a variant of List and Menzies' drug trial example.

For example, suppose that in the actual circumstances the patient is given a dose of exactly 150mg and recovers from the illness in five days. Just as in the original example, suppose that Doctor A wants to know why the patient recovers, rather than does not recover. In this case, given the explanatory goal of Doctor A, List and Menzies are correct to state that the dose's being exactly 150mg does not qualify as a proportionate DM cause of the recovery, given that it is overly specific, while the dose's being at least 150mg does qualify as a proportionate DM cause of the patient's recovery. However, now consider doctor B who is also on the patient's medical team and who instead wants to know why the patient recovered specifically in five days, rather than in some other specific time frame (for example, in two days, three days, four days, seven days, etc.) In this case, given the explanatory goal of Doctor B, the dose's being exactly 150g *does* now qualify as a proportionate DM cause of the effect, while the dose's being at least 150mg no longer qualifies as a proportionate DM cause of the effect, given that it

[28] Woodward (2011a) also makes this point (see footnote 1).

is changes to whether the dose is specifically 150mg or some other specific dose that is associated with changes to the specific length of recovery.

So, we can see that whether or not some property qualifies as a proportionate cause of some effect can depend on the goal of the enquirer and given that, according to List and Menzies, whether X causes Y depends on whether X is proportionate in relation to Y, the DM conditions for causation inevitably become subjective and anti-realist.

Secondly, by appealing to a similarity metric in terms of the closeness of possible worlds, I suggest that List and Menzies' theory is also open to the same problems of subjectivity that I mentioned in Chapter 4 in relation to Woodward's notion of insensitivity. For example, I explained in Chapter 4 that judgements of closeness can depend on the context of the situation, on social custom, the expectations of the subject and so on. Given the potential subjectivity of considerations of closeness and given that these considerations are central to List and Menzies' account of causation, their account inevitably becomes subjective and anti-realist.

As an illustration, remember that when considering whether the counterfactual 'target is not scarlet □→ pigeon does not peck' is true, List and Menzies assumed that the closest world in which scarlet is not present is a world in which another shade of red is instantiated, in which case the pigeon still pecks and in which case the property of scarlet fails to qualify as a proportionate DM cause of the effect. However, why should we think that the closest possible world is one in which another shade of red is instantiated? This seems to depend on the idea that some kind of back-up mechanism would be in place, which would guarantee that another shade of red would be instantiated if scarlet were not

instantiated. However, this is not stipulated in the original example and in any case, seems a fairly far-fetched possibility.[29] Although one *could* argue that the closest possible world in which scarlet is not instantiated is one in which some alternative shade of red is instantiated, this question is at least open to subjective debate and this inevitably opens List and Menzies' theory up to the problem of subjectivity and anti-realism.

So, we can see that according to List and Menzies, whether X causes Y depends on whether X is proportionate in relation to Y and given that I have suggested that the notion of proportionality is a somewhat subjective notion, in the sense that it depends on the subjective considerations of our goal as enquirers and on the somewhat subjective considerations of closeness of possible worlds, their account inevitably becomes subjective and anti-realist. Given the strong and reasonable intuition that our account of mental causation should *not* be anti-realist and given that I have suggested that any satisfactory response to Kim's exclusion problem will have to avoid being anti-realist, I conclude that this account fails to provide a satisfactory account of mental causation and solution to the exclusion problem.

By contrast, although Woodward's theory incorporates the notion of proportionality, it does not likewise generate an anti-realist conception of causation. Remember that according to Woodward's account, X causes Y so long as there is at least *some* intervention on X that changes Y, even if there is no intervention on X that is associated with a proportionate change in Y. For example, since there is *some* intervention on physical property $N_{11,}$ namely an

[29] This is essentially the same point that Shapiro (2011) and Woodward (2011a, see footnote 1) make.

intervention that changes N_{11} to $N_{15,}$ that changes physical effect R_1, N_{11} qualifies as a bona fide cause of R_1, even though there is no intervention on N_{11} that is associated with a proportionate change in R_1.

Remember that for Woodward, the fact that N_{11} does not appear to be proportionate to its effect merely provides us with reasons to consider mental property I_1 as providing a *better* causal claim and explanation of the effect in comparison to N_{11}, rather than providing grounds for the downwards exclusion of that physical property. In other words, while Woodward appeals to the notion of proportionality to distinguish between *better or worse* causal claims and explanations, on List and Menzies' account, proportionality is built into the very definition of DM causation and hence becomes a feature that distinguishes between *causal and non-causal* relationships, inevitably introducing a problematic kind of subjectivity and anti-realism into their theory, whilst this possibility is ruled out on Woodward's account.[30]

Moreover, remember that unlike List and Menzies' account, the counterfactual truth conditions for Woodward's version of interventionism do not appeal to a similarity metric based on the closeness of possible worlds, but instead appeals to the technical notion of an intervention, which makes no reference to the notion of closeness of possible worlds. As I explained in Chapter 4, considerations of closeness do enter into Woodward's theory via the notion of *insensitivity,* but since these judgements do not determine whether X causes Y, I argued that this degree of subjectivity is not problematic and does not generate

[30] Yet another way of describing this difference is in terms of the idea that those features that Woodward identifies as useful from the point of view of *causal selection* (i.e. as useful for identifying which causes strike us as most salient amongst various causes), List and Menzies take to determine whether X causes Y. This difference is brought out clearly in Menzies (2011) and in Woodward (2011a).

an anti-realist conception of causation, while the same cannot be said for List and Menzies' account.

5.4.2 Campbell and the Problem of Realism

John Campbell (2008a) appeals directly to the theory of interventionism and specifically to the notion of a 'control variable', (which is explained below), in order to illustrate that it is possible to have physical effects without physical causes, thereby directly refuting the thesis of causal closure.[31] Campbell claims that this provides a solution to Kim's exclusion problem, since it proves that physical effects can have psychological causes without there *necessarily* existing competing physical causes of the same effects. I argue that despite the attractive consequences of this theory for the non-reductive physicalist, Campbell's theory also generates an anti-realist conception of mental causation and so fails to provide a satisfactory account of mental causation and solution to the exclusion problem. Once again, it is not possible to examine in detail the complex and insightful arguments that Campbell presents in his papers, so I focus only on those features that are most relevant to my argument.

[31]It is not clear that Campbell actually does 'refute' the thesis of causal closure, but rather, he appears to refute a particular formulation of the thesis that appeals to causal notions, i.e. he refutes the thesis that 'every physical effect has a sufficient physical *cause*'. Remember that as I described it in Chapter 2, causal closure entails that every physical effect is sufficiently determined by purely physical prior occurrences. As I explained, although it is possible to define causal closure in this specific way (and although it is possible to generate the physicalist conclusion of the Causal Argument on either formulation), it would be harmless (and useful for the purposes of my argument) to define causal closure as the thesis that 'every physical effect has a sufficient physical cause', since this turns out to be true when causation is understood in terms of *Woodward's* version of interventionism. In other words, I take it that Campbell does not refute the thesis of causal closure per se (since he certainly does not seem to deny that all physical effects are sufficiently determined by purely physical prior occurrences), but rather refutes the particular formulation of the thesis that appeals to causal notions, since he argues precisely that sufficient determination is not identical to causation and argues that when causation is understood in terms of his specific version of interventionism, it is not guaranteed that every physical effect has a physical cause. Menzies (2008), Raatikainen (2010) and Hitchcock (2012) put forward somewhat similar arguments against the thesis of causal closure; however, the points that I have made in response to Campbell here apply equally well to those arguments.

Campbell's specific formulation of interventionism centres on the notion of a 'control variable', which he introduces in order to elucidate the interventionist relationship between causes and their effects. As Campbell writes,

> "The idea is that when we are trying to find 'the right level' at which to characterize the causal functioning of a complex system, what we are looking for is what you might think of as the 'control panel' for the system, with respect to the outcomes we are interested in." (Campbell, 2010: 1)

To illustrate the notion of a control variable, consider Campbell's (Ibid) example of the relation between the dials on a radio and the output: since it is possible to control the output of the radio in a stable and systematic way by intervening on the position of the dials, changing the dials qualifies as a control variable for the output and consequently qualifies as a cause of the effect. Conversely, since it is not possible to stably and systematically control the output of the radio by intervening on the level of the circuitry of the radio, the physical state of the circuitry (on which it is assumed that the varying positions of the dial supervenes) does not qualify as a control variable for the output and consequently fails to qualify as a cause of the effect. Put slightly differently, since there is a "systematic function" (Ibid: 6) between the various positions of the dials and the output, turning the dials qualifies as a control variable and cause of the output, while given that there is no "systematic function" (Ibid) between the various physical states of the circuitry and the output, the physical state of the circuitry fails to qualify as a control variable and cause of the output. Note that

Campbell does acknowledge (Ibid) that there would be *some* change to the output under an intervention on the physical state of the circuitry, for example, under one intervention the radio may be completely destroyed. However, since this function is not *systematic,* (i.e. since the changes to the physical state of the circuitry are not *systematically* associated with changes to the output), this physical variable fails to qualify as a control variable and hence cause of the effect.

What becomes clear is that according to Campbell, in order for X to cause Y it is not only necessary that *some* intervention on X changes Y (since it is true, for example, that there is some intervention on the state of the circuitry of the radio that changes the output), but those interventions on X must be associated with "large, specific and systematic" (Campbell, 2008a: 433) changes in Y, such that X acts as a control variable for Y. Given this understanding of causation, it becomes possible for physical effects to have mental causes (if those mental variables can be considered as control variables for those effects), without also having physical causes, (if there aren't any physical variables that meet the specific criteria set out by the notion of a control variable).

In order to illustrate this, Campbell (Ibid: 437-439) appeals to the following example: suppose there exists a Martian physicist who has complete physical knowledge of human beings, including complete knowledge of the physical laws governing the basic particles that constitute us. Despite this complete physical knowledge, the Martian physicist is unaware that human beings are sentient creatures. On one occasion the Martian physicist and student are considering the cause of the congregation of humans at a colloquium, every Friday at 11am. The Martian physicist might respond that this outcome is a

complete accident, since there is no physical process which can be identified as a control variable for the outcome, i.e. no physical process, interventions on which are associated with "large, specific and systematic" (Ibid: 433) effects on the outcome. The Martian physicist might consider constructing a 'gerrymandered' physical control variable for this outcome, consisting of the 'total microphysical state' of each individual, along with the total microphysical state of their environment, but as Campbell points out, as well as being highly complex and quite removed from the qualitative outcome space we are interested in explaining, interventions on this gerrymandered variable would also fail to have a systematic effect on the outcome, which is required by his definition of a control variable. Thus it appears that this physical outcome (congregation) has no physical cause, directly refuting the thesis of causal closure.

Furthermore, Campbell points out that although this effect does not have a physical cause, it does have a psychological control variable and cause, this being the place and time at which everyone agreed to meet. For Campbell, this example illustrates that psychological causation is possible without physical causation, apparently providing a solution to the exclusion problem, since it illustrates that psychological properties and their physical realizers do not necessarily stand in competition with one another.[32]

So what exactly is wrong with Campbell's argument and why does it fail to provide a satisfactory solution to the exclusion problem? As I explained above, according to Campbell's account, in order for X to cause Y, it is not only

[32] For Campbell, the question of whether there are physical control variables (and hence physical causation) becomes an entirely empirical matter to be determined on a case-by-case basis. However, since he claims to have proven that the thesis of casual closure is false, he argues that the discovery of such physical causes does not affect the causal status of psychological properties.

necessary that there is *some* intervention on X that changes Y, but those interventions must have "large, specific and systematic" (Ibid) effects on Y, such that X acts as a control variable for Y. This is what makes it possible for a physical effect to have a psychological cause, without also having a physical cause.

Why does Campbell impose the constraint that causes should act as control variables for their effects? In order to see why, note that the requirement that causes act as control variables for their effects just *is* the requirement that changes to the cause variable should be *proportionate* to changes to the effect variable. On a plausible reading, the notion of a control variable simply captures the idea that different values of the cause variable should be systematically, stably and *proportionately* associated with different values of the effect variable.

For example, take the case of the colloquium: since it is the specific time and place of the congregation that we are interested in explaining (e.g. why the congregation occurred at 11am, rather than, say, 12pm), in order for some physical variable, such as the total neurophysiological state of each individual, to qualify as a cause of this effect, it is necessary that interventions on the value of this physical variable are associated with specific, systematic and *proportionate* changes to the value of the effect variable (e.g. to the time of the congregation). Since there fails to exist some such physical property, interventions on which are associated with specific, systematic and proportionate changes to the effect, Campbell concludes that this physical effect has no physical cause.[33]

[33] This reasoning can lead to the equally unintuitive conclusion that it is possible for some physical effect to have *no* cause whatsoever. See Campbell's (2008a) example of the billiard table.

Now, just as with List and Menzies' account, I suggest that this version of interventionism also generates an anti-realist conception of mental causation, since whether X causes Y also depends on whether X is proportionate in relation to Y and as we have seen, whether X qualifies as a proportionate cause of Y can vary depending on the somewhat subjective consideration of our goal as enquirers.

I suggest that this can be illustrated by considering a variant of Campbell's colloquium example. For example, suppose that we are no longer interested in explaining why the colloquium occurs at 11am, rather than 12pm, but are instead interested more generally in why the colloquium occurs *at all*. Given this new explanatory goal, the total neurophysiological state of each individual *will* qualify as a proportionate cause of the effect, since interventions on this variable will now be stably, systematically and proportionately associated with changes to whether the colloquium occurs, or fails to occur.

Once again, we can see that whether some property qualifies as a proportionate cause of some effect is somewhat subjective and given that according to Campbell, whether X causes Y depends on whether X is proportionate in relation to Y, this theory inevitably becomes subjective and anti-realist. Given the strong and reasonable intuition that our account of mental causation should *not* be anti-realist and given that I have argued that any satisfactory response to Kim's exclusion problem will have to avoid being anti-realist, I conclude that this account also fails to provide a satisfactory account of mental causation and solution to the exclusion problem

By contrast, this problem simply does not arise for Woodward's account. This is because on Woodward's account, there is guaranteed to be some physical

property (for example, one relating to the total neurophysiological state of each individual) that qualifies as a cause of this physical effect, given that there will be *some* intervention on this physical property that is associated with *some* change to the effect (for example, the congregation may not occur at all under an intervention on this physical variable). In other words, the relationship between this physical property and the effect is guaranteed to be at least minimally invariant and hence causal.[34] This remains true even though this intervention will not be associated with "large, specific and systematic" (Ibid) changes to the effect variable (e.g. to the specific time of the congregation), i.e. it remains true even though the changes to this physical variable are not *proportionate* to the changes to the effect variable.

Remember that for Woodward, the fact that this physical cause would not be proportionate to its effect merely explains why this physical cause may be considered as *less preferable* in comparison to the psychological cause of this effect. In other words, just as in the case of List and Menzies, while Woodward appeals to the notion of proportionality to distinguish between *better or worse* causal claims and explanations, for Campbell, proportionality becomes a necessary condition for causation, (as evidenced by his notion of a control variable), that distinguishes between *causal and non-causal* relationships,

[34] Campbell (2008c) actually refers to Woodward's notion of invariance when providing his account of mental causation, but it is evident that Campbell confuses this notion with his specific notion of a control variable. For example, Campbell suggests (Ibid: 188) that if there exists some physical effect, for which there is no physical property, interventions on which are associated with "large, specific and systematic" (Campbell, 2008a: 433) changes to the effect, then there will fail to exist an even minimally invariant and hence causal relationship at the physical level. However, remember that for Woodward, so long as there is *some* intervention on some physical property that changes the effect, that relationship will qualify as minimally invariant and causal, even if there is no physical property, interventions on which are associated with "large, specific and systematic" (Ibid) changes to the effect, i.e. even if there is no physical property that is a *proportionate* cause of that physical effect.

thereby generating an anti-realist conception of causation, whilst this possibility is ruled out on Woodward's account.

What about the worry that in formulating such minimal requirements for causation, Woodward's theory provides *problematically* weak requirements for mental causation? For example, one could argue that merely knowing that there is *some* intervention on a property that changes the effect does not tell us very much about that relationship, nor does it guarantee that that property will provide a satisfactory explanation of the effect, or provide an effective means of control over the effect. Whereas, since on both List and Menzies' and Campbell's accounts, these features are built into the very definition of what it is for X to cause Y, it is guaranteed that causes will also provide 'good' explanations of their effects and potentially provide an effective means of control over those effects.

Now, it is important to make clear exactly why Woodward's account does not generate *problematically* weak conditions for mental causation. This is because Woodward's version of interventionism does successfully distinguish between genuinely causal and non-causal relationships (in terms of the idea of a minimal degree of invariance), whilst also appealing to the notion of proportionality to distinguish between *better or worse* causal claims and explanations. The difference between Woodward's theory and each of the theories discussed in this section is simply that Woodward does not take this *further* consideration to be a necessary condition for causation.

Nevertheless, what I hope to have shown is that in so far as Woodward's account does generate relatively minimal requirements for causation, this is not a problem for this account, since it is precisely because Woodward only appeals to

the notion of invariance (which is an entirely objective notion, unlike the notion of proportionality) to distinguish between causal and non-causal relationships that his theory is able to avoid the problem of realism, while the same is not true for List and Menzies' and Campbell's accounts. I hope to have shown that Woodward's account of mental causation therefore provides the *only* satisfactory account of mental causation and solution to the exclusion problem.[35]

5.5 Conclusion

In this chapter I presented Woodward's interventionist account of mental causation and demonstrated how this account avoids the exclusion problem, whilst upholding all of the minimal commitments of non-reductive physicalism, thereby providing a successful non-reductive *physicalist* solution to the exclusion problem.

I began by demonstrating that interventionism not only provides an account of causation by which *both* mental and physical properties can qualify as causes of the same effect, but that when causation is understood in interventionist terms, mental properties can actually be considered as preferable causes of their effects, in comparison to their subvenient physical realizers (when, for example, those mental properties are highly RIDR and relatively invariant and provide the correct contrastive focus). Most importantly, I demonstrated that when causation is understood in interventionist terms, the question of mental causation becomes an entirely *a posteriori*, not *a priori* question.

[35] It is important to be clear that this argument has not been intended as an outright rejection of all of the ideas presented by List and Menzies and Campbell in their papers, since each of these theories upholds a broadly interventionist approach to causation and provides many insights into the exclusion problem.

I also made explicit how this account of mental causation avoids Kim's a priori exclusion problem and argued, contra Kim, that although this account is "metaphysically modest" (Woodward, 2003: 121), it does provide a satisfactory account of mental causation and solution to the exclusion problem. I also suggested that it is precisely *because* this account is "metaphysically modest" (Ibid) that it is able to uphold all of the minimal commitments of non-reductive physicalism and hence provide a viable non-reductive *physicalist* solution to the exclusion problem. Finally, I compared this account to two alternative manipulationist accounts of mental causation and argued that since they each generate anti-realist conceptions of mental causation, they fail to provide satisfactory accounts of mental causation and solutions to the exclusion problem. I concluded that Woodward's interventionist account of mental causation therefore provides the *only* satisfactory non-reductive physicalist account of mental causation and solution to the exclusion problem.

6. Interventionist Causal Exclusion and the Underdetermination Argument

6.1 Introduction

In a series of recent papers, Michael Baumgartner (2009, 2010) argues that far from securing the causal status of mental properties and providing a non-reductive physicalist solution to the exclusion problem, interventionism actually generates a new kind of exclusion problem, which apparently rests on weaker premises than the original Kimian formulation of the exclusion problem. Moreover, Baumgartner (2010) argues that the proposed interventionist solution to this novel interventionist exclusion problem leads to an 'underdetermination' of mental causation, making this supposed solution not fit for the purposes of the non-reductive physicalist.

In the first half of this chapter I outline and examine the debate between Baumgartner (2009) and Woodward (2011a). I demonstrate that although Woodward's solution involves modifying the definition of interventionism proposed in his (2003), (which I appealed to in Chapters 4 and 5 of this thesis), it does offer a genuine solution to Baumgartner's a priori interventionist exclusion argument. With this interventionist solution outlined, I will then, in the second half of this chapter, present my argument against Baumgartner's (2010) underdetermination argument. I demonstrate that by clarifying the metaphysical

implications of interventionist mental causation and by clarifying the conditions under which we can acquire empirical *evidence* for mental causation, the non-reductive physicalist who hopes to use interventionism as a solution to the exclusion problem can avoid Baumgartner's underdetermination argument. I will therefore conclude that the interventionist is able to defend her position against *both* of Baumgartner's objections and uphold the interventionist solution to the exclusion problem outlined in the previous chapter. In fact, I will demonstrate that this discussion actually provides *further* support for the "metaphysically modest" (Woodward, 2003: 121) account of mental causation that I outlined in Chapter 5.

The chapter is organised as follows: in Section 6.2, I outline and examine Baumgartner's (2009) interventionist exclusion argument and in Section 6.2.1, I outline and examine Woodward's (2011a) interventionist response to this argument. In Section 6.2.1.1, I outline Woodward's proposed modification of interventionism and in Section 6.2.1.2, I address some worries regarding this modification. In Section 6.3, I outline Baumgartner's underdetermination argument and argue that by clarifying the metaphysical implications of interventionist mental causation and by clarifying the conditions under which we can acquire empirical *evidence* for mental causation, the interventionist can avoid the underdetermination argument. Section 6.4 follows with some concluding remarks.

6.2 The Interventionist Exclusion Argument

Baumgartner (2009) argues that the 'interventionist exclusion argument' follows a priori from the very definition of interventionism proposed by

Woodward in his (2003). It will be useful to remind ourselves of these definitions:

> "(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. A necessary and sufficient condition for *X* to be a (type-level) *contributing* cause of *Y* with respect to variable set V is that (i) there be a directed path from *X* to *Y* such that each link in this path is a direct causal relationship … and that (ii) there be some intervention on *X* that will change *Y* when all other variables in V that are not on this path are fixed at some value." (Woodward 2003: 59, cited in Baumgartner, 2009: 163-164)

> "(IV) *I* is an intervention variable for *X* with respect to *Y* iff
>
> 1. *I* causes *X*;
>
> 2. *I* acts as a switch for all the other variables that cause *X*. That is, certain values of *I* are such that when *I* attains those values, *X* ceases to depend on the values of other variables that cause *X* and instead depends only on the value taken by *I*;
>
> 3. Any directed path from *I* to *Y* goes through *X*. That is, *I* does not directly cause *Y* and is not a cause of any causes of *Y* that are distinct from *X* except, of course, for those causes of *Y*, if any, that are built into the *I* – *X* – *Y* connection itself; that is, except for (a) any causes of *Y* that are effects of *X* (i.e., variables that are causally between *X* and *Y*) and (b)

any causes of *Y* that are between *I* and *X* and have no effect on *Y* independently of *X*.

4. *I* is (statistically) independent of any variable *Z* that causes *Y* and that is on a directed path that does not go through *X*." (Woodward, 2003: 98, cited in Baumgartner, 2009: 164)

Now, according to Woodward (2003), (and according to the account of interventionism examined in this thesis so far), (M) and (IV) provide *necessary* and *sufficient* conditions for X to cause Y: (M) spells out what it is for X to cause Y by appealing to the notion of a 'possible' intervention, while (IV) spells out the criteria that an intervention must meet if it is to be considered as suitable for assessing the causal role of X in relation to Y. How then does this formulation of interventionism generate the a priori interventionist exclusion problem?

According to Baumgartner, this problem arises for the interventionist because (M) and (IV) entail two necessary conditions for causation, which as we shall see, cause trouble when applied to cases of mental causation. Baumgartner labels these necessary conditions (MAN) and (FIX):

"(MAN) There possibly exists an intervention $I = z_i$ on *X* with respect to *Y*.

(FIX) The possible intervention $I = z_i$ is such that, while it is performed on *X*, all variables in the pertaining variable set V that are not located on a causal path from *X* to *Y* are held fixed, i.e. the variables in V that are not

> located on a causal path from *X* to *Y* can be held fixed while $I = z_i$ is performed on *X*." (Ibid: 167)

Now, (MAN) states, quite simply, that in order for X to cause Y there must possibly exist an intervention *I* on X with respect to Y. As I explained in Chapter 4 (Section 4.2.2), this notion of possibility should be understood in a fairly permissive sense, in that it is only necessary, according to Woodward, that interventions be constrained under logical, conceptual and metaphysical possibility, rather than, say, practical, physical or nomological possibility. With the notion of possibility understood in this fairly permissive sense, Baumgartner is right to claim that (MAN) is entailed by (M) and (IV) and is a necessary condition for interventionist causation.

Next, note that (FIX) essentially appeals to criterion IV-4 above, which states that intervention *I* must be independent of any variable Z which causes Y, that is on a directed path that does not go through X. It does therefore look as though (FIX) is also entailed by (M) and (IV) and is also a necessary condition for interventionist causation. We may therefore agree with Baumgartner that "If either (MAN) or (FIX) cannot be satisfied by two variables *X* and *Y* and the variable set V, *X* and *Y* are not causally connected relative to V according to (M)." (Ibid)

With these two necessary conditions for causation in mind, Baumgartner formulates his interventionist exclusion argument, which apparently a priori rules out the possibility of mental causation. In order to introduce this argument, let us consider the supposedly paradigmatic example of mental causation that I introduced earlier in this thesis: my desire for a cup of tea causes me to form the

intention to walk to the kitchen and switch on the kettle. Let M and M* represent these two mental phenomena and let P and P* represent the two physical phenomena that realize them. According to the non-reductive physicalist, M and M* supervene on P and P* without being identical to them, M causes M* and P*, and P causes P*.[1] I illustrate this in Figure 6.1 below.

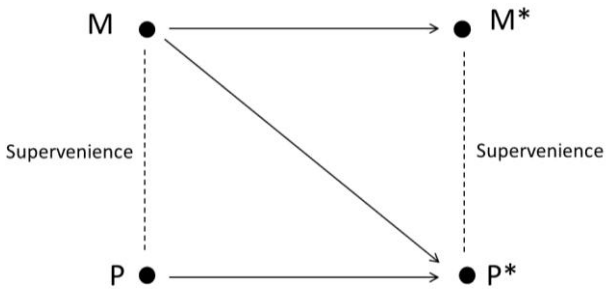

Figure 6.1: Supervenient Mental Causation

Solid arrows represent supposed causal relationships, while the broken lines depict supervenient relationships.

In relation to this set of variables (set V), Baumgartner formulates the interventionist exclusion argument as follows:

> "(1) *M* is causally relevant to *P** with respect to the variable set V = {*M*, *M**, *P*, *P**} iff there possibly exists an intervention $I_1 = z_1$ on *M* with respect to *P** such that all other variables in V that are not located on a

[1] It is worth noting that Baumgartner claims that his interventionist exclusion argument only excludes the causal relevance of M in relation to P*, but leaves it open as to whether M can cause M*. According to Baumgartner, this is because causation could never be transmitted through the P to M* route (given a supervenience relation between M* and P*) and without the idea that P is also a cause of M*, Baumgartner's exclusion argument would not go through. However, I agree with Woodward (2011a: 22) that causation can *sometimes* be transmitted through the P to M* route (if, for example, some intervention on the value of P changes the value of M*) and if that is correct, Baumgartner's exclusion argument would also rule out the possibility of causation between M and M*.

causal path from *M* to *P** are held fixed and the value or the probability distribution of *P** changes.

(2) *M* supervenes on MSB(*M*) = {$P = y_1$, $P = y_2$, … , $P = y_n$} without being identical to *P*.

(3) *P* is causally relevant to *P**.

[therefore] ¬(*M* is causally relevant to *P** with respect to the variable set V = {*M*, *M**, *P*, *P**}).” (Ibid: 169)

Let us examine this argument more closely. Although Woodward demonstrates that the crucial misstep in Baumgartner’s argument lies in his formulation of premise (1), according to (1), in order for M to cause P* there must possibly exist an intervention *I* on M such that all other variables in set V that are not located on a causal path that goes through M are held fixed and in which case P* changes. In other words, in order for M to cause P*, (MAN) and (FIX) must be satisfied in relation to all of the variables in set V.

According to premise (3), which is guaranteed by causal closure, P is causally relevant to P*.

Next, consider premise (2), which appeals to the theses of supervenience and non-identity. Now, Baumgartner correctly observes that any legitimate reading of supervenience has to maintain two things. Firstly, that supervenience is a non-causal relation and secondly, that any change at the supervenient level requires a change at the subvenient level (with disagreements concerning the modal force with which this is thought to hold). From these two minimal requirements, Baumgartner claims that the following holds:

"(2a) $M \neq P \wedge \neg$(*M* causes *P*) $\wedge \neg$(*P* causes *M*);

(2b) Every change in the values of *M* is necessarily accompanied by a change in the values of *P*.

(2a) and (3) imply:

(4) *P* is on a causal path to *P** that does not include *M*." (Ibid: 170)

Given the conjunction of these theses, the interventionist exclusion problem seems inevitable: assuming that P is an 'off-path' variable that does not go through M that causes P*, (MAN) and (FIX) require that P be held fixed while M is manipulated, however, this is clearly ruled out by the most minimal reading of supervenience. In other words, it turns out that because of supervenience, (MAN) cannot be satisfied in relation to set V, since supervenience guarantees that there fails to exist a possible intervention on M that meets even the most liberal reading of possibility. As Baumgartner explains,

"From the conjunction of (1a) and (4) it follows that, if *M* is causally relevant to *P**, there possibly exists a variable that causes changes in *M* while being statistically independent of changes in *P*. The latter, however, is excluded by (2b), which determines that the values of every variable that induces changes in *M* will necessarily be correlated with the values of *P*. Hence, there cannot possibly exist an intervention variable for *M* with respect to *P**. A straightforward application of modus tollens to (1a) then leads to the conclusion of the interventionist exclusion argument: ¬(*M* is causally relevant to *P** with respect to the variable set V = {*M*, *M**, *P*, *P**}) or *M* is causally irrelevant to *P**, for short. Put differently, *M*

> and *P** violate the first necessary condition for *M* to cause *P** according to reading (III) of (M), viz. (MAN)." (Ibid: 170-171)

We can see that (FIX) is also clearly violated for the same reason,

> "(4) states that *P* is located on a causal path to *P** that does not include *M*, which in virtue of (FIX) requires that *P* be fixed while *M* is manipulated. (2b), however, excludes just that fixability, i.e. (2b) excludes that *P* can possibly be held fixed while *M* is manipulated. Therefore, *M*, *P*, and V also violate (FIX)." (Ibid: 171)

Given that I accepted that (MAN) and (FIX) are necessary conditions for causation according to (M) and (IV), it does look as though supervenient causation (for example between M and M*, or between M and P*) is a priori ruled out by the very definition of causation outlined in (M) and (IV), since (MAN) and (FIX) *cannot* be satisfied when the variables under consideration stand in a supervenience relation. Moreover, Baumgartner claims that since this argument requires only a minimal reading of supervenience, it applies to *all* cases of supervenient causation, including *all* mental causation. For Baumgartner, this generates a novel, a priori interventionist exclusion problem.

Before moving on to discuss the solution to this problem, it is worth noting that according to Baumgartner, this exclusion argument rests on even weaker premises than the traditional Kimian formulation of the exclusion problem, apparently making it all the more decisive against the non-reductive physicalist who hopes to use interventionism to solve the exclusion problem.

Baumgartner notes that his exclusion argument differs from the traditional Kimian formulation in two crucial ways. Firstly, it does not involve a premise ruling out overdetermination and secondly, it does not presuppose that physical property P is sufficient for P*. As Baumgartner explains,

> "The mere causal relevance of *P* for *P** suffices that *P* would need to be fixable while *M* is manipulated in order for the latter to be a cause of *P** in the sense of (M). As we have seen above, such a fixing of *P* is impossible. In consequence, even though there may well exist countless systematically overdetermined effects and even though micro causes may not fully determine their micro effects, the currently most popular version of interventionism does not allow for any downward causal influence of supervening macro properties." (Ibid)

If correct, the conclusion of Baumgartner's interventionist exclusion argument would indeed have disastrous consequences for the interventionist solution to the exclusion problem that I outlined in the previous chapter. What then is the solution for the interventionist?

6.2.1 The Interventionist Solution to the Interventionist Exclusion Problem

In a recent paper, Woodward (2011a) defends the interventionist response to the traditional exclusion problem and responds directly to Baumgartner's interventionist exclusion argument. Woodward argues that Baumgartner's mistake is to assume that cases of causation involving what Woodward calls

'non-causal dependency relations', such as logical, conceptual and *supervenient* dependencies, should be treated in exactly the same way as cases of causation involving no non-causal dependencies. More specifically, Woodward argues that Baumgartner's argument relies on a mistaken assumption about what it is appropriate to control for, or hold fixed, when assessing causal systems that include supervenient dependencies and argues, contra Baumgartner, that it is *not* appropriate to control for the subvenient bases of supervenient properties when assessing the causal status of the latter. In other words, Woodward argues that (M) and (IV) (and (MAN) and (FIX)) cannot simply be applied to cases of causation involving supervenient dependencies.[2] Instead, Woodward proposes a modified version of (M) and (IV) that *can* be applied to causal systems that include supervenient dependencies and demonstrates that this formulation of interventionism does not lead to Baumgartner's interventionist exclusion argument.

Of course, the success of this solution will depend on how plausible one finds Woodward's argument that causal systems that include non-causal dependencies *should* be treated differently to causal systems that do not include any non-causal dependencies and that we should not require that subvenient bases be held fixed when assessing the causal relevance of supervenient properties. Luckily for the interventionist, Woodward does put forward a convincing argument in support of this. I outline and examine Woodward's argument in the remainder of this section.

[2] Woodward (2011a) acknowledges that this confusion surrounding whether (M) and (IV) do imply that the subvenient bases of supervenient properties must be controlled for might well be due to the fact that he simply does not make a distinction between the two kinds of causal systems that are discussed here in his (2003).

To begin, Woodward points out that the formulation of interventionism that he provides in his (2003), which appeals to (M) and (IV), is intended to apply to causal systems that include variables that all stand (or can potentially stand) in causal relationships with one another. In other words, it is presumed that the causal systems do not include any variables that stand in non-causal dependency relationships.[3] As Woodward explains,

> "…it is generally assumed that the variables occurring in a graph may be causally related or not or correlated or not, but that such variables are *not* connected by relationships of non-causal dependency (such as logical, conceptual, or mathematical relationships or supervenience relationships) of a sort that are inconsistent with their standing in causal relationships…In other words, it is assumed that we are dealing with variables [that] are "distinct" in a way that allows them to be potential candidates for relata in causal relationships." (Woodward, 2011a: 6)

Woodward formulates the principle of 'independent fixability' (or IF), which essentially captures this requirement:

> "(**IF**): a set of variables **V** satisfies independent fixability of values if and only if for each value it is possible for a variable to take individually, it is possible (that is, "possible" in terms of their assumed definitional, logical, mathematical, or mereological relations or "metaphysically possible") to

[3] In fact, Woodward adds that it is standard practice in causal theory to assume that the causal graphs and systems that are examined do not involve any non-causal dependencies.

> set the variable to that value via an intervention, concurrently with each of the other variables in **V** also being set to any of its individually possible values by independent interventions." (Ibid: 11-12)[4]

Now, on the assumption that the system under consideration includes no non-causal dependency relations (i.e. assuming that the system meets the preconditions spelled out in (IF)), in order to determine whether X causes Y, (M) and (IV) do state that there must exist some possible intervention on X that changes Y, where the notion of an intervention is defined in terms of the criteria outlined in (IV). In other words, when dealing with causal systems that include no non-causal dependencies, Baumgartner is correct that (MAN) and (FIX) do require that *all* of the variables in that set that also cause the effect, and which are not on a causal path that goes through the purported cause variable, must be held fixed while the cause variable is manipulated.

As an illustration, consider the following example[5]: suppose that we want to find out whether smoking, S, causes lung cancer, C. According to (IV), (which, remember, assumes that all of the variables in the system are independent of one another, with no non-causal dependency relations holding between them) this requires (amongst other things) that *I* should intervene on S independently of any other variable Z that also causes C, that does not go through S. For example, imagine that Z is a variable representing either the

[4] As Woodward (2011a: 13) points out, some writers, such as Brad Weslake (2011) have argued that we should in fact restrict the application of interventionism to causal systems that meet (IF). However, I agree with Woodward that although dealing with systems that do not meet (IF), for definitional, or metaphysical reasons, for example, does complicate matters, we can acquire genuine and novel causal knowledge from examining such systems and it is therefore justifiable to modify interventionism along the lines suggested by Woodward to deal with such systems.

[5] This is a variant of an example that Woodward (2011a: 7-8) uses to illustrate this point.

presence or absence of asbestos in the subject's environment. Assuming that Z is also a cause of C, the intervention on S must be independent of Z in order to rule out the possibility that Z could confound the relationship between S and C. If there does not exist such a possible intervention on S, then the causal relationship between S and C will be ruled out on a priori grounds.

Now, as Woodward explains, the 'crucial misstep' in Baumgartner's argument is to assume that a causal system that includes *both* causal and non-causal dependency relations, such as the one depicted in Figure 6.1 above, (which Woodward calls a 'mixed structure'), can be treated in the same way as a causal system that includes no non-causal dependencies, such as the one described in the example above. More specifically, he explains that Baumgartner's mistake is to assume that (M) and (IV) can be applied to causal systems that include non-causal dependencies and hence that we should hold fixed the subvenient bases of supervenient properties when considering the causal status of the latter, i.e. that (MAN) and (FIX) must be satisfied in relation to these causal systems.

So, why should we think that these two kinds of causal system *should* be treated differently and that it is *not*, in fact, appropriate to control for the subvenient bases of supervenient properties when assessing the causal status of the latter? Woodward provides the following arguments in support.

Firstly, Woodward argues that when one assumes that both of these kinds of causal system can be treated in exactly the same way and hence assumes that one must control for the subvenient bases of supervenient properties, it leads to mistaken causal inferences, suggesting that there is in fact an important disanalogy between the two cases.

In support of this argument, Woodward begins by appealing to an example which does not involve supervenience, but which involves another kind of non-causal dependency, this being 'definitional dependency'. Woodward then argues that the same argument applies to cases involving supervenient dependencies.

Consider Woodward's example:

> "Suppose…that heart disease (*D*) is causally influenced by high density cholesterol (*HDC*), which lowers the probability of disease and low density cholesterol (*LDC*) which raises the probability of disease. Suppose that we also have a variable representing total cholesterol (*TC*) which is *defined* as the arithmetic sum of *HDC* and *LDC* (i.e., *TC=HDC+LDC*). Assume for the sake of argument that we think of *TC* as also (causally) influencing *D*, although its overall impact on any given occasion of course will depend on the precise mix of *HDC* and *LDC* that taken together realize *TC*." (Ibid: 20)

This can be illustrated as follows:

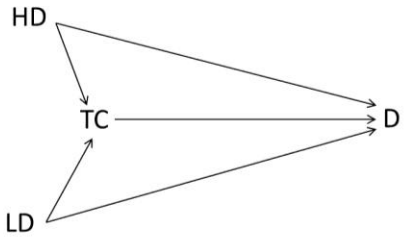

Figure 6.2: Cholesterol

Illustration copied from Woodward (2011a). The arrows from HD and LD to TC represent the definitional dependency of TC on HD and LD and the arrows from

HD, TC and LD to D represent the causal dependence of D on these variables. As Woodward notes, Figure 6.2 does not therefore make a distinction between causal and definitional relationships.

As Woodward explains, assuming, as Figure 6.2 does, that all of the variables in the system stand in causal, or potentially causal relationships with one another, in order to determine, for example, whether LD causes D, (M) and (IV) require that we consider the outcome of an intervention on LD that holds both HD and TC fixed, since both of these variables cause D and are on a causal path that does not go through LD. However, given the definitional relationship between HD, LD and TC, such interventions would be impossible. Arguing along Baumgartner's lines, it would seem that LD is a priori ruled out as a potential cause of D, given that the definitional relationships between the variables make IV-interventions impossible relative to that set.

However, as Woodward points out, this conclusion seems plainly mistaken and highly counterintuitive. Woodward argues that what this example actually suggests is that by introducing non-causal dependencies, such as definitional dependencies into a causal system, that causal system becomes somewhat complex and suggests that we need to be careful about how we treat such cases. More specifically, he argues that it suggests that we should be cautious in applying (M) and (IV) to these causal systems. For Woodward, the fact that the standard reading of (M) and (IV) delivers the judgement that a causal relationship between LD and D can actually be ruled out on *a priori* grounds, when this is clearly an issue to be settled on empirical grounds, strongly supports this point.

Now, Woodward does seem right to conclude that causal systems that include variables that stand in definitional relationships should be treated differently to those causal systems that do not include any such variables and that it is not therefore appropriate to hold fixed all of the variables in that set at independent values (which will be impossible for definitional reasons), since this leads to mistaken causal inferences regarding that system.

Can the same be said for causal systems that include supervenient dependencies? Woodward argues that the same argument can be applied to the supervenient case. As an illustration, consider again set V: given that two of the variables in that set, namely P and P*, are the subvenient bases of the other two variables in the set, namely M and M*, it will be impossible to intervene on M and M* while holding P and P* fixed. Once again, arguing along Baumgartner's lines, it would seem that M is a priori ruled out as a potential cause of P*, given that the supervenient relationships between the variables make IV-interventions impossible relative to that set.

Woodward concludes that the fact that a causal relationship between M and P* is ruled out on *a priori* grounds suggests, just as in the case involving definitional dependencies, that causal systems that include supervenient dependencies *should* be treated differently to those causal systems that do not include such dependencies and that it is not therefore appropriate to control for the subvenient bases of supervenient properties.[6]

[6] Baumgartner (2013) argues that there is an important disanalogy between the cholesterol case and a case of mental causation, because in the latter case, the non-reductive physicalist claims that a mental property, such as intention I_1, has *distinct* causal powers from its subvenient base, whereas in the cholesterol case, it would be appropriate to consider TC's causal powers to be identical and hence reducible to the causal powers of HD and LD. Consequently, Baumgartner argues that it *is* still appropriate to hold fixed the subvenient bases of mental properties, while the

Although I do not think that this example decisively proves this point, I agree with Woodward that given that a causal relationship between M and P* is ruled out on *a priori* grounds, it at least *suggests* that causal systems that include supervenient dependencies should be treated differently to those causal systems that do not include such dependencies and that it is not therefore appropriate to control for the subvenient bases of supervenient properties.

Secondly, and perhaps most importantly, Woodward argues that the original motivation that we had for controlling for variables in causal systems that include no non-causal dependencies, does not transfer to cases that include variables that do stand in non-causal dependencies and that it is not therefore appropriate to control for the subvenient bases of supervenient properties when assessing the causal status of the latter.

As an illustration, consider again the example of smoking introduced above: when considering the causal relevance of smoking, S, in relation to lung cancer, C, the motivation for controlling for variable Z, (which represents the presence/absence of asbestos in the subject's environment), was to rule out the possibility that the correlation between S and C was not due to the effect of intervention *I* on S, but was due to the effect of variable Z, which is also a cause of Y. This is why (M) and (IV) require that intervention *I* should manipulate S while holding Z fixed.

Now, imagine if in place of variable Z, we introduce variable B into the causal system, which represents the biological process on which S supervenes.

---

same may not be true for the case of HD and LD. My argument in the second half of this chapter, against Baumgartner's underdetermination argument, will address this objection.

Woodward suggests[7] that it no longer seems appropriate to control for B as it was for Z, since by being the subvenient base of S, B is simply not the kind of variable that could stand in a potential causal relationship with S and hence it is simply not the kind of variable that could confound the relationship between S and C.

Put slightly differently, Woodward's suggestion is that it seems wrong to assume that the motivation that we had for controlling for variables, such as Z, transmits to variables that are the subvenient bases of the properties under consideration, since subvenient properties are not the kind of properties that can stand in causal relationships with the properties that supervene on them and hence they are not the kind of properties that could act as confounders in the ordinary sense to those supervenient properties. Although I will demonstrate, in the second half of this chapter, that the issue of potential confounding in the case of mental causation is somewhat complex, Woodward is nonetheless right to conclude that it is *not* appropriate to control for the subvenient bases of supervenient properties when considering the causal status of those supervenient properties, as it is in the case of causal systems that include no non-causal dependencies.[8]

In summary, what these two arguments both suggest is that there are important differences between causal systems that include non-causal dependencies and causal systems that do not include any non-causal dependencies and that it is *not* appropriate to control for the subvenient bases of

[7] Woodward (2011a: 37) does not explicitly appeal to the example of smoking to illustrate this point, but instead appeals to a general set of variables (*X, Y, Z*).

[8] The argument that I present against Baumgartner's underdetermination argument in the second half of this chapter provides further support for the claim that the subvenient bases of supervenient properties should not be treated as confounders in the ordinary sense.

supervenient properties when considering the causal status of the latter and hence wrong to conclude from the fact that such interventions are impossible, that those supervenient properties are thereby a priori excluded as causes.

6.2.1.1 Modifying (M) and (IV)

However, as Woodward himself points out, this does not as yet offer a positive proposal of how we *should* deal with such systems within an interventionist framework, nor does it prove that the interventionist can provide such an account without running into Baumgartner's interventionist exclusion argument. How then should we understand (M) and (IV) when applied to causal systems that include non-causal dependencies, specifically supervenient dependencies?

Woodward's simple suggestion is that we should modify the requirements of (M) and (IV) so that they only consider, as relevant for assessing the causal status of some property within such a system, those interventions that set the variables within that set to values that respect the non-causal dependencies that hold between the variables. For example, when considering the causal status of some property in a causal system that includes supervenient dependencies, Woodward suggests that we should *only* consider the outcome of interventions that are possible given the supervenient dependencies that hold between the variables, i.e. we should not consider as relevant those interventions that cannot be carried out for metaphysical reasons, given the supervenient relationships that hold between the variables. (Remember that the motivation for this was outlined above.)

This idea of a relevant intervention suggests how we should modify (IV) to deal with causal systems that include supervenient dependencies. As Woodward explains,

> "To be more explicit, when (non-causal) supervenience relationships are present, the characterization **IV** should be interpreted in such a way that in condition (I3) a directed path counts as "going from *I* to *Y* through *X*" even if *I* also changes (as it must) the supervenience base *SB(X)* of *X*, as well as the value of *X*. Similarly, the reference in (I4) to "any variable *Z*" should be interpreted as "any variable *Z* other than those in the supervenience base *SB(X)* of *X*". Put slightly differently, an intervention *I* on *X* with respect to *Y* will (a) fix the value of *SB(X)* in a way that respects the supervenience relationship between *X* and *SB(X*), and (b) the requirements in the definition (**IV**) are understood as applying only to those variables that are causally related to *X* and *Y* or are correlated with them but [not] to those variables that are related to *X* and *Y* as a result of supervenience relations or relations of definitional dependence. Call this characterization of interventions (**IV***) and an intervention meeting these conditions an *IV**-intervention*." (Ibid: 34)

In other words, we should not only understand the notion of an intervention in condition as IV-3 as allowing that the intervention will necessarily bring about a change to the subvenient base of the supervenient property being considered (which respects the supervenient relationship that holds between the variables), but we should also crucially reinterpret criterion

IV-4 in such a way that it makes those variables that are the subvenient bases of the supervenient properties under consideration, (or those properties that are on causal paths that go through the subvenient properties), exempt from being considered as relevant off-path variables that need to be held fixed. Again, the justification for this can be drawn from the arguments outlined above.

With all of this in mind, Woodward provides the following modification of (M) and (IV), which incorporates the idea of a 'relevant' intervention. (I refer directly to the formulation provided by Baumgartner (2010)[9]):

> (M*) "*X* is a cause of *Y* with respect to the variable set V iff there possibly exists an (IV*)-defined intervention $I_1 = z_1$ on *X* with respect to *Y* such that all other variables in V that are not located on a causal path from *X* to *Y* and that are not part of the supervenience base of *X* are held fixed and the value or the probability distribution of *Y* changes.
>
> (IV*) *I* is an intervention variable for *X* with respect to *Y* iff *I* satisfies (IV.1), (IV.2), (IV.3), and (IV.4*):
> (IV.4*) *I* is (statistically) independent of any variable *Z* such that *Z* is a cause of *Y*, *Z* is not located on a causal path from *X* to *Y*, and *Z* is not part of the supervenience base of *X*." (Baumgartner, 2010: 17)

Thus, (M*) and (IV*), unlike (M) and (IV) make clear exactly which kinds of interventions are relevant for assessing the causal status of variables within

[9] Baumgartner actually refers to these modified principles as M** and IV*, but for the sake of continuity, I refer to them as (M*) and (IV*).

causal systems that include supervenient dependencies and also make explicit exactly which variables it is appropriate to control for within such systems.

Before I demonstrate how this revised formulation helps the interventionist to avoid Baumgartner's exclusion argument, it is important to emphasise that one direct consequence of this understanding of an intervention, (which I noted above), is that any IV*-intervention that changes a supervenient property will *automatically* cause a change in the subvenient base of that property (this follows given that any reading of supervenience requires that any change at the supervenient level requires a change at the subvenient level). In other words, IV*-interventions on mental properties are always *common causes*[10] of their subvenient physical realizers. Thus an intervention that respects the supervenient relationship between, for example, mental property I_1 and physical realizer N_{11}, will respect the requirement that an intervention that changes the value of the intention (for example, from I_1 to I_2) must also change the value of the neural realizer of that intention (for example, from N_{11} to N_{14}, or whatever physical property realizes I_2 on this occasion). I will return to this issue in Section 6.3 below, but what this in effect means is that when an intervention changes the value of some supervenient property, for example, from I_1 to I_2 and changes physical effect variable from R_1 to R_2 and therefore establishes that I_1 is a cause of R_1, the *same* intervention will establish that N_{11} is also a cause of R_1. This follows since the intervention on I_1, (which changes the value of the intention from I_1 to I_2), necessarily changes the value of the physical realizer, from N_{11} to N_{14} (or whatever physical property realizes I_2 on this occasion) and

[10] This was helpfully pointed out by Baumgartner in correspondence.

since this change in the value of N_{11} is also associated with a change in the value of R_1, N_{11} also qualifies as a cause of R_1, under the *same* intervention.[11]

How then do (M*) and (IV*) help us to avoid Baumgartner's interventionist exclusion argument? Consider again set V, which contained mental properties M and M* and their subvenient realizers, P and P*: suppose we want to find out whether M causes P*. Although it will be impossible to intervene on M independently of P, (which is also a cause of P* and is on a causal path that does not go through M), (IV*) *does not* require that P be held fixed while intervening on M, since P is the subvenient base of M and is therefore exempt from being held fixed. Since both (IV*) and (M*) will be satisfied relative to this set, (since there will exist a possible intervention on M that meets the requirements of (IV*)), supervenient causation between M and M*, or between M and P* will not be a priori excluded on the grounds that it fails to meet the basic requirements of interventionism. Furthermore, what the discussion above should have shown is that this *is* the right way to interpret the minimal requirements of interventionism when dealing with causal systems that include supervenient dependencies and it is clear that when interventionism is understood in this way, Baumgartner's interventionist exclusion argument does not go through.

6.2.1.2 Some Further Worries

One immediate worry that arises, however, is whether this solution, which modifies the definition of interventionism proposed by Woodward (2003),

[11] I address an issue concerning whether these properties can still be considered as *causally* distinct in Section 6.3 below.

implies that Woodward's (2003) account of interventionism, which I appealed to in Chapters 4 and 5 of this thesis, is false.

In order to see why Woodward's original definition of interventionism is not falsified by this modification, remember firstly that according to Woodward, (M) and (IV) were intended to apply to causal systems that do not include any non-causal dependencies and they continue to provide necessary and sufficient conditions for causation when applied to such causal systems.

Secondly, it is important to note that in formulating (M*) and (IV*) Woodward really is just *modifying*, or *extending* (M) and (IV) to include additional clauses, so that they can be unproblematically applied to causal systems that include variables that stand in non-causal dependency relations. What the previous discussion should have shown is that the problem was that (M) and (IV) did not make clear exactly how we are to understand the notion of an intervention when applied to such causal systems and this left interventionism open to Baumgartner's exclusion argument. However, by modifying (M) and (IV) along the lines suggested by Woodward (2011a), it is clear that the interventionist can avoid Baumgartner's interventionist exclusion argument. Moreover, it should be clear that this proposed modification does not falsify, but rather *extends* the definition of interventionism outlined in Woodward (2003).

6.3 The Underdetermination Argument

A more serious worry regarding this solution, which I will address in the remainder of this chapter, is proposed by Baumgartner (2010). Baumgartner (2010) objects that this proposed solution to his exclusion argument fails to fit the purposes of the non-reductive physicalist, since this modified formulation of

interventionism apparently results in an 'underdetermination' of mental causation. In the remainder of this chapter, I argue that by making clearer the metaphysical implications of interventionist mental causation and by making clearer the conditions under which we can acquire empirical *evidence* for mental causation, the interventionist can avoid Baumgartner's underdetermination argument.

To begin, what exactly is Baumgartner's objection? Baumgartner presents his argument as follows:

> "Assume we perform an (IV*)-defined intervention on the mental property M_1 and assume furthermore that we find this intervention to be followed by a change in the value of P_2. Does this test result reveal that M_1 is a cause of P_2? Certainly not. For by (IV*)-manipulating M_1 we explicitly allowed for changes in P_1 which the non-reductive physicalist takes to be another cause of P_2. This other cause is not located on a path from M_1 to P_2 and, above all, is determined to be causally sufficient for P_2 by the causal closure of the physical. In consequence, our test result significantly underdetermines a causal inference. At least two structures can generate the result of our hypothetical test: either (i) the change in the value of P_2 is only caused by a change in the value of P_1 which necessarily accompanied our intervention on M_1 or (ii) the change in the value of P_2 is overdetermined by P_1 and M_1. Of course, this ambiguity does not only arise due to a misguided intervention in one particular experimental context, rather, (IV*)-defined interventions, in general, are not required to be independent of all other causes of an effect under

> investigation. Supervenience bases of macro variables may vary and thereby causally influence investigated effects at will when those macro variables are (IV*)-manipulated. Hence, all empirical data that result from (IV*)-interventions and that could stem from macro-to-micro causation might just as well stem from a structure that only features micro-to-micro causation. (IV*)-manipulations never induce an unambiguous inference to macro-to-micro causation. Or differently: to every causal structure S_1 that involves at least one macro-to-micro dependency in the sense of non-reductive physicalists there exists a causal structure S_2 that is only composed of micro-to-micro dependencies such that S_1 and S_2 generate the exact same (IV*)-manipulability relations, notwithstanding the fact that they differ in causal respects. That is, somebody who subscribes to (M**) and (IV*) and conceives of the relationship between macro and micro properties in terms of non-reductive supervenience renounces one of the core principles behind interventionism, viz. 'no causal difference without a difference in manipulability relations'." (Ibid: 18-19)

There is a lot going on in this passage, but we can summarise Baumgartner's argument as follows[12]: because IV*-interventions on mental properties are always common causes of the subvenient bases of those mental properties, which non-reductive physicalists also take to be causes of the effects of mental properties, one could never tell on the basis of an IV*-intervention

[12] In keeping with the rest of the discussion in this chapter, I refer to variables M, M*, P and P*, rather than variables M_1, M_2, P_1 and P_2.

whether a mental property, such as M, causes a physical effect, such as P*, since the same evidence that is produced by the IV*-intervention on M would apparently support either (1) that only P causes P*, or (2) that P* is overdetermined by both M and P. Since IV*-interventions provide no evidence for mental causation, the definition of interventionism outlined by (M*) and (IV*) *underdetermines* mental causation and is not therefore suitable for the purposes of the non-reductive physicalist, who hopes to use interventionism to refute Kim's exclusion problem. Moreover, (M*) and (IV*) apparently violate the interventionist maxim 'no causal difference without a difference in manipulability relations', since the interventionist claims that there is a causal difference between each of these causal scenarios, even though there would apparently be no difference in manipulability relations between them.

In order to respond to this argument, some clarification is firstly in order. Firstly, is it true that it would be impossible to tell on the basis of an IV*-intervention, whether M causes P*, or whether *only* P causes P* (i.e. would it be impossible to distinguish between a causal scenario in which M causes P* and causal scenario (1))? The answer, quite simply, is 'no': according to interventionism, if M is *not* a cause of P* and only P is a cause of P* then there would *not* exist an IV*- intervention on M that changes P*, while there would exist some intervention on P that changes P*. As I explained in Chapter 5, this happens when the relationship between the mental property and the effect is non-RIDR and hence non-invariant and non-causal. For example, this occurs in the example of the general psychological concept 'fear'. See Figure 5.2 from Chapter 5 below as an illustration.

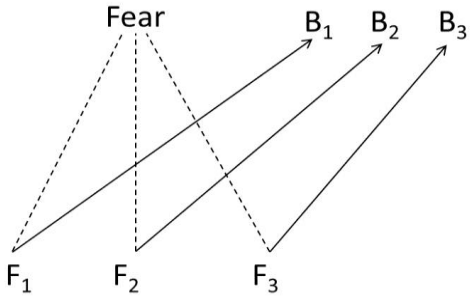

Figure 5.2: Non-RIDR/Fear

Solid arrows represent genuine causal relationships. Broken lines represent supervenient relationships. Fear systems F_1, F_2 and F_3 represent the different realizers of the supervenient concept 'fear'. B_1, B_2 and B_3 represent different behavioural effects.

As we can see, given that the relationship between the general psychological concept 'fear' and behavioural effect B_1 is not realization independent, there would *not* exist any IV*-intervention on 'fear' that changes B_1, while there would exist some intervention on physical property F_1 that changes B_1; hence *only* physical property F_1 will qualify as a cause of B_1.

On the other hand, if some mental property M *is* a cause of some physical effect P*, then according to interventionism, there *will* exist some IV*-intervention on M that changes P*. As I explained in the previous chapter, this happens when the relationship between M and P* is realization independent, since this ensures that the relationship between M and P* is at least minimally invariant and causal and ensures that the relationship between M and P* exhibits a distinct level of invariance in comparison to the relationship between P and P*

(more on this below). For example, this occurs in the example of the research of Andersen et al. See Figure 5.1 from Chapter 5 below as an illustration.

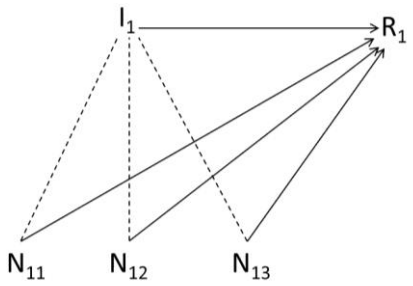

Figure 5.1: RIDR/Intention I_1

Solid arrows represent genuine causal relationships. Broken lines represent supervenient relationships.

As we can see, given that the relationship between mental property I_1 and physical effect R_1 is realization independent, there would exist some IV*-intervention on I_1 that changes R_1; hence I_1 will qualify as a bona fide cause of R_1.

So, it is simply not true that a causal scenario in which, for example, M causes P* would be indistinguishable from a causal scenario in which *only* P causes P* because according to the interventionist, in the latter case there would *not* exist any IV*-intervention on M that changes P*, while there would exist some intervention on P that changes P* and in the former case there *would* exist some IV*-intervention on M that changes P*. Or, to put this another way, it is not true that a causal scenario in which M causes P* would be indistinguishable from a causal scenario in which M is merely *epiphenomenal*, since according to the interventionist (and as demonstrated by the two examples above), there

would be a difference in manipulability and hence causal relations between the two cases.

Moreover, we can see that Baumgartner's worries about data confounding in this context are misguided, since it is simply not true that when some IV*-intervention on M supposedly establishes that M is a cause of P* that this could all be due to P's causal influence on P*, with M being merely epiphenomenal, because, once again, according to the interventionist, if this were the case, there would not exist any such IV*-intervention on M that changes P*.

What about causal scenario (2), in which P* is overdetermined by both M and P? Is it true that it would be impossible to tell on the basis of an IV*-intervention on M, whether M causes P*, or whether *both* M and P cause P*? As we shall see, this question does require more careful treatment, but I will argue that it does not lead, as Baumgartner suggests, to an underdetermination of mental causation.

Now, it is true that when an IV*-intervention on some mental property, such as intention I_1, brings about a change to physical effect R_1 and, ex hypothesi, establishes that I_1 is a cause of R_1, the same IV*-intervention will cause a change in N_{11}, the physical realizer of I_1, and hence will establish that N_{11} is *also* a cause of R_1. In other words, it is true that *any* IV*-intervention on a mental property that establishes that that mental property is a cause of some effect will also establish that the physical realizer of that mental property is a cause of the effect.

However, this does not lead to an underdetermination of mental causation, but merely reflects a fact that I have emphasised throughout this thesis, which is that mental properties and their physical realizers are not

metaphysically distinct causes of their effects. In other words, it merely reflects that the non-reductive physicalist is committed to the fact that whenever some mental property qualifies as a cause of some effect, it is guaranteed (by supervenience and causal closure) that the physical realizer of that mental property also qualifies as a cause of that effect (i.e. that mental causation *entails* physical causation). (Remember that this would not lead to a problematic form of overdetermination, given a supervenience relation between mental properties and their physical realizers.)

Moreover, the interventionist maxim would *not* be violated in these kinds of cases, since there is no difference in manipulability relations between a case in which, for example, M causes P* and a case in which both M and P cause P*, precisely *because* there is no causal difference between these two cases. In other words, I suggest that Baumgartner's initial question about whether it would be impossible to distinguish between these two causal scenarios on the basis of an IV*-intervention is simply misguided in the context of mental causation, given that according to the non-reductive physicalist, there is nothing *to* distinguish between these cases. To conclude from the fact that mental causation entails physical causation and from the fact that there is no empirical evidence that could distinguish between a case in which M causes P* and a case in which *both* M and P cause P* that mental causation is thereby underdetermined, is simply to misunderstand the commitments of non-reductive physicalism.

However, there is one potential problem with this response and this concerns whether what I have said actually undermines the argument that I made in Chapter 5, which was that mental properties and their physical realizers can be considered as *causally* distinct, i.e. as causes that cannot be identified, or reduced

(thereby upholding the non-reductive physicalist's commitment to the thesis of non-identity). This is because I have accepted that any IV*-intervention on a mental property will be a common cause of the physical realizer of that mental property and will establish that that physical property is also a cause of that effect. However, under *these* IV*-interventions, mental properties and their physical realizers will enter into exactly the same manipulability relations and as I have explained, according to interventionism, those properties will therefore be considered as the *same* cause.

In other words, I suggest that the problem that Baumgartner's underdetermination argument highlights isn't that IV*-interventions always establish that the physical realizers of mental properties also qualify as causes of their effects, (since I have argued that this is perfectly consistent with and is in fact guaranteed given the minimal commitments of non-reductive physicalism), nor is it that the evidence that would be produced when some IV*-intervention establishes that some mental property is a cause of some effect is the same evidence that would be produced if that mental property were merely epiphenomenal, (since I have argued that this is simply not true). Rather, the problem is that when some IV*-intervention establishes that some mental property is a cause of some physical effect, it *looks* as though that mental property is reducible to its physical realizer[13] (i.e. that IV*-interventions do not

[13] It is worth noting that the main target of Baumgartner's (2010) paper is in fact Shapiro and Sober's (2007) argument against epiphenomenalism. Shapiro and Sober argue somewhat similarly that it is wrong to hold fixed the subvenient bases of supervenient mental properties when assessing the causal status of the latter, but go on to argue that it is also wrong to assume that mental properties have causal powers in addition to those of their subvenient bases and wrong to conclude from the fact that mental properties do not have any such additional causal powers that they are thereby epiphenomenal. If what I have said in this chapter is right, Shapiro and Sober's argument would fail to provide a satisfactory *non-reductive* physicalist solution to the exclusion problem, since they accept that the causal powers of mental properties are identical

provide any *evidence* for mental causation). Baumgartner is right that in this *specific* sense (i.e. in the sense that IV*-interventions seem to support the conclusion that mental properties are *not* irreducible causes of their effects, when they in fact are), IV*-interventions would underdetermine mental causation and would not fit the purposes of the non-reductive physicalist who hopes to use interventionism precisely to avoid the threat of reduction.

How then can the interventionist avoid this problem? In order to *prove* that some mental property M and its physical realizer P are *causally* distinct (i.e. that M and P are causes that cannot be identified or reduced), I suggest that we not only consider whether some IV*-intervention on M (and P) changes P*, but *also* consider the outcome of an additional intervention on P, in order to determine whether there is a difference in manipulability (i.e. difference in degree of invariance) and hence causal relations between the M (and P) to P* relationship and the P to P* relationship.

As the discussion in the previous chapter should have made clear, whenever M *is* a cause of P* (which will occur when the relationship between M and P* is realization independent), this additional intervention on P *will* establish that the M (and P) to P* relationship displays a distinct level of invariance in comparison to the P to P* relationship and hence would demonstrate that M and P *are* genuinely causally distinct in interventionist terms. As I explained in Chapter 5, this is guaranteed given that it is the realization independence of the supervenient relationships that mental properties stand in with their effects that ensures that mental properties exhibit *distinct* levels of invariance and hence

---

and hence reducible to those of their physical realizers. This argument is echoed in Shapiro (2010, 2011).

distinct manipulability and causal relations in relation to their effects. (Remember also that I explained that depending on the nature of the realization independent dependency relationship between mental properties and their physical effects (i.e. depending on whether they are *highly* realization independent or possess only a *low* degree of realization independence), mental properties may exhibit either more or less invariance in relation to their effects in comparison to their physical realizers. However, in either case I argued that this varying degree of invariance is sufficient to distinguish the causal roles of those properties.)

In other words, whenever any mental property stands in some RIDR relationship to some physical effect (and hence qualifies as a cause of that effect), it is *guaranteed* that that mental property and its physical realizer qualify as causally distinct. What I have argued in this chapter is that Baumgartner's argument proves that the empirical *evidence* for such mental causation cannot, however, be acquired from single IV*-interventions on mental properties alone, but will be acquired from *both* the IV*-intervention on the mental property and the IV-intervention on the physical realizer of that mental property.

To elucidate these ideas further, we can appeal to the example of Andersen et al: in order to determine whether mental property I_1 is causally distinct from its physical realizer, N_{11}, I suggest that we not only consider whether there is some IV*-intervention on I_1 (and N_{11}) that changes physical effect R_1, but also consider the outcome of an additional intervention on N_{11}, in order to determine whether there is a difference in manipulability (i.e. difference in degree of invariance) and hence causal relations between the I_1 and R_1 relationship and the N_{11} and R_1 relationship. Since the relationship between I_1

and R_1 is realization independent, this additional intervention on N_{11} *will* establish that the I_1 (and N_{11}) to R_1 relationship displays a distinct (and in this case relatively high) level of invariance in comparison to the N_{11} to R_1 relationship and hence would demonstrate that I_1 and N_{11} *are* genuinely causally distinct in interventionist terms. (As Figure 5.1 above illustrates and as I explained in Chapter 5 (Section 5.3.1), the high level of realization independence of the relationship between I_1 and R_1 guarantees that there is a distinct (and in this case relatively high) level of invariance between I_1 and R_1, in comparison to the relationship between N_{11} and R_1, because it simply guarantees that there will be a wider range of interventions on I_1 that change R_1 than there are for physical property N_{11}.)

To summarise, although I accept that IV*-interventions on mental properties *alone* cannot distinguish between the causal roles of supervenient and subvenient properties, I suggest that the empirical evidence for mental causation can be acquired from *both* the IV*-intervention on the mental property and the IV-intervention on the physical realizer of that mental property. By considering the outcome of both of these interventions, interventionist mental causation would not be underdetermined by the definition of interventionism outlined in (M*) and (IV*).

Moreover, we can see that Woodward's argument about potential data confounding that I discussed in the previous section still stands in the sense that although, for example, mental property I_1 and its physical realizer N_{11} qualify as causally distinct (the evidence for which, I have suggested, is acquired from *both* interventions on I_1 (and N_{11}) and N_{11}), given that they are not *metaphysically* distinct and given that I have accepted that supervenient causation entails

physical causation (i.e. that whenever I_1 is a cause of R_1, so is N_{11}), it would be wrong to treat N_{11} as a potential confounder of I_1 in the ordinary sense. So long as the interventionist considers the outcome of *both* interventions on I_1 (and N_{11}) and N_{11}, there is no sense in which N_{11} could confound the relationship between I_1 and R_1.

What about the worry that this nonetheless undermines the argument that I made in Chapter 5 that it is by intervening directly at the mental level, for example, on intention I_1 that we often discover highly invariant and hence highly useful causal relationships, rather than by intervening directly at the physical level, for example on physical realizer N_{11}, given that under *any* IV*-intervention on I_1, I_1 and N_{11} (or whatever physical property realizes I_1 on some occasion) enter into exactly the same manipulability relations with respect to physical effect R_1?

In response, I would emphasise that my suggestion is that it is precisely I_1's, *not* N_{11}'s, distinct causal influence on R_1 that generates the distinct (and high) level of invariance under the common cause IV*-intervention on I_1 and N_{11} and which ensures that mental property I_1 qualifies as a preferable cause of R_1, in comparison to N_{11}. The fact that the additional intervention on N_{11} uncovers a relatively low invariant relationship between N_{11} and R_1 supports this hypothesis. Nothing about the account of mental causation that I outlined in Chapter 5 is undermined by the fact that when some IV*-intervention on I_1 establishes that I_1 is a cause of physical effect R_1 (and establishes that the relationship between I_1 and R_1 is highly invariant), the same IV*-intervention on I_1 establishes that N_{11} also qualifies as a cause of R_1, nor is it undermined by the fact that in order to

prove that I_1 and N_{11} are causally distinct, we must consider the outcome of an additional intervention on N_{11}.

Finally, it is worth emphasising that far from undermining the account of mental causation that I outlined in the previous chapter, this discussion actually provides further support for the "metaphysically modest" (Woodward, 2003: 121) account of mental causation that I outlined. This is because it emphasises that mental properties cannot cause their effects in some metaphysically rich sense, for example, via the transfer of some conserved physical quantity, (since then Baumgartner would be right to insist that it should be possible to intervene on mental properties independently of their subvenient bases)[14], but can only cause their effects in the "metaphysically modest" (Ibid) sense that they exhibit a distinct level of invariance in relation to their effects, in comparison to their physical realizers.

So, once again, in so far as the target of the exclusion problem, both Kim's and Baumgartner's, is some such metaphysically rich notion of mental causation, then both arguments prove that this is ruled out for the non-reductive physicalist. However, so long as the non-reductive physicalist is willing to accept this "metaphysically modest" (Ibid) account of mental causation and is willing to accept that the empirical *evidence* for mental causation cannot be acquired from IV*-interventions on mental properties alone, her position will not be undermined.

[14] Shapiro (2010) captures this point nicely in the following passage: "Thus, the idea that a supervening property might contribute causal force in addition to that which its base property possesses is at least untestable and, quite possibly, incoherent." (Ibid: 601-602)

6.4 Conclusion

In this chapter I examined two objections put forward by Michael Baumgartner against the interventionist account of mental causation and solution to the exclusion problem. I began by outlining the first objection put forward by Baumgartner (2009) and examined the interventionist response to this objection proposed by Woodward (2011a). I demonstrated that although Woodward's solution involved modifying the definition of interventionism that he proposes in his (2003), (which I appealed to in Chapters 4 and 5 of this thesis), it does offer a genuine solution to Baumgartner's a priori interventionist exclusion argument. I then argued that by clarifying the metaphysical implications of interventionist mental causation and by clarifying the conditions under which we can acquire empirical *evidence* for mental causation, the non-reductive physicalist who hopes to use interventionism as a solution to the exclusion problem can avoid Baumgartner's underdetermination argument. Moreover, I demonstrated that this discussion actually provides further support for the "metaphysically modest" (Ibid: 121) account of mental causation that I outlined in the previous chapter. It is therefore possible to conclude that the interventionist is able to defend her position against *both* of Baumgartner's objections and uphold the interventionist solution to the exclusion problem outlined in the previous chapter.

Nevertheless, this is not to undermine the significance of Baumgartner's arguments for the interventionist: Baumgartner's first objection highlighted that interventionists must modify the definition of interventionism outlined in (M) and (IV) in order to accommodate cases of supervenient causation. Baumgartner's second objection proved that the empirical evidence for mental

causation cannot be acquired from IV*-interventions on mental properties *alone*, but that in order to prove that mental properties and their physical realizers are causally distinct, we must consider the outcome of additional interventions on the physical realizers of those mental properties. However, I hope to have shown that so long as the interventionist is willing to make such adjustments, the interventionist solution to the exclusion problem that was outlined in the previous chapter will not be undermined.

7. Conclusion

7.1 Summary

In this thesis, I have argued that Woodward's (2003, 2008a, 2011a) version of interventionism not only provides an account of mental causation that avoids Kim's a priori exclusion problem, but also provides a genuine non-reductive *physicalist* solution to this problem, since it upholds all of the minimal commitments of non-reductive physicalism. In order to demonstrate this, I addressed a number of key issues and questions.

In Chapter 2, I began by demonstrating how Kim's a priori exclusion problem follows from five apparently inconsistent theses of non-reductive physicalism, namely mental causation, non-identity, supervenience, causal closure and non-overdetermination. I examined two of these theses in detail, namely causal closure and supervenience. I demonstrated that although the thesis of causal closure faces the problem of defining what it is to be physical and despite having had a complex history, causal closure is a true a posteriori thesis that does entail that every physical effect has a sufficient physical cause. I argued that this thesis provides the grounds for physicalism itself and concluded that it is therefore a minimal commitment of non-reductive physicalism that cannot be rejected in order to overcome the exclusion problem.

I then examined the thesis of supervenience in detail in order to determine exactly which formulation of supervenience the non-reductive physicalist is

minimally committed to and what its implications are. I argued that the non-reductive physicalist is minimally committed to a form of strong supervenience that holds with metaphysical necessity across all possible worlds, which implies that mental properties are entailed by and dependent on physical properties. After addressing some potential problems with this thesis, I argued that it is a minimal commitment of non-reductive physicalism that cannot be rejected in order to overcome the exclusion problem. I concluded that all five theses are in fact minimal commitments of non-reductive physicalism that cannot be rejected in order to overcome the exclusion problem and that they do appear to *a priori* lead to the exclusion problem.

In Chapter 3, I examined the assumptions that I take to underlie the exclusion problem. I argued that despite its apparent inevitability, the exclusion problem only follows a priori from these minimal commitments when they are combined with an assumption regarding causation, this being the assumption that causation is identical to sufficient production. I began by examining the SP concept of causation and demonstrated that Kim makes the assumption of SP. I then demonstrated how Kim's exclusion problem, as it is most commonly presented, depends crucially upon this assumption and that without it, the minimal commitments of non-reductive physicalism *do not* lead to the a priori exclusion of the mental. Finally, I demonstrated that even when Kim acknowledges that genuine overdetermination is not possible in the case of mental causation, he nonetheless generates the a priori exclusion problem because of the assumption of SP.

At this stage, I had yet to offer a solution to the exclusion problem. This is because if it turned out that the assumption of SP was in fact true, the non-

reductive physicalist would nevertheless be forced to accept the conclusion of the exclusion problem. In Chapter 4, I therefore outlined and examined Woodward's version of interventionism and presented an argument that undermined the assumption of SP. I began by outlining Woodward's version of interventionism and in particular, examined those features of the theory that would be especially relevant to my argument in Chapter 5, in which I presented the interventionist account of mental causation as a solution to the exclusion problem. Secondly, I highlighted some problems that the SP concept faces and presented interventionism as a viable alternative theory of causation that avoids these problems, undermining the assumption of SP and thereby demonstrating that the non-reductive physicalist need not accept Kim's a priori exclusion problem. I also addressed the worry that despite the problems that the SP concept faces, interventionism fails to provide a viable alternative to this theory and so fails to undermine the assumption of SP, since it faces serious problems of its own. I argued that not only can interventionism avoid these problems, but that it can actually deal with many of these problems in a more satisfying way than the SP concept. I concluded that interventionism *does*, after all, provide a viable alternative theory of causation to the SP concept and does undermine the assumption of SP, demonstrating that the non-reductive physicalist need not accept Kim's a priori exclusion problem. Lastly, I addressed some problems concerning the potentially anthropocentric, anti-realist and circular nature of interventionist causation, in order to demonstrate that interventionism can provide a coherent account of mental causation and satisfactory solution to the exclusion problem.

In Chapter 5, I outlined Woodward's interventionist account of mental causation and demonstrated that it provides an account of mental causation that not only avoids the exclusion problem, but also upholds all of the minimal commitments of non-reductive physicalism, thereby providing a successful non-reductive *physicalist* solution to the exclusion problem.

I began by demonstrating that interventionism not only provides an account of mental causation by which *both* mental and physical properties can qualify as causes of the same effect, but that when causation is understood in interventionist terms, mental properties can actually be considered as preferable causes of their effects, in comparison to their subvenient physical realizers (when, for example, they are highly realization independent and hence relatively invariant and provide the correct contrastive focus). Most importantly, I demonstrated that when causation is understood in interventionist terms, the question of mental causation becomes an entirely *a posteriori*, not *a priori* question.

I then made explicit how this account of mental causation avoids Kim's a priori exclusion problem and argued, contra Kim, that although this account is "metaphysically modest" (Woodward, 2003: 121), it does provide a satisfactory account of mental causation and solution to the exclusion problem. I also suggested that it is precisely *because* this account is "metaphysically modest" (Ibid) that it is able to uphold all of the minimal commitments of non-reductive physicalism and hence provide a viable non-reductive *physicalist* solution to the exclusion problem. Finally, I compared this account to two alternative manipulationist accounts of mental causation and argued that since they each generate anti-realist conceptions of mental causation, they fail to provide

satisfactory accounts of mental causation and solutions to the exclusion problem. I concluded that Woodward's interventionist account of mental causation therefore provides the *only* satisfactory non-reductive physicalist account of mental causation and solution to the exclusion problem.

Finally, in Chapter 6, I examined two objections put forward by Michael Baumgartner (2009, 2010) against the interventionist account of mental causation and solution to the exclusion problem. I began by providing an outline and analysis of Baumgartner's first objection and the response proposed by Woodward (2011a). I demonstrated that although Woodward's solution involves modifying the definition of interventionism that he proposes in his (2003), (which I appealed to in Chapters 4 and 5 of this thesis), it does offer a genuine solution to Baumgartner's a priori interventionist exclusion argument. I then argued that by clarifying the metaphysical implications of interventionist mental causation and by clarifying the conditions under which we can acquire empirical *evidence* for mental causation, the interventionist can avoid Baumgartner's underdetermination argument. In fact, I demonstrated that this discussion actually provides *further* support for the "metaphysically modest" (Ibid: 121) account of mental causation that I outlined in the previous chapter. I concluded that both of these objections can be overcome and that it is therefore possible to uphold the interventionist solution to the exclusion problem outlined in Chapter 5.

7.2 Implications for Mental Causation

I have argued that within an interventionist framework it is possible to provide an account of mental causation that not only avoids the exclusion

problem, but that also upholds all of the minimal commitments of non-reductive physicalism, thereby providing a viable non-reductive physicalist solution to the exclusion problem. What I also hope to have made clear is that it is precisely *because* this account is "metaphysically modest" (Woodward, 2003: 121) that it is able to uphold all of these minimal commitments and hence provide a viable non-reductive *physicalist* solution to the exclusion problem.

For example, I have demonstrated that this account of mental causation provides an account by which *supervenient* mental properties can count as genuine causes of physical effects, in addition to their physical realizers. I demonstrated that this account respects the theses of *causal closure* and *non-overdetermination* by guaranteeing that mental properties cannot contribute to or interact with the sufficient physical causes of physical effects, or qualify as metaphysically distinct sufficient productive causes of those effects. Moreover, I demonstrated that this account also upholds causal closure in the sense that it remains true that every physical effect has a sufficient physical cause, even when causation is understood in interventionist terms. Lastly, I demonstrated that this account nonetheless upholds the theses of *non-identity* and *mental causation*, since it assigns genuinely distinct causal roles to mental properties, such as intentions, beliefs and desires.

As I hope to have made clear, any metaphysically richer account of mental causation is simply ruled out given the minimal commitments of non-reductive physicalism. For example, as I made clear in Chapters 2 and 3, mental properties cannot be thought to exert any force or energy into the physical domain to produce or determine their effects, since this would directly violate causal closure. Moreover, since overdetermination is not possible given a

supervenience relation between the mental and the physical (and since this kind of overdetermination would be an implausible model for mental causation in any case), mental properties cannot be considered as metaphysically distinct sufficient productive causes of their effects. *A productive or generative conception of mental causation, as captured by the SP concept, for example, is therefore ruled out given the minimal commitments of non-reductive physicalism.*

While Kim took this fact to lead to the conclusion of the exclusion problem, I argued (in Chapter 3) that the exclusion problem only follows from this fact when it is combined with the assumption that causation is identical to sufficient production. Interestingly, what this discussion should therefore have made clear is that this limitation on mental causation (this being that mental properties cannot be thought of as metaphysically distinct sufficient productive causes of their effects) is not actually a result of Kim's a priori exclusion problem, but is in fact a result of the minimal commitments of non-reductive physicalism.

In fact, remember that I suggested that as non-reductive *physicalists* we should not actually be surprised to discover that mental properties cannot be thought to cause their effects in this productive, generative sense, but can only be considered to produce their effects, or be considered as sufficient causes of those effects, in virtue of the fact that they supervene on physical properties. This is because it was our commitment to causal closure (which implies that mental properties cannot exert any force or energy into the physical domain to produce or determine physical effects) and our commitment to the idea that the widespread overdetermination of physical effects by two metaphysically distinct, sufficient causes would be implausible, that we accepted that the mental must

supervene on the physical and hence that we should be physicalists in the first place (c.f. the Causal Argument from Chapter 2). This "metaphysically modest" (Ibid) account of mental causation may not be satisfactory for some, but I hope to have shown that it *does* nonetheless provide a satisfactory account of mental causation and solution to the exclusion problem and that it is in fact the *only* viable account of mental causation and solution to the exclusion problem that we can give as serious *physicalists*.

Bibliography

Anscombe, G.E.M. (1981) 'Causality and Determination', *Metaphysics and the Philosophy of Mind: Collected Philosophical Papers Volume II*, Oxford, Basil Blackwell Publisher.

Armstrong, D. (1968) *A Materialist Theory of the Mind*, London, Routledge & Kegan Paul.

Baker, L. R. (2003) 'Metaphysics and Mental Causation', *Mental Causation*, eds. Heil, J. Mele, A. Oxford, Oxford University Press, pp. 75-97.

Baker, L. R. (2009) 'Nonreductive Materialism', *The Oxford Handbook of Philosophy of Mind*, eds. McLaughlin, B. Beckermann, A. Walter, S. Oxford, Oxford University Press, pp. 109-128.

Baumgartner, M. (2009) 'Interventionist Causal Exclusion and Non-reductive Physicalism', *International Studies in the Philosophy of Science,* Vol. 23, Issue 2, pp. 161-178.

Baumgartner, M. (2010) 'Interventionism and Epiphenomenalism', *Canadian Journal of Philosophy*, Vol. 40, Issue 3, pp. 359-383.

Baumgartner, M. (2013) 'Rendering Interventionism and Non-Reductive Physicalism Compatible', *Dialectica,* Vol. 67, Issue 1, pp. 1-27.

Beebee, H. (2004) 'Causing and Nothingness', *Causation and Counterfactuals*, eds. Collins, J. Hall, N. Paul, L.A. Massachusetts, MIT Press, pp. 291-309.

Bennett, J. (2003) *A Philosophical Guide to Conditionals,* Oxford, Oxford University Press.

Bennett, K. (2003) 'Why The Exclusion Problem Seems Intractable, and How, Just Maybe, to Tract It', *Nous,* Vol. 37, Issue 3, pp. 471-497.

Block, N. (2003) 'Do Causal Powers Drain Away?', *Philosophy and Phenomenological Research,* Vol. 67, Issue 1, pp. 133-150.

Brewer, B. (1998) 'Levels of Explanation and the Individuation of Events: A Difficulty for the Token Identity Theory', *Acta Analytica,* Vol. 13, Issue 20, pp. 7-24.

Brewer, B. (2011) 'Realism and Explanation in Perception', *Perception, Causation and Objectivity*, eds. Roessler, J. Lerman, H. Eilan, N. Oxford, Oxford University Press, pp. 68-82.

Bromberger, S. (1966) ‘Why Questions’, *Mind and Cosmos*, ed. Colodney, R.G. Pittsburgh, University of Pittsburgh Press.

Bub, J. (Winter 2010) ‘Quantum Entanglement and Information’, *The Stanford Encyclopaedia of Philosophy*, ed. Zalta, E.N.

Burge, T. (2003) ‘Mind-Body Causation and Explanatory Practice’, *Mental Causation*, eds. Heil, J. Mele, A. Oxford, Oxford University Press, pp. 97-121.

Campbell, J. (2003) ‘Philosophy of Mind’, *Philosophy of Science Today*, Oxford, Oxford University Press, pp. 131-146.

Campbell, J. (2007) ‘An Interventionist Approach to Causation in Psychology’, *Causal Learning: Psychology, Philosophy and Computation*, eds. Gopnik, A. Schulz, L. Oxford, Oxford University Press, pp. 58-66.

Campbell, J. (2008a) ‘Interventionism, Control Variables and Causation in the Qualitative World’, *Philosophical Issues*, Vol. 18, Issue 1, pp. 426-445.

Campbell, J. (2008b) ‘Causation in Psychiatry’, *Philosophical Issues in Psychiatry: Explanation, Phenomenology, Nosology*, eds. Kendler, K. S. Parnas, J. Maryland, Johns Hopkins University Press, pp. 196-235.

Campbell, J. (2008c) ‘Comment: Psychological Causation Without Physical Causation’, *Philosophical Issues in Psychiatry: Explanation, Phenomenology, Nosology*, eds. Kendler, K. S. Parnas, J. Maryland, Johns Hopkins University Press, pp. 184-195.

Campbell, J. (2010) ‘Control Variables and Mental Causation’, *Proceedings of the Aristotelian Society*, Vol. 110, Issue 1, pp. 15-30.

Carroll, J. W. (Spring 2012) ‘Laws of Nature’, *The Stanford Encyclopaedia of Philosophy*, ed. Zalta, E.N.

Cartwright, N. (1979) ‘Causal Laws and Effective Strategies’, *Nous*, Vol. 13, Issue 4, pp. 419-437.

Cartwright, N. (1980) ‘Do the Laws of Physics state the Facts’, *Pacific Philosophical Quarterly*, 61, pp. 75–84.

Chalmers, D. (2003) ‘Consciousness and its Place in Nature’, *The Blackwell Guide to the Philosophy of Mind*, eds. Stich, S.P. Warfield, T.A. Oxford, Blackwell Publishing Ltd., pp. 102-143.

Chan, D. Ge, R. Gershony, O. Hesterberg, T. Lambert, D. (2010) ‘Evaluating Online Ad Campaigns in a Pipeline: Causal Models at Scale’, *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 7-16.

Child, W. (1994) *Causality, Interpretation, and the Mind*, Oxford, Oxford University Press.

Collins, J. Hall, N. Paul, L.A. (eds.) (2004) *Causation and Counterfactuals,* Massachusetts, MIT Press.

Crane, T. Mellor, D.H. (1990) 'There is No Question of Physicalism', *Mind 99,* pp. 185-206.

Crane, T. (1995) 'The Mental Causation Debate', *Proceedings of the Aristotelian Society,* Vol. 69, pp. 211-236.

Davidson, D. (1963) 'Actions, Reasons and Causes', *Journal of Philosophy*, Vol. 60, pp. 685-700.

Davidson, D. (1967) 'Causal Relations', *Journal of Philosophy,* Vol. 24, Issue 21, pp. 691-703.

Davidson, D. (2001) *Essays on Actions and Events,* Oxford, Oxford University Press.

Davidson, D. (2003) 'Thinking Causes', *Mental Causation*, eds. Heil, J. Mele, A. Oxford, Oxford University Press, pp. 3-19.

Dowe, P. (1999) 'The Conserved Quantity Theory of Causation and Chance Raising', *Philosophy of Science,* Vol. 66, pp. 486-501.

Feigl, H. (1958) 'The 'Mental' and the 'Physical'', *Minnesota Studies in the Philosophy of Science Vol. II*. eds. Feigl, H. Scriven, M. Maxwell, G. Minneapolis, University of Minnesota Press, pp. 370-497.

Fine, K. (1975) 'Critical notice: Counterfactuals', *Mind,* Vol. 84, pp. 451–458.

Fleming, S.M, Mars, R.B, Gladwin, T.E, Haggard, P. (2009) 'When the Brain Changes Its Mind: Flexibility of Action Selection in Instructed and Free Choices', *Cerebral Cortex,* October Issue.

Fodor, J. (1974) 'Special Sciences (or: the disunity of science as a working hypothesis)', *Synthese,* Vol. 28, Issue 2, pp. 97-115.

Gillet, C. Loewer, B. (eds.) (2001) *Physicalism and Its Discontents,* Cambridge, Cambridge University Press.

Glennan, S. (2010) 'Mechanisms, Causes, and the Layered Model of the World', *Philosophy and Phenomenological Research*, Vol. 81, Issue 2, pp. 362-381.

Godfrey-Smith, P. (2003) *Theory and Reality: An Introduction to the Philosophy of Science*, Chicago, University of Chicago Press.

Godfrey-Smith, P. (2007) 'Causal Pluralism', *Oxford Handbook of Causation*. eds. Beebee, H. Hitchcock, C. Menzies, P. Oxford, Oxford University Press, pp. 1-15.

Godfrey-Smith, P. (2008) 'Reduction in Real Life', *Being Reduced*, eds. Hohwy, J. Kallestrup, J. Oxford, Oxford University Press, pp.52-75.

Gopnik, A. Schulz, L. (eds.) (2007) *Causal Learning: Psychology, Philosophy and Computation*, Oxford, Oxford University Press.

Hall, N. Paul, L.A. (2003) 'Causation and Preemption', *Philosophy of Science Today*, pp. 100-130.

Hall, N. (2004) 'Two Concepts of Causation', *Causation and Counterfactuals*, eds. Collins, J. Hall, N. Paul, L.A. Massachusetts, MIT Press, pp. 225-276.

Hall, N. (Fall 2010) 'David Lewis's Metaphysics', *The Stanford Encyclopaedia of Philosophy*, ed. Zalta, E.N.

Haug, M. (2009) 'Two Kinds of Completeness and the Uses (and Abuses) of Exclusion Principles', *The Southern Journal of Philosophy*, Vol. 47, Issue 4, pp. 379-401.

Hausman, D. Woodward, J. (1999) 'Independence, Invariance and the Causal Markov Condition', *British Journal for the Philosophy of Science*, Vol. 50, Issue 4, pp. 521-583.

Heil, J. Mele, A. (eds.) (2003) *Mental Causation*, Oxford, Oxford University Press.

Hempel, C. (1969) 'Reduction: Ontological and Linguistic Facets', *Essays in Honour of Ernest Nagel*, eds. Morgenbesser, S. et al., New York, St Martin's Press.

Hitchcock, C. (1995) 'The Mishap at Reichenbach Fall: Singular vs. General Causation', *Philosophical Studies*, Vol. 78, Issue 3, pp. 257–291.

Hitchcock, C. (2012) 'Theories of Causation and the Causal Exclusion Argument', *Journal of Consciousness Studies*, Vol. 19, No. 5-6, pp. 40-56.

Hoerl, C. (2009) 'Causal Reasoning', *Philosophical Studies*, Vol. 152, Issue 2, pp. 167-179.

Hoerl, C. McCormack, S. Beck, S.R. (eds.) (2011) *Understanding Counterfactuals, Understanding Causation: Issues in Philosophy and Psychology*, Oxford, Oxford University Press.

Hohwy, J. Kallestrup, J. (eds.) (2008) *Being Reduced*, Oxford, Oxford University Press.

Horgan, T. Woodward, J. (1985) ‘Folk Psychology is Here to Stay’, *Philosophical Review,* Vol. 94, Issue 2, pp. 197-226.

Horgan, T. (1993) ‘From Supervenience to Superdupervenience: Meeting the Demands of a Material World’, *Mind,* Vol. 102, pp. 555-586.

Hornsby, J. (2003) ‘Agency and Causal Explanation’, *Mental Causation*, eds. Heil, J. Mele, A. Oxford, Oxford University Press, pp. 161-189.

Hornsby, J. (2004) ‘Agency and Actions’, *Agency and Action*, eds. Steward, H. Hyman, J. Cambridge, Cambridge University Press, pp.1–23.

Jackson, F. (1986) ‘What Mary Didn't Know’, *Journal of Philosophy,* Vol. 83, Issue 5, pp. 291–295.

Jackson, F. Pettit, P. (1988) ‘Functionalism and Broad Content’, *Mind,* Vol. 97, Issue 387, pp. 381-400.

Jackson, F. Pettit, P. (1990a) ‘Program Explanation: A General Perspective, *Analysis,* Vol. 50, Issue 2, pp. 107-117.

Jackson, F. Pettit, P. (1990b) ‘Causation in the Philosophy of Mind’, *Philosophy and Phenomenological Research,* Vol. 50, pp. 195-214.

Kallestrup, J. (2006) ‘The Causal Exclusion Argument’, *Philosophical Studies,* Vol. 131, Issue 2, pp. 459-485.

Kim, J. (1973) ‘Causation, Nomic Subsumption, and the Concept of Event’, *Journal of Philosophy,* Vol. 70, Issue 8, pp. 217-236.

Kim, J. (1982) ‘Psychophysical Supervenience’, *Philosophical Studie*s, Vol. 41, Issue 1, pp. 51–70.

Kim, J. (1984) ‘Concepts of Supervenience’, *Philosophy and Phenomenological Research,* Vol. 45, pp. 153-176.

Kim, J. (1985) ‘Supervenience, Determination, and Reduction’, *Journal of Philosophy,* Vol. 82, Issue 11, pp. 616–618.

Kim, J. (1987) ‘“Strong” and “Global” Supervenience Revisited’, *Philosophy and Phenomenological Research,* Vol. 48, No. 2, pp. 315-326.

Kim, J. (1989a) ‘Mechanism, Purpose, and Explanatory Exclusion’, *Philosophical Perspectives,* Vol. 3, pp. 77-108.

Kim, J. (1989b) ‘The Myth of Nonreductive Materialism’, *Proceedings and Addresses of the American Philosophical Association,* Vol. 63, No. 3, pp. 31-47.

Kim, J. (1992) ‘Multiple Realization and the Metaphysics of Reduction’, *Philosophy and Phenomenological Research*, Vol. 52, No. 1, pp. 1-26.

Kim, J. (1998a) *Mind in a Physical World*, Massachusetts, MIT Press.

Kim, J. (1998b) *Philosophy of Mind*, Colorado, Westview Press.

Kim, J. (2002) ‘Responses’, *Philosophy and Phenomenological Research*, Vol. 65, Issue 3, pp. 674-677.

Kim, J. (2003a) ‘Can Supervenience and Non-Strict Laws Save Anomalous Monism?’, *Mental Causation*, eds. Heil, J. Mele, A. Oxford, Oxford University Press, pp. 19-27.

Kim, J. (2003b) ‘The Non-Reductivist’s Troubles with Mental Causation’, *Mental Causation*, eds. Heil, J. Mele, A. Oxford, Oxford University Press, pp. 189-211.

Kim, J. (2005) *Physicalism, or Something Near Enough*, Princeton, Princeton University Press.

Kim, J. (2008) ‘Reduction and Reductive Explanation’, *Being Reduced*, eds. Hohwy, J. Kallestrup, J. Oxford, Oxford University Press, pp. 93-114.

Kim, J. (2009) ‘Mental Causation’, *The Oxford Handbook of Philosophy of Mind*, eds. McLaughlin, B. Beckermann, A. Walter, S. Oxford, Oxford University Press, pp. 29-53.

Kim, J. (2010) *Essays in the Metaphysics of Mind*, Oxford, Oxford University Press.

Kim, J. (2010a) ‘Causation and Mental Causation’, *Essays in the Metaphysics of Mind*, Oxford, Oxford University Press, pp. 243-263.

Kim, J. (2010b) ‘Explanatory Realism, Causal Realism, and Explanatory Exclusion’, *Essays in the Metaphysics of Mind*, Oxford, Oxford University Press, pp. 148-167.

Kim, J. (2010c) ‘Why There Are No Laws in the Special Sciences: Three Arguments’, *Essays in the Metaphysics of Mind*, Oxford, Oxford University Press, pp. 282-311.

Kripke. S. (1980) *Naming and Necessity*, Cambridge, Massachusetts, Harvard University Press.

Le Pore, E. Loewer, B. (1987) ‘Mind Matters’, *Journal of Philosophy*, Vol. 84, No. 11, pp. 630-642.

Lewis, D. (1966) ‘An Argument for the Identity Theory’, *Journal of Philosophy,* Vol. 63, Issue 2, pp. 17-25.

Lewis, D. (1973a) *Counterfactuals,* Oxford, Basil Blackwell.

Lewis, D. (1973b) ‘Causation’, *Journal of Philosophy,* Vol. 70, Issue 17, pp. 556-567.

Lewis, D. (1973c) ‘Counterfactuals and Comparative Possibility’, *Journal of Philosophical Logic,* Vol. 2, Issue 4, pp. 418-446.

Lewis, D. (1977a) ‘Counterfactual Dependence and Time’s Arrow’, *Nous,* Vol. 13, Issue 4, pp. 455-476.

Lewis, D. (1977b) ‘Possible-World Semantics for Counterfactual Logics: A Rejoinder’, *Journal of Philosophical Logic,* Vol. 6, pp. 359-363.

Lewis, D. (1986) *On the Plurality of Worlds,* Oxford, Basil Blackwell.

Lewis, D. (2000) ‘Causation as Influence’, *The Journal of Philosophy,* Vol.97, Issue 4, pp. 182-197.

List, C. Menzies, P. (2009) ‘Non-reductive Physicalism and the Limits of the Exclusion Principle’, *Journal of Philosophy,* Vol. 106, Issue 9, pp. 475-502.

Loewer, B. (2002) ‘Comments on Jaegwon Kim's Mind and the Physical World’, *Philosophy and Phenomenological Research,* Vol. 65, Issue 3, pp. 655-662.

Loewer, B. (2007) ‘Mental Causation or Something Near Enough’, *Contemporary Debates in the Philosophy of Mind,* eds. McLaughlin, B. Cohen, J. Oxford, Blackwell Publishing, pp. 243-265.

Loewer, B. (2008) ‘Why There Is Anything Except Physics’, *Being Reduced,* eds. Hohwy, J. Kallestrup, J. Oxford, Oxford University Press, pp.149-164.

Lowe, E.J. (2000) ‘Causal Closure Principles and Emergentism’, *Philosophy,* Vol. 75, Issue 4, pp. 571-585.

Mandel, D.R. Hilton, D.J. Catellani, P. (eds.) (2005) *The Psychology of Counterfactual Thinking,* Oxon, Routledge.

McDermott, M. (1995) ‘Redundant Causation’, *The British Journal for the Philosophy of Science,* Vol. 46, Issue 4, pp. 523–544.

McLaughlin, B. (1995) ‘Varieties of Supervenience’, *Supervenience: New Essays,* eds. Savellos, E. Yalcin, U. Cambridge, Cambridge University Press, pp. 16- 60.

McLaughlin, B. (2003) ‘On Davidson’s Response to the Charge of Epiphenomenalism’, *Mental Causation*, eds. Heil, J. Mele, A. Oxford, Oxford University Press, pp. 27-41.

McLaughlin, B. Bennett, K. (Summer 2010) ‘ Supervenience’, *The Stanford Encyclopaedia of Philosophy,* ed. Zalta, E.N.

Mellor, D.H. (1995) *The Facts of Causation*, London, Routledge.

Menzies, P. Price, H. (1993) ‘Causation as a Secondary Property’, *The British Journal for the Philosophy of Science,* 44, pp.187-203.

Menzies, P. (2008) ‘The Exclusion Problem, the Determination Relation, and Contrastive Causation’, *Being Reduced*, eds. Hohwy, J. Kallestrup, J. Oxford, Oxford University Press, pp.196-218.

Menzies, P. List, C. (2010) ‘The Causal Autonomy of the Special Sciences’, *Emergence in Mind,* eds. Macdonald, C. Macdonald, G. Oxford, Oxford University Press, pp.108-129.

Menzies, P. (2011) ‘The Role of Counterfactual Dependence in Causal Judgements’, *Understanding Counterfactuals, Understanding Causation: Issues in Philosophy and Psychology*, eds. Hoerl, C. McCormack, S. Beck, S. R. Oxford, Oxford University Press, pp. 186-208.

Montero, B. Papineau, D. (2005) ‘A Defence of the Via Negativa Argument for Physicalism’, *Analysis*, Vol. 65, Issue 287, pp. 233-237.

Mulliken, G.H. Musallam, S. Andersen, R.A. (2008) ‘Decoding Trajectories From Posterior Parietal Cortex Ensembles’, *Journal of Neuroscience,* Vol. 28, Issue 48, pp. 12913–12926.

Musallam, S. Corneil, B.D. Greger, B. Scherberger, H. Andersen, R.A. (2004) ‘Cognitive Control Signals for Neural Prosthetics’, *Science*, Vol. 350, Issue 5681, pp. 258-262.

Nagel, E. (1961) *The Structure of Science*, New York, Harcourt, Brace and World.

Nagel, T. (1974) ‘What is it Like to be a Bat?’, *Philosophical Review*, Vol. 83, Issue 4, pp. 435–450.

Papineau, D. (1990) ‘Why Supervenience?’, *Analysis,* Vol. 50, No. 2, pp. 66-71.

Papineau, D. (2001) ‘The Rise of Physicalism’, *Physicalism and Its Discontents*, eds. Gillet, C. Loewer, B. Cambridge, Cambridge University Press, pp. 3-37.

Papineau, D. (2004) *Thinking about Consciousness*, Oxford, Oxford University Press.

Papineau, D. (2008) 'Must a Physicalist be a Microphysicalist?', *Being Reduced,* eds. Hohwy, J. Kallestrup, J. Oxford, Oxford University Press, pp.126-149.

Papineau, D. (2009) 'The Causal Closure of the Physical and Naturalism', *The Oxford Handbook of Philosophy of Mind,* eds. McLaughlin, B. Beckermann, A. Walter, S. Oxford, Oxford University Press, pp. 53-65.

Papineau, D. (2013) 'Causation is Macroscopic but not Irreducible', *Mental Causation and Ontology,* eds. Gibb, S.C. Lowe, E.J. Ingthorsson, R.D. Oxford, Oxford University Press, pp. 26-153.

Pearl, J. (2009) 'Causality in the Social and Behavioral Sciences', [Online], Available: http://www.cs.ucla.edu/~kaoru/r355-long.pdf.

Place, U. T. (1956) 'Is Consciousness a Brain Process?', *British Journal of Psychology*, Vol.47, Issue 1, pp. 44–50.

Putnam, H. (1960) 'Minds and Machines', *Dimensions of Mind,* ed. Hook, S. New York, New York University Press.

Putnam, H. (1975a) 'The Nature of Mental States', *Mind, Language, and Reality: Philosophical Papers, Vol. 2,* Cambridge, Cambridge University Press, pp. 429-440.

Putnam, H. (1975b) 'The Meaning of "Meaning"', *Language, Mind and Logic; Minnesota Studies in Philosophy of Science,* Vol. 7, pp. 131-193.

Raatikainen, P. (2010) 'Causation, Exclusion, and the Special Sciences', *Erkenntnis,* Vol. 73, Issue 3, pp. 349-363.

Roessler, J. (2011) 'Perceptual Causality, Counterfactuals, and Special Causal Concepts', *Understanding Counterfactuals, Understanding Causation: Issues in Philosophy and Psychology,* eds. Hoerl, C. McCormack, S. Beck, S.R. Oxford, Oxford University Press, pp. 75-90.

Russell, B. (1912) 'On the Notion of Cause', *Proceedings of the Aristotelian Society,* Vol. 13, pp. 1-26.

Salmon, W. (1978) 'Why Ask 'Why?'? An Inquiry Concerning Scientific Explanation', *Proceedings and Addresses of the American Philosophical Association,* Vol. 51, No.6, pp. 683-705.

Salmon, W. (1984) *Scientific Explanation and the Causal Structure of the World,* Princeton, Princeton University Press.

Schaffer, J. (2003) 'Overdetermining Causes', *Philosophical Studies,* Vol. 114, Issue 1, pp. 23-45.

Scriven, M. (1962) ‘Explanations, Predictions, and Laws’, *Scientific Explanation, Space, and Time*, eds. Feigl, H, Maxwell, G. Minneapolis, University of Minnesota Press, pp. 170-230.

Shapiro, L. Sober, E. (2007) ‘Epiphenomenalism – the Do’s and the Don’ts’, *Thinking About Causes: From Greek Philosophy to Modern Physics*, eds. Wolters, G. Machamer, P., Pittsburgh, University of Pittsburgh Press.

Shapiro, L. (2010) ‘Lessons from Causal Exclusion’, *Philosophy and Phenomenological Research*, Vol. 81, Issue 3, pp. 594-604.

Shapiro, L. (2011) ‘Mental Manipulations and the Problem of Causal Exclusion’, *Australasian Journal of Philosophy,* Vol. 90, Issue 3, pp. 507-524.

Smart, J.J.C. (1959) ‘Sensations and Brain Processes’, *Philosophical Review*, Vol. 68, Issue 2, pp. 141–156.

Spurrett, D. Papineau, D. (1999) ‘A Note on the Completeness of ‘Physics’’, *Analysis*, Vol.59, pp. 25-29.

Statnikov, A. Henaff, M. Lytkin, N.I. Aliferis, C.F. (2012) ‘New Methods for Separating Causes from Effects in Genomics Data’, *BMC Genomics 2012,* Vol. 13, Supplement 8.

Steward, H. (1997a) *The Ontology of Mind*, Oxford, Oxford University Press.

Steward, H. (1997b) ‘On the Notion of Cause ‘Philosophically Speaking’’, *Proceedings of the Aristotelian Society,* Vol. 97, Issue 2, pp. 125–140.

Stoljar, D. (Fall 2009) ‘Physicalism’, *The Stanford Encyclopaedia of Philosophy,* ed. Zalta, E.N.

Stoljar, D. (2010) *Physicalism*, Oxford, Routledge.

Strawson, P. F. (1992) ‘Causation and Explanation’, *Analysis and Metaphysics: An Introduction to Philosophy*, Oxford, Oxford University Press, pp. 109-133.

Weatherson, B. (Summer 2010) ‘David Lewis’, *The Stanford Encyclopaedia of Philosophy*, ed. Zalta, E.N.

Welshon, R. (2002) ‘Emergence, Supervenience, and Realization’, *Philosophical Studies,* Vol.108, Issue 1, pp. 39-51.

Weslake, B. (2011) ‘Exclusion Excluded’, [Online], Available: http://bweslake.s3.amazonaws.com/research/papers/weslake_exclusion.pdf.

Williamson, T. (1998) ‘The Broadness of the Mental: Some Logical Considerations’, *Noûs*, Vol. 32, pp.389–410.

Williamson, T. (2000) *Knowledge and Its Limits*, Oxford, Oxford University Press.

Williamson, T. (2007) *The Philosophy of Philosophy*, Oxford, Blackwell Publishing.

Williamson, T. (2010) 'Reclaiming the Imagination', *The New York Times*, [Online], Available: http://opinionator.blogs.nytimes.com/category/the-stone/.

Woodward, J. (1984) 'A Theory of Singular Causal Explanation', *Erkenntnis*, Vol.21, Issue 1, pp. 231-262.

Woodward, J. (1992) 'Realism about Laws', *Erkenntnis*, Vol. 36, Issue 2, pp. 181-218.

Woodward, J. (2002) 'There is No Such Thing as a Ceteris Paribus Law', *Erkenntnis*, Vol. 57, Issue 3, pp. 303-328.

Woodward, J. (2003) *Making Things Happen*, Oxford, Oxford University Press.

Woodward, J. (2006) 'Sensitive and Insensitive Causation', *Philosophical Review*, Vol. 115, Issue 3, pp. 273-316.

Woodward, J. (2008a) 'Mental Causation and Neural Mechanisms', *Being Reduced*, eds. Hohwy, J. Kallestrup, J. Oxford, Oxford University Press, pp. 218-263.

Woodward, J. (2008b) 'Cause and Explanation in Psychiatry: An Interventionist Perspective', *Philosophical Issues in Psychiatry: Explanation, Phenomenology, Nosology*, eds. Kendler, K.S. Parnas, J. Maryland, Johns Hopkins University Press, pp. 132-195.

Woodward, J. (2010a) 'Causation in Biology: Stability, Specificity, and the Choice of Levels of Explanation', *Biology and Philosophy*, Vol. 25, Issue 3, pp. 287-318.

Woodward, J. (2010b) 'Scientific Explanation', *The Stanford Encyclopaedia of Philosophy*, ed. Zalta, E.N.

Woodward, J. (2011a) Preprint: 'Interventionism and Causal Exclusion', [Online], Available: http://philsci-archive.pitt.edu/id/eprint/8651.

Woodward, J. (2011b) 'Causal Perception and Causal Cognition', *Perception, Causation and Objectivity*, eds. Roessler, J. Lerman, H. Eilan, N. Oxford, Oxford University Press, pp. 229-264.

Woodward, J. (2011c) 'Psychological Studies of Causal and Counterfactual Reasoning', *Understanding Counterfactuals, Understanding Causation: Issues in*

Philosophy and Psychology, eds. Hoerl, C. McCormack, S. Beck, S. R. Oxford, Oxford University Press, pp. 17- 54.

Yablo, S. (1992) 'Mental Causation', *Philosophical Review*, Vol. 101, Issue 2, pp. 245-280.

Yablo, S. (1997) 'Wide Causation', *Philosophical Perspectives,* Issue 11, pp. 251-281.

Yablo, S. (2003) 'Causal Relevance', *Philosophical Issues,* Vol. 13, Issue 1, pp. 316-328.

Yablo, S. (2004) 'Advertisement for a Sketch of an Outline of a Prototheory of Causation', *Causation and Counterfactuals*, eds. Collins, J. Hall, N. Paul, L.A. Massachusetts, MIT Press, pp. 119-139.

Zhong, L. (2010) 'Can Counterfactuals Solve the Exclusion Problem?', *Philosophy and Phenomenological Research*, Vol. 83, Issue 1, pp. 129-147.

List of Abbreviations

(WS) weak supervenience

(SS) strong supervenience

(SS*mn*) strong supervenience that holds with metaphysical necessity across all possible worlds

(SP) the sufficient production concept of causation

(EP) the exclusion principle

(M) designates Woodward's definition of the interventionist, or manipulationist concept of causation

(IV) designates the criteria for a suitable intervention

(AC) actual causation

(RIDR) realization independent dependency relation

(MAN) "There possibly exists an intervention $I = z_i$ on X with respect to Y." (Baumgartner, 2009: 167)

(FIX) "The possible intervention $I = z_i$ is such that, while it is performed on X, all variables in the pertaining variable set V that are not located on a causal path from X to Y are held fixed, i.e. the variables in V that are not located on a causal path from X to Y can be held fixed while $I = z_i$ is performed on X." (Ibid)

(IF) "a set of variables **V** satisfies independent fixability of values if and only if for each value it is possible for a variable to take individually, it is possible (that is, "possible" in terms of their assumed definitional, logical, mathematical, or mereological relations or "metaphysically possible") to set the variable to that value via an intervention, concurrently with each of the other variables in **V** also being set to any of its individually possible values by independent interventions." (Woodward, 2011a: 11-12)