

MENTAL CAUSATION IS NOT COUNTERFACTUAL:

A DEFENSE OF THE EXCLUSION ARGUMENT

AGAINST COUNTERFACTUAL SOLUTIONS



SANGHUN HAN

BOĞAZIÇI UNIVERSITY

2018

MENTAL CAUSATION IS NOT COUNTERFACTUAL:

A DEFENSE OF THE EXCLUSION ARGUMENT

AGAINST COUNTERFACTUAL SOLUTIONS

Thesis submitted to the

Institute for Graduate Studies in Social Sciences

in partial fulfillment of the requirements for the degree of

Master of Arts

in

Philosophy

by

Sanghun Han

Boğaziçi University

2018

Mental Causation Is Not Counterfactual:

A Defense of the Exclusion Argument Against Counterfactual Solutions

The thesis of Sanghun Han has  
been approved by:

Prof. Stephen Voss



---

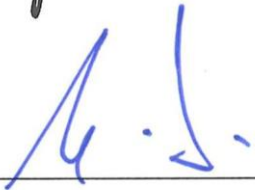
(Thesis Advisor)

Assist. Prof. Sun Demirli



---

Assist. Prof. István Aranyosi



---

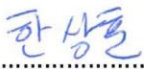
(External Member)

September 2018

## DECLARATION OF ORIGINALITY

I, Sanghun Han, certify that

- I am the sole author of this thesis and that I have fully acknowledged and documented in my thesis all sources of ideas and words, including digital resources, which have been produced or published by another person or institution;
- this thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- this is a true copy of the thesis approved by my advisor and thesis committee at Boğaziçi University, including final revisions required by them.

Signature.....

Date .....17.10.2018

## ABSTRACT

### Mental Causation Is Not Counterfactual:

#### A Defense of the Exclusion Argument Against Counterfactual Solutions

The modern debate in philosophy of mind revolves around reductive and nonreductive physicalism. Especially the reducibility of the mental to the physical has been one of the central issues in the debate. The so-called exclusion argument put forward by Jaegwon Kim is thought to raise grave challenges to nonreductive physicalism, for it shows that the theses of nonreductive physicalism are incoherent. As a response, some nonreductive physicalists have attempted to refute the argument by way of counterfactual analysis. More specifically, those counterfactual solutions are to falsify one of the core assumptions of the exclusion argument, viz., overdetermination. In this study, it is argued that the counterfactual solutions are not successful given that the mental and the physical are tied by a special sort of relation, namely, supervenience. To this end, three arguments are presented: (1) the counterfactual solutions give rise to semantic emptiness as to counterfactual talks. (2) the counterfactual solutions actually entail a certain version of the exclusion argument. (3) the counterfactual solutions wrongly interpret supervenience as a causal relation. The conclusion is drawn that the defense of the exclusion argument against the counterfactual solutions give some plausible reasons why reductive physicalism is a more viable option, and that instead of the dependence view of causation, the production view of causation seems to be a better candidate when it comes to the relation between the mental and the physical.

## ÖZET

### Mental Nedensellik Karşılgusal Değildir

#### Dışlama Argümanının Karşılgusal Çözümlere Karşı Bir Savunması

Zihin felsefesinin modern tartışması indirgemeci ve indirgemeci olmayan fizikselcilik etrafında dönmektedir. Özellikle zihinsel olanın fiziksel olana indirgenmesi tartışmanın merkezinde olagelmıştır. Jaegwon Kim tarafından ortaya atılan dışlama argümanının, indirgemeci olmayan fizikselciliğin tutarsız olduğunu gösterdiği sebebiyle bu görüşe büyük bir sorun çıkardığı düşünülmektedir. Bazı indirgemeci olmayan fizikselciler cevap olarak argümanı karşılgusal analiz ile çürütmeye çalışmışlardır. Karşılgusal çözümler dışlama argümanının merkezi varsayımı olan aşırı belirlenme varsayımını yanlışlamaya çalışmaktadır. Bu çalışmada karşılgusal çözümlerin başarılı olmadığı, çünkü zihinsel olan ve fiziksel olanın özel bir ilişki ile, ardıllık ile bağlı oldukları savunulacaktır. Bunu yapmak için üç argüman sunulacaktır: (1) karşılgusal çözümler karşılgusal cümleleri anlamca boş hale getirmektedir. (2) karşılgusal çözümler aslında dışlama argümanının başka bir biçimine sebebiyet vermektedir. (3) karşılgusal çözümler ardıllığı yanlış bir biçimde nedensel olarak yorumlamaktadır. Sonuç olarak, dışlama argümanının karşılgusal çözümlere karşı savunusunun indirgemeci fizikselciliğin neden doğru olduğuna dair makul sebepler verdiği ve zihinsel ile fiziksel olanın ilişkisi için nedensellik ile ilgili bağımlılık görüşünün değil üretim görüşünün daha uygun bir aday olduğu savunulacaktır.

## ACKNOWLEDGMENTS

My gratitude goes first to my thesis advisor, Prof. Stephen Voss. I have no doubt that my work would have been impossible without his support. Stephen is a true example of what the Korean word *seu-seung* (teacher) stands for: a person whose way of living exemplifies how one should live one's life. Not only is he a great scholar, but also a great person. I would also like to express my gratitude to Prof. Sun Demirli and Prof. István Aranyosi for their help as thesis committee members, and to Prof. Andrew Hansang Kim, my undergrad advisor in South Korea. My parents and siblings, who have always been supportive of me, should be mentioned, too. I'd like to give a big shout-out to Hashem Ramadan and Özcan Karabağ, my dearest friends, who stayed beside me and emotionally supported me during my hard times. Lastly, I want to thank Adil Alibaş for his help with the translation of the abstract.

## TABLE OF CONTENTS

|  |    |
|--|----|
| CHAPTER1: INTRODUCTION.....  | 1  |
| CHAPTER2: PHYSICALISM OR NON-PHYSICALISM?.....                         | 4  |
| 2.1 Cartesian dualism.....   | 5  |
| 2.2 Physicalism: reductive or non-reductive?.....                      | 7  |
| CHAPTER3: REDUCTIVE AND NONREDUCTIVE PHYSICALISM.....                  | 12 |
| 3.1 Physicalism.....   | 12 |
| 3.2 Supervenience.....   | 15 |
| 3.3 Reductive and nonreductive physicalism defined.....                | 22 |
| 3.4 More on reduction: traditional reduction and Kimian reduction..... | 25 |
| CHAPTER4: THE EXCLUSION ARGUMENT.....                                  | 31 |
| 4.1 What is the exclusion argument?.....                               | 31 |
| 4.2 Can overdetermination be allowed?.....                             | 33 |
| CHAPTER5: OBJECTIONS TO THE EXCLUSION ARGUMENT.....                    | 36 |
| 5.1 Challenges to the exclusion argument.....                          | 36 |
| 5.2 Versions of the counterfactual solution.....                       | 39 |
| CHAPTER6: OBJECTIONS TO THE COUNTERFACTUAL SOLUTIONS.....              | 47 |
| 6.1 Counterfactuals and semantics.....                                 | 47 |
| 6.2 Counterfactuals and downward causation.....                        | 51 |
| 6.3 Counterfactuals and supervenience.....                             | 56 |
| CHATPER7: CONCLUSION.....  | 66 |
| REFERENCES.....  | 68 |



# CHAPTER 1

## INTRODUCTION

Much responsibility is owed to Rene Descartes when it comes to sparking the modern questions on the mind-body problem. Especially the questions as to the nature of the mind and body and as to the causal relation, if there is, between them have perplexed people ever since. Worse yet, it seems that these questions have become more intractable than ever due to the rise of neuroscience and empirical evidence it suggests.

It is no wonder the questions as to the relationship between mind and body have become since then one of the main subjects in philosophy of mind. And I think that mental causation that surveys whether the mind or, put more precisely, mental properties are causally efficacious is of the utmost importance among those subjects. For the belief that our behaviors causally have to do with mental states or the “mind” is perhaps the most fundamental assumption of folk psychology that underlies the way we interact with the world. And this point is as clear as it gets when we consider how we explain our behaviors when asked why. For instance, I’m drinking a coffee because I have a desire to drink a coffee; my desire directs my bodily movements in such a way that leads to a coffee shop, to get a coffee, and to drink a coffee. Indeed, it seems that the causal explanation of our behaviors is incomplete without invoking some mental properties that figure in the process. Yet, the effects of the mental on our behaviors often go unnoticed. I believe this is because the thesis that the mental causally interacts with the physical is that of folk psychology which lay people

subscribe to without recognition. Thus, people just take it for granted that the mental is causally efficacious as to the physical while ignoring why and how. If mental causation is so prevalent in our daily life, then it needs good qualifications instead of taking it as a brute fact. I believe this gives a good enough reason why mental causation is worthy of exploring. So this was my personal motivation why I want to discuss about mental causation in this thesis. Now I'd like to mention specifically what topic in mental causation I want to discuss.

Broadly speaking, what I want to achieve here is defending one of the ontological and metaphysical positions involved in mental causation: reductive physicalism. More specifically, I argue that a certain version of reductive physicalism, viz., Jaegwon Kim's version of reductive physicalism, seems to be the only plausible position that *makes sense* in the face of our common sensical understanding of a scientific world and certain philosophical issues that will be presented in this thesis. Reductive physicalism, however, has been on the downside for some decades; during its downfall, nonreductive physicalism considered as its rival view has gained quite a popularity in the debate. What lies at the heart of nonreductive physicalism's popularity is an alleged success of refuting the exclusion argument.<sup>1</sup> Many nonreductive physicalists refute the exclusion problem by arguing that the exclusion argument which takes a form of *reductio* and argues for the causal inertness of mental states has a false premise in its construction. These nonreductive physicalists falsify the premise by a counterfactual analysis of causation and are thus

---

<sup>1</sup> The original term for the exclusion problem is the 'causal-explanatory exclusion' and 'explanatory exclusion' found in Kim (1989a). But since its first appearance, the term has taken various similar names such as the 'exclusion problem', 'exclusion argument', 'exclusion principle', and 'supervenience argument'. In this thesis the 'exclusion argument' will be the standard term but some other terms might be used interchangeably in some chapters.

standardly called ‘counterfactualists’. In this thesis, I want to defend reductive physicalism by pointing out that the counterfactual solutions to the exclusion argument is far from being right.

In sum, I aim at achieving two interconnected goals. On the one hand, in a narrow sense I attempt to defend the exclusion argument. In a broad sense, on the other hand, by defending the exclusion argument, I attempt to show that reductive physicalism remains as a viable option in the mental causation debate.

Lastly, here is the order of the discussions to follow in this thesis: In chapter 2 as an introductory part, I discuss alternative views other than reductive and nonreductive physicalism in the debate. Chapter 3 will be a detailed discussion of reductive and nonreductive physicalism; their central doctrines and implications will be discussed. The exclusion argument will be the topic of Chapter 4. There I explicate how the *reductio* is completed with the explanations on the core premises of the exclusion argument. Followed by Chapter 5 is several counterfactualist approaches that try to render the exclusion argument moot. Chapter 6 examines the criticisms of the counterfactual approaches. Three criticisms will be raised against them. Chapter 7 will summarize the whole discussions and draw some implications and lessons from them.

## CHAPTER 2

### PHYSICALISM OR NON-PHYSICALISM?

The issue I will take up in this thesis revolves around physicalism. Thus, the discussions to follow in the other chapters will take physicalism for granted and unfold explicitly *within* the physicalist scheme. Other alternative metaphysical views will thus be mostly neglected. But those alternative views need a considerable attention too. For paying attention to them can help us draw the big picture as to the mind-body problem. Specifically, I believe that discussing those alternative views can be beneficial for the following reasons: first, it can help us see where we are currently standing as to the mind-body problem. Second, it can show why physicalism—reductive and nonreductive physicalism in particular—matters for us, thus giving some credits to the importance of what I will attempt to do in this thesis.

The most basic distinction as to the mind-body problem is between physicalism and non-physicalism. Broadly speaking, physicalism is a view that the mind-body problem and its solution are explicitly *within* the realm of the physical, signifying that the mind is grounded in the physical. There is much more to be discussed about what “being physically grounded” means. And this topic will be discussed later. On the other hand, non-physicalism is a view that the mind-body problem is not the problem that invokes only a physical domain; rather, there is another realm, viz., a mental realm, that enjoys its own metaphysical life. As far as I can see, all the theories about the mind-body problem can be categorized under either physicalism or non-physicalism.

## 2.1 Cartesian dualism

The paradigmatic case of non-physicalism, or its alternative name dualism, can be found in the writings of the 17<sup>th</sup> century philosopher Rene Descartes. Indeed, it wouldn't be an exaggeration to say that the modern-day dualists are one way or another all heir to his view on the mind-body problem. More specifically, in *Meditations on First Philosophy* he explains why the mind and body, or the mental and the physical, are distinct substances that are independent of each other (1996):

And accordingly simply from this, that I know that I exist and that at the same time I notice absolutely nothing else to belong to my nature or essence but only that I am a thinking thing, I correctly conclude that my essence consists in this one [thing], that I am a thinking thing. And although (or rather, as I shall soon say, assuredly) I have a body that is very closely joined to me, nevertheless, because on the one hand I have a clear and distinct idea of myself, insofar as I am merely a thinking thing and not an extended thing, and because on the other hand I have a distinct idea of a body, insofar as it is merely an extended thing and not a thinking thing, it is certain that I am really distinct from my body, and can exist without it. (p. 64) (the translation of the first sentence is by Stephen Voss.)

The quoted passage clearly shows that the mind and body are two distinct substances that can exist independently of each other; that the mind (or the soul in Descartes' term) is clearly and distinctly conceivable separate from the body is the evidence that the mind is really a distinct substance.

Leaving aside his ultimate philosophical goal, that is, to prove that God exists, I would like to elaborate more on his mind-body dualism and the implications we can derive from it. Firstly, although Descartes thought that the mind and body, viz., the mental and the physical, stand in a causal relation, Cartesian definition of them doesn't support his own claim. According to Descartes' *Principles of Philosophy*, as a distinct substance that can independently exist with nothing else, the mind and body have thought and extension as a necessary attribute respectively

(Descartes, 1998, pp. 131-132). A challenge to Descartes is how to explain the causal work between the mind and body because the claim that the *nonextensional* mind and *extensional* body are distinct substances yet they causally interfere with one another seems to be problematic. As far as the causality of physical objects that have extension are concerned, we can think of two possibilities: (1) a physical world is causally closed. (2) a physical world is causally open, i.e., the mental as a distinct substance can causally interfere with a physical world.

The first possibility can be easily dismissed given that Descartes does think that the mind has causal effects on the physical; if a physical world is closed, then whatever happens in a physical world has sufficient physical causes and explanations. So there is no room for the mind in this picture. The second possibility is no better than the first. The perplexing question is this: if the nonextensional mind can causally affect the extensional body, exactly how is that possible? For it seems that a physical object can be a cause or an effect *only* of some other physical object that has extension. The motivation behind this claim is that causality in a physical world requires that a cause and an effect have spatiotemporal physical coordinates. Kim (2005) also points out this, by putting forth that

Causality requires a domain with a space-like structure—that is, a “space” within which objects and events can be identified by their “locations”—and, as far as we know, the domain of physical objects is the only domain with a structure of that kind. (p. 151)

But the mind in Descartes’ picture doesn’t have any physical coordinates because it is definitionally nonextensional. As a response, Descartes tries to answer the problem by supposing that a pineal gland in a brain is where the mind is physically “located”. But this begs the question because supposing so means that the mind has extension after all, which is against his definition of the mind.

There is nothing wrong to say that the mind is a substance and the physical is another substance, but it is wrong to say that their being two distinct substances and their being able to causally affect one another are compatible. The problem of substance dualism is thus its internal incoherency as to the causal influence on one another. Moreover, the commonsense of our era seems to suggest that it is very unlikely that the mind is beyond the body; the growing body of empirical evidence from science points out that the mind is after all located in a physical brain. For these reasons, substance dualism of Cartesian sort doesn't receive much attention in the modern-day mind-body debate. Rather, the debate revolves around the alternative view, viz., substance monism.

## 2.2 Physicalism: reductive or non-reductive?

The substance monism referred to in the above passage is physicalism. Physicalism is a view that there is only a physical realm in the world.<sup>2</sup> Naturally, the mind also belongs to the physical realm, and physicalists try to deal with the mind-body problem exclusively in the physical domain. However, there are disputes as to the nature of the mental in the physical domain that separate physicalists into two groups: reductive and nonreductive physicalists. As I have said earlier, however, I won't dig into them in this chapter. Being the main subjects of this thesis, reductive and nonreductive physicalism will be investigated in detail in the next chapter. Instead, I would like to give a brief overview of some of the other alternative views in physicalism: eliminativism and epiphenomenalism.

---

<sup>2</sup> One can of course be a mental monist or be a monist that is different from both—like Spinoza. But I simply skip these because I believe discussing them isn't essential for my thesis.

### 2.2.1 Eliminativism

One central assumption of the hitherto discussion is that the mental exists. One may say that it isn't even an assumption because it seems just too obvious that there is the mental. Indeed, the fact that we, I and the readers, are now reading and thinking about this thesis seems to show the very existence of the mental; for the medium through which we can read and think in the first place seems to be the mental thoughts that arise from reading the thesis. This way of thinking—that the mental guides and regulates our behaviors—is prevalent in our daily life.

Eliminativists, however, would say that once looked carefully there is no such thing as the mental. Paul Churchland is perhaps one of the most prominent philosophers who have defended eliminativism. For instance, Churchland (1981) famously argues that the existence of the mental is a myth.<sup>3</sup> Churchland's argument turns on the claim that folk psychology in which the mental is expressed through propositional attitudes is to be taken as a scientific theory. If folk psychology is a scientific theory, then propositional attitudes like beliefs and desires are theoretical assumptions postulated as explanatory tools in the theory. According to this view, we believe the existence of the mental because it *helps* explaining why we did such and such things. Thus the defense of its existence mostly leans on its explanatoriness. A crude analogy might help here: the ancient Greeks believed in Zeus, Poseidon, and other gods and thought that they are responsible for changes in nature *because* they needed some mechanism by which changes in the nature are explained. Likewise, the

---

<sup>3</sup> A caution is required as to what mental properties are subsumed under Churchland's definition of the mental. His use of the term 'the mental' is limited only to those mental states that take propositional forms among which beliefs and desires are the most important. So his goal is to eliminate mental states that involve propositional attitudes while the phenomenal mental properties or qualia are left aside in his project.



mental in folk psychology is no different in that people are in need of some mechanism by which the behaviors of people are explained.

However, the mental as part of a scientific theory of course renders it vulnerable to falsification, just like the substance called phlogiston in a phlogiston theory of combustion by Georg Stahl was a hypothetical assumption so as to explain combustion but later discarded as a better model came into prominence, viz., Antoine Lavoisier's oxygen theory of combustion. Churchland believes that the development in neuroscience and the possibility of a better model of the mind therefrom which won't invoke the mental as propositional attitudes are only to show that the fate of a phlogiston theory will be likely to befall folk psychology that invokes the mental.

Another indirectly related topic as to eliminativism, I believe, is pragmatism as found in Daniel Dennett's works. Dennett (1971) defends the existence of the mental states expressed through propositional attitudes on the ground that supposing so has so far yielded the best outcome in explaining and predicting behaviors of various forms of living and non-living things in the world. And indeed, Dennett *does* believe that the mental is real and Churchland admits that he stands on the other side of his position. However, I think there is room for an eliminative interpretation of Dennett's account. It seems to me that according to Dennett, the fact that there hasn't been any theory that better explains and predicts behaviors of various sorts than folk psychology is the sole ground why we stick to the theory of the mind that invokes the mental. In other words, it remains open that if there is a pragmatically better model that doesn't invoke the mental, then pragmatists like Dennett wouldn't hesitate to discard the old and to adopt the new model.

### 2.2.2 Epiphenomenalism

Another branch in physicalism is epiphenomenalism. It is less extreme than eliminativism given that epiphenomenalism doesn't deny the existence of the mental while eliminativism does. However, as will be shown shortly, its ontological and causal implications as to the mental are no less agreeable than the eliminativist implications.

As the name implies, epiphenomenalists (e.g., Huxley, 1874, and James, 1879) argue that although the mental exists and does so by virtue of being physically grounded, it is merely epiphenomenal in the sense that it doesn't have any effect on the physical. This has a close connection with the late 19<sup>th</sup> and early 20<sup>th</sup> philosopher Samuel Alexander (1927) and the so-called Alexander's Dictum: to be is to have causal powers. In later chapters where I discuss ontology this dictum will reappear and be investigated in more detail. In any case, the causal status of mental properties is what distinguishes epiphenomenalism from emergentism. They are often mistakenly thought as more or less the same claim for a simple reason that both of them subscribe to the existence of mental properties. However, they crucially differ in that epiphenomenalists argue that mental properties are causally inert—thus epiphenomenalism—whereas emergentists think that they do have some genuine causal power. Let me give a rough example to illustrate the difference. Imagine a baseball shattering the window. The baseball is physically grounded in that it is made up of a group of countless many atoms at a physical level. Epiphenomenalists would say in this case that the shattering of the window is solely accounted for by its atoms, not the baseball. On the other hand, emergentists would say that the baseball composed of those atoms has some new causal power that the mere assembly of the

atoms doesn't possess. I think that the endorsement of causal power on mental properties is an important issue and that this is why epiphenomenalism raises some serious questions about the ontological status of mental properties. Specifically, when epiphenomenalists argue that mental properties are mere epiphenomena of their physical bases, we should answer the following question: what is, then, their role in the world? It seems that things without causal power are, in Herbert Feigl's term, nomological "danglers". Together with Alexander's Dictum, epiphenomenalism is no better option than eliminativism; no causal work means no existence. Answering exactly why epiphenomenalism implies the causal impotence of mental properties requires the discussion of reduction and the exclusion argument which will be discussed in due course. But epiphenomenalism taken at a face value already signifies that it cannot be an attractive option if we want that our beliefs and desires give rise to changes in our actions and behaviors.

## CHAPTER 3

### REDUCTIVE AND NONREDUCTIVE PHYSICALISM

Now it is time to deal with what we mean by “reductive and nonreductive physicalism”. The first thing to point out is the scope of the term. By doing so, however, we only alleviate the wide interpretability of the terms. This is evident given that when various philosophers of mind talk about reductive and nonreductive physicalism, what they mean are often more or less different from one another. For example, in one sense physicalism can be understood as incorporating not only the bottom-level physics, but also all the higher-order sciences such as chemistry, biology, and psychology. Or in another sense, physicalism can simply be about the bottom-level physics. Given these circumstances, that is, that there are multiple definitions of the terms and that which version is correct hasn’t been clearly settled yet, I hope to achieve at best outlining some uncontroversial generalizations on each view that all, or most, would agree with. This will establish some fair, neutral ground that favors none to begin with.

#### 3.1 Physicalism

It would be convenient to start with something that the contradicting positions have in common: physicalism. So here is a rough definition of it: as the name implies, physicalism is a view that the world is wholly made up of physical matters. Although it gives some useful insight about physicalism, this definition is not specific enough; after all, what does it mean for something to be physical? We may answer it by

invoking folk-physics, that is, something is physical iff it is wholly composed by elementary physical matters such as atoms, photons, electrons, and so on. It is important to remark that the only thing required for me to proceed my argument is that there *be* a bottom level constituent of physics. Be that as it may, atoms, photons, electrons, or something unknown to us might be such a constituent. But it wouldn't affect the validity of my argument even a bit; the argument I'd like to espouse is successful as long as there is *any* bottom level physical matter to which higher level physical or mental properties can be reduced. Thus, although it would be an interesting task, answering what really is the bottom level physical matter is not something I will attempt to do here.

To recapitulate, physicalism is a view that the world is wholly made up of the bottom level physical matter. Or to borrow Kim's definition (2005):

The core of contemporary physicalism is the idea that all things that exist in this world are bits of matter and structures aggregated out of bits of matter, all behaving in accordance with laws of physics, and that any phenomenon of the world can be physically explained if it can be explained at all. (pp. 149-150)

One corollary of physicalism understood as such is that the world is physically closed in terms of causality. This seems evident given that if the world is physical *simpliciter*, then any explanation of events that happen in it must ultimately be physical; to invoke something other than physical is to violate the rule of the game, as Descartes mistakenly did. We may put this point as a principle (Kim, 2005, p. 16):

*Physical Closure*: everything that happens at *t* has a sufficient physical cause at *t*.

Physical Closure will recur throughout the discussion since it is required for the exclusion argument to succeed.

At this point, one might ask the nature of (allegedly or seemingly) nonphysical properties in this picture. For example, do mental properties have any

roles in this picture? Clearly, there is something different about mental properties *contra* physical properties. But there is room for accommodating mental properties in physicalism. That is, if mental properties are somehow grounded in the physical, then there is nothing contradictory or problematic for us to talk about mental properties in a physical world. More will be said about this.

One important remark is in order before moving on: why do we have to stick to the supposition that the world is made up of the *bottom level* physical matters? Apart from ontological simplicity, is there any reason to suppose exactly as such? Yes, as far as the exclusion argument is concerned. According to Lei Zhong, it is necessary that the physical in physicalism is only the bottom level physical constituents. If the physical incorporates higher-order properties like mental properties, then the targets of the exclusion argument, namely, those higher-order mental properties cannot be attacked, making the exclusion argument useless in the first place (Zhong, 2011, pp. 135-136). It would be too quick to discuss exactly why this is the case because that would require a complete elaboration on the exclusion argument. For now, the following remark is enough: if the exclusion problem were to be successful, there should be a distinction between lower- and higher-order physical properties because Physical Closure, one of the necessary premises of the exclusion argument, draws its force from the supposition that there is such a distinction. Without the distinction, Physical Closure wouldn't have any meaningful bearing on physicalism, and the exclusion argument doesn't arise in the first place.

For most people, the surface definition of physicalism is easy to digest. The problem rather lies in the exact articulation of implications that follow from it—like the nature of mental properties in that picture. As I promised, besides the talk of

reduction and non-reduction, I would like to deal more with the nature of mental properties in physicalism in the next section. These two go hand in hand given that the clarification of the nature of mental properties will help defining what reductive and nonreductive physicalism are.

### 3.2 Supervenience

It is hardly an exaggeration to say that supervenience is the most central, underlying metaphysical concept in the mental causation debate. Much of the credit goes to G. E. Moore's work on ethics (1922) when it comes to the concept of supervenience coming into philosophers' attention. However, Moore himself didn't coin the term 'supervenience'; it was R. M. Hare who first used the term (1952). Another important figure in the discussion of supervenience is Donald Davidson (1970). It wasn't until Donald Davidson's introduction of supervenience to philosophy of mind that supervenience was recognized in philosophy of mind. Since then, supervenience has been thought to offer a neat framework in making sense of the mind.

Imagine Chuck has some characteristics such as hospitality, kindness, and empathy. Typically, we take these characteristics as comprising the criteria for judging someone to be good (or at least let's suppose so). In that case, we say Chuck is a good person because he has those characteristics. Or alternatively, those characteristics determine or entail Chuck's being a good person. Indeed, we are justified to call anyone with those characteristics a good person as long as there are no background assumptions that counter his or her good character; in the absence of those extra assumptions, whoever with hospitality, kindness, and empathy *must* be a good person. We can say that in the current case supervenience is the intimate

relationship between goodness and those characteristics. More specifically, supervenience tells us about what kind of determinate relationship between two distinct properties, events, or facts is held. Examples: you draw 3 straight lines that meet each other's end, it becomes a triangle; you create a molecular structure by combining one oxygen atom and two hydrogen atoms, it becomes a water molecule. In each case, we may say that a triangle supervenes on the way the lines are drawn, and that a water molecule supervenes on the way those atoms are combined. One thing to note is that the relation between supervenient and subvenient properties is asymmetrical. For instance, in the triangle example you can change the shape of a triangle by altering the lines but you can't do the other way around. In short, a simple catch phrase holds in any case of supervenience: no alteration in the determined if no alteration in the determining.

That there is a certain interconnectedness between distinct events, objects, properties, and whatnot in the world is the underlying assumption that makes it possible for us to make sense of the world (Kim, 1984, p. 153). Making sense of what happens in the world without such interconnectedness seems utterly impossible. For example, imagine a world without causation as falling under one type of this interconnectedness. It would be impossible or sheer luck if there is any explanation that works at all. In that world without the notion of causation, I wouldn't be able to account for why I eat food when I'm hungry. So it seems to me that supervenience is worthy of attention given that just like causation, it regulates the way we interact with the world. But there needs to be some qualification if we are to apply supervenience to the issue at hand, viz., the mental causation debate, because supervenience can be thought to carry different implications depending on what one



means by it. Accordingly, what kind of supervenience is appropriate in assessing the mind-body problem needs to be investigated. More specifically, there are three types of supervenience that I have in mind, and I will discuss each of them to the effect that only one (or arguably two) of them is appropriate for the mind-body problem.

Kim has extensively written on the notion of supervenience (Kim, 1984, 1987, 1988, 1998, and 2005). And in his discussion of supervenience, three types of supervenience are distinguished which he calls as ‘weak’, ‘strong’, and ‘global’ supervenience. I think this classification comes in handy given that it catches the intuition behind each view. As will be clear, only strong supervenience can accommodate the degree of determination relation that we expect to hold in talking about supervenience in the current context. But let us first start with weak supervenience. To do so, let me first make sure of the use of terms and symbols in this section. Let italicized lower case letters  $x$  and  $y$  designate individuals. And italicized capital letters  $A$  and  $B$  will be terms for kind predicates of an individual such that  $A$  designates a mental kind and  $B$  designates a physical kind. Lastly, let capital letters  $F$  and  $G$  be properties that fall under  $A$  and  $B$  respectively. For instance, let  $x$  be Chuck, and  $F$  be a desire to have a coffee. Then Chuck’s desire to have a coffee =  $F(x)$ . All the quotes in this section are modified only to the extent that I changed those symbols to the ones described above. Now, following Kim (1984), we may formally define weak supervenience:

Weak Supervenience:  $A$  *weakly supervenes* on  $B$  if and only if necessarily for any property  $F$  in  $A$ , if an object  $x$  has  $F$ , then there exists a property  $G$  in  $B$  such that  $x$  has  $G$ , and if any  $y$  has  $G$  it has  $F$ .<sup>4</sup> (p. 163)

---

<sup>4</sup> Throughout this thesis supervenience will be about supervenience between mental and physical properties. If one wants to be more specific, we can say that an event where a mental property is instantiated supervenes on an event where a physical property is instantiated.

According to Kim, Hare (1952) and one of the versions of supervenience found in Davidson (1970) imply this weak supervenience. However, weak supervenience doesn't promise a level of determination we would like to mean when we talk about supervenience; weak supervenience cannot guarantee a fixed supervenience relation between F and G across various possible worlds. For instance, suppose that in the actual world, a C-fiber firing is correlated with pain such that whenever one has a C-fiber firing, one experiences pain. Let G be a C-fiber firing predicate and F be a pain predicate. Then pain supervenes on a C-fiber firing =  $(\forall x)[G(x) \rightarrow F(x)]$ . But imagine a possible world where a law pertaining to pain realization is slightly different, to the effect that in that world a C-fiber firing is correlated with happiness—whenever one has a C-fiber firing in that world, one experiences happiness. Let H be a predicate about happiness. Consequently,  $(\forall x)[G(x) \rightarrow F(x)]$  doesn't hold in that world because  $(\forall x)[G(x) \rightarrow H(x)]$  holds instead.

But why is the instability of weak supervenience across various possible worlds problematic in the first place? In order to see why, there has to be a question as to what kind of determinate relation we want to hold when we talk about supervenience. Although it is intuitive, the following passage by Kim (1984) shows what sort of problem weak supervenience raises:

Determination or dependence is naturally thought of as carrying a certain modal force: if being a good man is dependent on, or is determined by, certain traits of character, then having these traits must insure or guarantee being a good man (or lacking certain of these traits must insure that one not be a good man). The connection between these traits and being a good man must be more than a de facto coincidence that varies from world to world. (p. 160)

Generally, we want the laws that hold in our actual world to hold in other possible worlds as well for two reasons. The first is semantical. Without supervenience stable

across possible worlds, there will be a significant constraint on our use of counterfactual talk. As Kim points out, saying, for example, that “Chuck would have been a good person if he had had such and such characteristics” would be meaningful only if we suppose that being a good person consists in having exactly the same such and such characteristics in other possible worlds too. The second has to do with our intuition. If we suppose weak supervenience, we can imagine a possible world where the psychophysical laws are drastically different from those of the actual world. Imagine further that in the actual world  $w_1$  and some possible world  $w_2$  there is an exact duplicate of an individual  $x$  and of an occurrence of a physical property  $P$  in  $x$  such that  $P(x, w_1) = P(x, w_2)$ . However, the psychophysical laws governing these two worlds are different;  $P(x, w_1) \rightarrow M(x, w_1)$ , where  $M$  stands for an instantiation of some mental property. On the other hand,  $P(x, w_2) \rightarrow \sim M(x, w_2)$ . We can further complicate the situation by supposing that this disparity in psychophysical laws between  $w_1$  and  $w_2$  covers all the instantiation of physical properties in  $x$ . In other words, imagine a possible world in which there is an exact physical duplicate of me that lacks mentality. Although such a scenario is conceptually possible, we tend to believe that in those possible worlds where  $x$  has the same instantiations of physical properties as the actual world,  $x$  should have the same instantiations of mental properties as  $x$  in the actual world would.

Contrary to weak supervenience, strong supervenience holds that the correlation between base and supervenient properties be stable across various possible worlds. In Kim’s words (1984), formally put, strong supervenience states the following:

Strong supervenience: *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  $F$ . (p. 165)

As can be seen, the modal operator “necessarily” newly appears in the last sentence, stipulating that the correlation between  $G$  and  $F$  holds no matter what. In other words, the occurrence of the modal operator in the last sentence *fixes* the stableness of a given psychophysical correlation across various possible worlds. In addition, necessity here can be understood as metaphysical necessity. But the first thing to point out is that the specific contents of supervenience, that is, specific correlation laws that govern mental and physical properties are nomological. This seems true given that a C-fiber firing doesn’t have to be correlated with pain in all other possible worlds because we can think of a possible world where pain is correlated with, say, a D-fiber firing. However, the basic idea of supervenience that those specific contents are tied together is something invariable. In other words, pain might be correlated with different physical realizers such as C-, D-, E-fiber varying from world to world. However, it is metaphysically necessary that *if* there is a physical property and a psychophysical law in virtue of which this physical property is a supervenience-base property for some mental property, *then* whenever there is an instantiation of this physical property there *must* also be an instantiation of a mental property. Suppose that because of the physical laws governing the actual world we are living in, a C-fiber firing is correlated with pain. Then it is nomologically necessary that any possible worlds that share the same physical laws have the same pain realization correlation with a C-fiber firing. But it is metaphysically necessary that if there is such a correlation, it applies to any instance of C-fiber firings and pain.

Lastly, there is one more version of supervenience called ‘global’ supervenience. Simply put, it says the following:  $A$  globally supervenes on  $B$  iff two worlds are indiscernible with respect to  $A$  due to  $B$ -indiscernibility. I shall not,

however, investigate it deeply and will simply take strong supervenience to be the standard definition of supervenience that will be invoked throughout the discussion here. The main reason is that global supervenience is somewhat controversial as to its applicability to the mind-body problem. For instance, some authors, e.g., Bennett (2004), have pointed out that just like weak supervenience, global supervenience doesn't carry the degree of modal force that determination relation between base and supervenient properties is supposed to carry, and that a version of global supervenience that does provide such modal force is *just* strong supervenience couched in different terms. Along this line, authors like Leuenberger (2009) take the issue even further; that no version of global supervenience that has been proposed so far, i.e., weak, intermediate, and strong global supervenience, aren't able to catch the concept of global supervenience. For instance, weak global supervenience requires only that if there is the indiscernibility in *B* between two worlds, then there is the indiscernibility in *A* too (Leuenberger, 2009, p. 116). However, weak global supervenience interpreted as such doesn't promise the cross-world stableness in the distribution of properties. In other words, weak global supervenience requires only that there be one-to-one isomorphism in the instantiation of a property, leaving aside the possibility that such isomorphism doesn't obtain to the *same* individual (Leuenberger, 2009, p. 117). So it's possible that in the actual world if I have a C-fiber firing I have pain, but in some other possible world, if I have a C-fiber firing, Chuck has pain. Intermediate and strong global supervenience, on the other hand, face different difficulties. The definition of intermediate and strong global supervenience differs in that the intermediate version puts that the indiscernibility in *B* between two worlds entails the indiscernibility of *some* properties in *A* while the

strong version asserts that the indiscernibility in *B* entails the indiscernibility of *every* property in *A* (Leuenberger, 2009, p. 117). Be that as it may, the truth or falsity of those claims matters little: as far as the discussion in this thesis goes, strong supervenience alone is sufficient. Thus, although it is an interesting task to investigate the nature of global supervenience, I will remain silent on this issue. Those who want to use global supervenience instead of strong supervenience are welcome to do so as long as global supervenience is cross-world stable.

In any case, given supervenience we can talk more specifically about the mental in a physical world: *mental properties exist in the actual and other possible worlds similar to the actual world, physically grounded by supervenience*. For instance, now I have a certain mental property—a desire to grab a cup of coffee. And this desire is supervenient on certain neurochemical property in my brain; my having this mental property supervenes on my having this neurochemical property given that there is a law-like correlation between the two. And supervenience is neutral as to reductive and nonreductive physicalism due to the fact that the only thing it requires is that the mental is supervenient on the physical. It will thus be safe to say that supervenience defines minimal physicalism (Kim, 2005, p. 13).

### 3.3 Reductive and nonreductive physicalism defined

Once supervenience is understood, we can make sense of the adjective “reductive” and “nonreductive” in the current context with it. First to note is that what I try to do is just like the previous section; there are various thoughts on what constitutes the nonreductive part of nonreductive physicalism, and the same for reductive physicalism. I merely try to suggest some uncontroversial generalization. Given this,

I think the following three theses are piercing through any nonreductive physicalists' heart (List and Menzies, 2009, p. 475):

(NRP1): Mental properties are not identical to physical properties.

(NRP2): Mental properties cause other mental properties and physical properties.

(NRP3): Mental properties supervene on physical properties.

(NRP3) states supervenience just explained above, so it is accepted both by reductive and nonreductive physicalists. The controversy rather lies in (NRP1) and (NRP2).

Indeed, vindicating or falsifying these two theses *is* the mental causation debate.

How they bear such a consequence will naturally appear as the discussion here unfolds. Now contrast these theses with the following reductive physicalist theses:

(RP1): Mental properties are reducible to physical properties.

(RP2): Mental properties do not cause other mental and physical properties—they are nothing over and above the physical.

(RP3): Mental properties supervene on physical properties.

(RP1) and (RP2) claim exactly the opposite of (NRP1) and (NRP2)—that mental properties are reducible to physical properties and thus that if mental properties cause anything, they do so *qua* physical, not *qua* mental.

To see why the reducibility and causal power of mental properties have caused so much trouble in philosophy of mind, more have to be said about (NRP1) and (NRP2). These two notions go hand in hand, in that to argue for (NRP1), (NRP2) is required, and vice versa. To start off, why should we subscribe to (NRP2)? The intuitive answer seems to be: because mental properties *seem* to really cause things in the world. Initially, it seems too trivial that my desire to grab a coffee makes me move my body to a coffee shop. But nonreductive physicalists need more than folk-

psychological explanation. For it is exactly this point that reductive physicalists would like to reject—that my desire to grab a coffee does cause my action; rather, they would argue that it is the neurochemical mechanism that plays the causal role.

So what other reasons are there for (NRP2)? Answering this requires a commitment to a certain ontological picture, viz., Alexander's Dictum that causal power is what defines being real. I think this ontological commitment is right. First, we would like the world to be ontologically parsimonious, that is, we would like unnecessary things out in our ontology. And by unnecessary things, pragmatically speaking, I mean things that are not causally efficacious. Consider two options: you can either account for something with one real cause or account for it with one real cause plus one extra causally inert thing. The former seems a better option; for why would you add an unnecessary item in explaining something when you can still fully explain without it? I think this speaks well for why ontological parsimony and causal efficaciousness matter. Nonreductive physicalists claim (NRP2) because they want mental properties to be meaningful in the world. Without real causal work, there is no meaningful contribution to the world, so it must be gone. In other words, causal efficaciousness as a necessary and sufficient condition for something to be real, (NRP2) sets forth an ontological reality of mental properties.

What about (NRP1)? It's a natural consequence from (NRP2). If you want to say that mental properties are real, you have to argue for their causal efficaciousness. And if you want to say that mental properties are causally efficacious, you have to argue for their irreducibility to the physical. To see why, for *reductio ad absurdum*, suppose that mental properties are efficacious while they are reducible to physical properties. Given that mental properties are nothing over and above physical



properties, all the causal power that mental properties have is from physical properties that ground them. This ultimately begs the question, what causal power do mental properties have in this case? All the causal powers they allegedly have are from the physical properties, so it only points to the existence of the physical properties, not the mental properties. For the mental properties to exist, they must have some causal power. As long as the two assumptions—the causal efficacious of mental properties and their reducibility to physical properties—hold, there is no way to argue for mental properties’ unique causal efficaciousness.<sup>5</sup> In short, in order for (NRP2) to be consistent, (NRP1) is required, and vice versa. They go hand in hand.

#### 3.4 More on reduction: traditional reduction and Kimian reduction

Lastly, before I move on to the next chapter, I want to say something more about reduction. Up until now, I have taken reduction for granted without qualification. In this section, I will try to say more about what kind of reduction I meant.

The first to note is that throughout the discussion so far what I meant by ‘reduction’ is actually a version of reduction model that Kim has espoused (Kim,

---

<sup>5</sup> One may object by saying that when mental properties supervene on physical properties, a new causal power that belongs to mental properties emerges from physical properties by which mental properties can be distinguished in a causal sense. I take this as one form of emergentist arguments and think that this isn’t a viable option. At the heart of emergentism is the idea that the newly emerged powers (or properties) cannot be explained in terms of underlying physical properties. Think of vitalism as a version of emergentism: individual cells do not give rise to life if they are separate but once they are combined in a certain way, there arises a living organism with life. Yet, we can’t explain how life emerges from a combination of cells by tracing each cell separately. In other words, emergentism conceived as such carries some mystical element in its core in that there is an unknown, mysterious “missing link” that explains how underlying physical properties give rise to emerging properties. I think this point goes against the physical closure principle; for it states that things that happen in the physical realm must be *fully* explainable physically. The emergentism argument above seems to fail to pass this principle because it doesn’t offer a fully physical explanation of how the newly emerged properties come to exist. In addition, this way of viewing mental properties seems to block the functional reduction of mental properties to physical properties in the first place given that a new causal power of a mental property doesn’t have anything to do with its physical base property (Kim, 1998, p. 12).

1998 and 2005). The modern debate on reduction dates back to the 1950s and 60s, and Ernest Nagel (1961) played an important role in the debate. And as will be shown, in the face of certain philosophical doctrines and concepts, the reductive physicalist project that tries to reduce the mental to the physical by way of Nagelian reduction has lost its philosophical merit. What I will try to do here is two-fold: first, I review Nagelian reduction along with type-identity theory, multiple realization, and functionalism as threatening the prospect of Nagelian reduction in the mind-body problem. Second, I distinguish Nagelian reduction and Kimian reduction by showing how Kim overcomes the previous challenges raised against the Nagelian model. At the end, the discussion of Kimian reduction will shed some light on what kind of reductive physicalism Kim has in mind. The main thesis of this paper is merely to lend a hand to support Kim's reductive physicalism. I hope this section is helpful not only in that it discusses the history of modern-day debate on reduction, but also in that it broadly sketches the philosophical picture that Kim has favored. But this will be brief given that it's only for a survey; it won't affect any of the arguments to follow in this thesis.

The modern-day debate on the mind-body problem has its roots in so-called type identity theory suggested by authors like U. T. Place (1956), Herbert Feigl (1958), and J. J. C. Smart (1959). We don't have to go in detail about the theory, but the basic claim is this: all the mental properties that belong to the mental kind *are* just physical properties that belong to the physical kind. The view that mental and physical properties stand in a strict identity can be found in the works of the abovementioned authors. For instance, Smart (1959) puts that

When I say that a sensation is a brain process or that lighting is an electric discharge I do not mean just that the sensation is somehow spatially or

temporally continuous with the brain process or that the lightning is just spatially or temporally continuous with the discharge. (p. 145)

I take it this to imply that mental and physical properties *are* just one and the same things, meaning that mental properties are nothing “over and above” physical properties. In this picture, pain = C-fiber firing *tout court*. And this identity covers every instance of pain experience, meaning that different species with different physical systems have the same pain-realizing mechanism.

While type identity theory was discussed, Nagelian reduction was thought to offer a neat logical method of reduction for type identity theory. According to the Nagelian model, a theory to be reduced is reducible to the underlying basic theory if and only if there is a bridge law that connects the two.<sup>6</sup> The requirement of a bridge law is based on the claim that it is possible for the reduced theory (or secondary) theory to contain in it descriptive predicates that do not occur in the reducing (or primary) theory (Nagel, 1961, p. 342). And the following quote from Nagel (1961) shows why the reduction of two theories is impossible without a bridge law:

As has already been indicated in this chapter, a reduction is effected when the experimental laws of the secondary science (and if it has an adequate theory, its theory as well) are shown to be the logical consequences of the theoretical assumptions (inclusive of the coordinating definitions) of the primary science. (p. 352)

According to the quote, a given set  $L_1$  of laws in the reduced theory S is reduced to the reducing theory T if there is a set  $L_2$  of laws in T by which  $L_1$  is logically entailed. But this reduction is impossible unless the two theories are made intelligible to one another by sharing some common language. Thus, Nagel (1961) puts that

.. if the laws of the secondary science contain terms that do not occur in the theoretical assumptions of the primary discipline (and this is the type of reduction to which we agreed earlier to confine the discussion), the logical

---

<sup>6</sup> This bridge law is standardly taken to be a biconditional law although Nagel himself didn't explicitly say so (Kim 2005, 99). For a detailed discussion, see Kim 1998.

derivation of the former from the latter is *prima facie* impossible. (p. 352)

The situation depicted in the previous quote is like a situation where we need to derive  $R$  from  $Q$  without any further premises but  $Q$  alone. More specifically, suppose that we want to derive  $R$  from  $Q$ . First of all, to do so in a formal derivation,  $Q$  must occur in premises. However,  $Q$  alone doesn't tell anything about any logical relations whatsoever it bears to  $R$ . Such relations are made in virtue of some additional stipulations that occur in the premise. In the current case, we may add in the premise that  $Q \rightarrow R$ . Then from these two premises we can successfully derive  $R$  from  $Q$ . The additional premise " $Q \rightarrow R$ " functions as a bridge law whereby  $R$  is derived from  $Q$ . Moreover, it also functions as a "common language" given that  $Q$  and  $R$  cooccur in the same law. More formally, in Kim's word (2005, p. 98), there must be bridge laws by which each statement of the reduced theory  $M$  is connected with some statement of the reducing theory  $P$ , and that bridge laws have the following form: For any  $x_1, \dots, x_n$ ,  $M(x_1, \dots, x_n)$  if and only if  $P(x_1, \dots, x_n)$ .

However, Nagelian reduction has several problems. The most pernicious one is the claim that higher-order properties can be multiply realized. Originated from Hilary Putnam (1973), multiple realization in our current context is a claim that a given higher-order mental property, say, being in pain, can be realized by various lower-order physical properties. For instance, being in pain,  $M$ , is realized by  $P_1$  in me, but being in pain, the same  $M$ , can be realized in you by  $P_2$ . Moreover, this is possible even within the same individual; my being in pain *now* is realized by  $P_1$ , but my being in pain *yesterday* was realized by  $P_2$ . The point is clear: there is a set of lower-order physical properties  $P_1, P_2, P_3 \dots P_n$  such that the properties belong to this set are all *equal* candidates or realizers of a higher-order mental property  $Q$ .

Thus, multiple realizability naturally threatens the prospect of Nagelian reduction; bridge laws that figure in Nagelian reduction only state that there is one-to-one correlation between a predicate in the reduced theory and a predicate in the reducing theory. Moreover, it also challenges type identity theory given that type identity theory doesn't allow a correlation difference between different species; as was said, pain = C-fiber firing *tout court*.

Facing these difficulties, Kim tries to offer a better, refined model of reduction that avoids the problems mentioned above. Kim's solution is this: functionalize mental properties and reduce them species-specifically.<sup>7</sup> Firstly, Kim (1998) puts that

Functionalism takes mental properties and kinds as functional properties, properties specified in terms of their roles as causal intermediaries between sensory inputs and behavioral outputs, and the physicalist form of functionalism takes physical properties as the only potential occupants, or "realizers," of these causal roles. (p. 19)

In other words, a functionalized mental property is a second-order property, that is, a property of having some property that meets certain causal specifications, and only a physical property in the current context is able to meet those specifications.

So let's take a look at exactly how this reduction based on functionalization is carried out (Kim, 1998, pp. 98-99). Take pain as an example. Being in pain as a mental property typically causes one who experiences it to wince and groan. Again, let  $M$  be being in pain, and  $H$  be its causal specification of it, viz., causing wince and groan. Suppose now that there is a neurophysical property  $P$  such that when it's instantiated, it meets  $H$ .  $M$  then is a property of having  $P$ . And this is nothing more than to say that  $M$  is  $P$ . In this way, a functional reduction of  $M$  to  $P$  is completed.

---

<sup>7</sup> But Kim is not the first to come up with functionalization of mental properties. See, e.g., David Armstrong (1968 and 1981).

However, the thesis of multiple realizability has shown that  $M$  can be realized by some physical properties other than  $P$ . This remark is the very reason Kim speaks of species-specificity; if  $P$  is a reduction base that meets a causal specification for  $M$  for *humans*, then be it a *human* reduction of pain. On the one hand, the accumulation of a sufficient number of empirical data that correlates  $P$  and  $H$  in humans will ultimately be the human-pain reduction of  $M$  to  $P$ . On the other hand, it remains as an open question if other species have different physical properties other than the causal specification of  $M$ . For instance, Martians might have a physical property  $Q$  as that which meets a causal specification  $H$ . Then a survey of the Martians can shed light on the Martian-pain reduction of  $Q$  to  $M$ . In this way the multiple realization that troubled type identity theory can be avoided in Kim's version.

In sum, multiple realizability poses two-fold problems for reduction: first, within the same individual, second, among different species. Functionalization of mental properties solves the first problem by allowing any physical properties that meet a certain causal specification to be a realizer of a mental property. And species-specific reduction solves the second problem by allowing that different species may have a different realization system; in human,  $P_1, P_2, P_3 \dots P_n$  meet a causal specification of a mental property  $M$ , but in octopus,  $Q_1, Q_2, Q_3 \dots Q_n$  might meet  $M$ . Lastly, Kim remains neutral on the possibility of reducing qualia to physical base properties, saying that this is as far as we can get for now: although we don't know if phenomenal properties in the mental are reducible, intentional properties in the mental are functionally reducible to the physical. And this way of reduction doesn't generalize to all species; different species have different reduction.

## CHAPTER 4

### THE EXCLUSION ARGUMENT

#### 4.1 What is the exclusion argument?

The previous chapter elaborated on the central assumptions and claims that underlie physicalism. Chapter 4 to 6 will deal with the main topic of my thesis—the exclusion argument. For the ease of the argument, in this chapter I first present the exclusion argument itself.

The exclusion argument as posed first by Kim (e.g., 1989a, 1989b, 1998, and 2005) is thought to raise some grave challenges to nonreductive physicalism. It is a form of a *reductio* argument which claims that the following nonreductive physicalist assumptions which are essential result in a contradiction (Kim, 2005, pp. 39-43):

- (1) Physical Closure: everything that happens at  $t$  has a sufficient physical cause at  $t-1$ .
- (2) Irreducibility of the Mental: the mental is irreducible to the physical.
- (3) Supervenience:  $A$  *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  $F$ .
- (4) Exclusion: no single event can have more than one sufficient cause occurring at any given time—unless it is a genuine case of causal overdetermination.

To say the conclusion first, Kim thinks that the second assumption, that is, Irreducibility of the Mental that defines nonreductive physicalism, has to go, given that the other three assumptions are non-negotiable.

To demonstrate, suppose *ex hypothesi* that a mental state  $M_1$  causes another mental state  $M_2$ . But Supervenience indicates that  $M_2$  has a (sufficient) physical base, say,  $P_2$  on which  $M_2$  supervenes. This being said, if  $M_1$  were to cause  $M_2$ ,  $M_1$  does so by causing its physical base property  $P_2$  (Kim, 2005, p. 40). More generally, “in order to cause a supervenient property to be instantiated, you must cause one of its base properties to be instantiated (Kim, 2005, p. 20).” This consequently creates a tension as to the occurrence of  $M_2$ ;  $M_2$  occurs either because of the alleged mental cause  $M_1$  or because of its physical realizer  $P_2$ .<sup>8</sup>

Worse yet, there is another problem lurking in this construction. For  $P_2$  should have a sufficient physical cause, say,  $P_1$  by Physical Closure. This means that there is yet another causal tension between  $P_1$  and  $M_1$  as to  $P_2$ . What it shows is that higher-level causation such as mental-to-mental causation implies downward causation (Kim, 2005, p. 40). Now, by Irreducibility of the Mental,  $P_1$  and  $M_1$  are not identical, which makes  $P_2$  causally overdetermined. However, by Exclusion, except genuine causal overdetermination, an effect has only one sufficient cause at a time; one of them has to go. Deciding on which is to go, Physical Closure comes in and favors  $P_1$  as a cause of  $P_2$ ; given that physicalism is taken for granted, the physical comes primary rather than the mental. Consequently, the physical should remain as a sufficient cause. The following summarizes the line of reasoning:

- (1)  $M_1$  causes  $M_2$ .
- (2)  $M_2$  has a supervenient base physical property  $P_2$ . (By Supervenience)

---

<sup>8</sup> One thing to be noticed is that the two appearances of ‘because of’ in this sentence carries a different meaning; when it’s applied to  $M_1$ , it’s interpreted causally; however, when it’s applied to  $P_2$ , it shouldn’t be interpreted causally. This is because a supervenience relation isn’t a causal relation. Naturally,  $M_2$  and  $P_2$  likewise shouldn’t be thought of as causal. I will discuss more on this point later given that this point is quite crucial for my argument in a later chapter. For now, it’s enough to recognize that there is a tension as to  $M_2$ .



- (3)  $M_1$  causes  $M_2$  by causing  $P_2$ . (Given that  $P_2$  is a sufficient for the occurrence of  $M_2$ )
- (4)  $M_1$  is a cause of  $P_2$ . (From (3))
- (5)  $P_2$  has a sufficient physical cause  $P_1$ . (By Physical Closure)
- (6)  $M_1$  is not identical to  $P_1$ . (By Irreducibility of the Mental)
- (7)  $P_2$  is not causally overdetermined. (By (4), (5), (6), and Exclusion)
- (8)  $M_1$  is excluded, and  $P_1$  is the cause. (By Physical Closure and Exclusion)

#### 4.2 Can overdetermination be allowed?

In chapter 3, we saw that Physical Closure and Supervenience are not for sale; they define (minimal) physicalism. So what is at stake are Irreducibility of the Mental and Exclusion. And the exclusion argument claims that the former should go rather than the latter. But one may reasonably question why it should be the case—that is, why isn't Exclusion problematic? One may have noticed that it requires some important assumptions about overdetermination and ontology in general for the argument to succeed. These include assumptions such as that there is no overdetermination in mental causation, and that overdetermination is undesirable. Indeed, the views towards overdetermination are not unilateral; depending on one's ontological picture, some philosophers think that overdetermination is fine, or that mental causation is actually overdetermined (e.g., Roche, 2014). Here I try to persuade the readers why the assumptions made in the argument are more convincing.

More specifically, I will discuss overdetermination macro and micro; at a macro level, I discuss why overdetermination is problematic in an ontological sense; at a micro level, I present one argument that claims that mental causation is actually

always overdetermined, and then evaluate the argument to the effect that it is not successful. These in turn will show that just as Kim suggests, the only viable option for nonreductive physicalists is to give up Irreducibility of the Mental. In the chapter to follow, I will discuss problems related with overdetermination in more detail; I will discuss two different approaches as to overdetermination by which the exclusion argument is challenged. Thus, it will suffice for now to examine overdetermination in terms of its ontological implication and to see why it is problematic.

The idea of overdetermination is that an effect can have two distinct causes. A good example of a genuine overdetermination is two bullets hitting a person each of which is sufficient to cause his or her death. No doubt overdetermination is logically possible; the bullet example above is conceivable without any contradiction and possible to actually happen. The real question is whether mental causation is subject to overdetermination, and if so what the implication is.

Going back to the previous discussion, we have seen that the exclusion argument draws a tension between  $P_1$  and  $M_1$  as to the causal efficacy of  $P_2$ . The previous discussion proceeded as if there are only two options, but in reality several options emerge as to the question “what caused  $P_2$ ?”: (1)  $P_1$  (2)  $M_1$  (3) both  $P_1$  and  $M_1$  (overdetermination), and (4) none of them. (4) is dismissed at first sight because an effect without a cause or joint causes is not conceivable. What then of (3)? Why is overdetermination not an option? The first answer is that we don’t want everything we intentionally do to be overdetermined. This again has to do with ontological parsimony I have talked about before. If our actions are overdetermined, then most of what we do is overdetermined, then most things that matter in this world are overdetermined. The causal explanation of that world would be a messy

one; in that world most meaningful behaviors would involve two overdetermining causes each of which claims its full causal ownership. However, this answer is only sketchy in that it merely appeals to our insight on ontology.

The second answer is that if overdetermination is allowed in mental causation, the physical closure principle is violated (Kim, 2005, pp. 46-50). Suppose that  $P_2$  is (genuinely) overdetermined by  $P_1$  and  $M_1$ . Then  $P_2$  can still happen without the presence of either  $M_1$  or  $P_1$ . For by overdetermination, each alleged cause is sufficient for the occurrence of  $P_2$ . However, this will result in the violation of Physical Closure due to the fact that the occurrence of  $P_2$  with  $M_1$ , but without  $P_1$  would be a physical effect without a physical cause. One possible objection to this line of reasoning is that  $M_1$  without  $P_1$ , or any of its physical realizers is itself inconceivable. For we have seen that by Supervenience,  $M_1$  occurs because  $P_1$  occurs. However, this rejection renders the problem moot in the first place; our initial supposition was that  $P_1$  and  $M_1$  are *each sufficient* causes of  $P_2$ . But invoking supervenience makes the distinctness of  $P_1$  and  $M_1$  impossible. The point is that if we want to say that  $P_1$  and  $M_1$  are each sufficient causes of  $P_2$ , then it should be possible to imagine a situation in which either of them is absent in the picture. Then this leads to  $M_1$  being a cause of  $P_2$  *on its own* without the presence of any physical bases. And in that case, Physical Closure is violated.

## CHAPTER 5

### OBJECTIONS TO THE EXCLUSION ARGUMENT

#### 5.1 Challenges to the exclusion argument

Up until now, the foregoing chapters have been mostly introductory. In this and the next chapter, issues and questions that are germane to the main topic of this thesis will be presented; this chapter will be dedicated to the counterfactual solutions to the exclusion argument, and the next one will be criticisms of them. It is important to notice that all the solutions that will be presented here aim at falsifying the principle of overdetermination in Exclusion; one way or another those who try to save mental causation argue either that there is no overdetermination of a problematic sort in mental causation or that there is no overdetermination in mental causation. My main target is those who invoke counterfactuals to defeat Overdetermination by the first method. It is worth noting why I take the counterfactual solutions as of great importance above all: first, the counterfactual solutions have been the most popular among nonreductive physicalists. Second, whereas other solutions to the exclusion argument tend to be limited to solving the problem raised by the argument, the counterfactual solutions take the discussion further—it raises the question of the concept of causation in general. More specifically, from the counterfactualist discussion of mental causation there arise questions regarding causation as production and as dependence. I take the issue of causation as of great importance in that the issue I'm dealing with here is subsumed under a broad discussion of causation. The issue of causation as production and as dependence will recur briefly

in the last chapter where I will be drawing the lessons and implications from the foregoing chapters.

In any case, I plan first to briefly cover some other alternative approaches other than the counterfactual solutions. For instance, Michael Roche (2014) argues that mental causation is always overdetermined. His point turns on the claim that Kim's interpretation of overdetermination leaves it open that dependent events (or causes) can overdetermine an effect (Roche, 2014, p. 816). And dependent-overdetermination doesn't cause the ontological problem mentioned above; the ontological inconceivability and implausibility from overdetermination are warranted only in so far as two independent causes overdetermine an effect (Roche, 2014, p. 817). However, the mental causation is obviously not the case involving two independent causes given that the mental and the physical are tied by supervenience.

A couple of points as to Roche's claim are in order. First of all, I don't see why Roche takes it that Kim allows dependent overdetermination. The fact that Kim (2005, p. 42) constructs Overdetermination as saying that *unless* it is a genuine case of overdetermination, no effect has two sufficient causes at once *ipso facto* excludes dependent overdetermination in Kim's picture. On many occasions Kim clearly says that what he means by a "genuine case of overdetermination" is cases like two *causally independent* bullets hitting the same victim each of which is sufficient to cause his or her death. So we can substitute "genuine" with "causally independent". I think this much is quite clear from his construction of Overdetermination. However, Roche goes further to point out that the fact that Kim uses an additional argument for the claim that there is no overdetermination is a reason to believe that Kim doesn't take the implication just mentioned (Roche, 2014, p. 816). In any case, even if Roche

is right about Kim's interpretation of overdetermination and thus we take it that Kim allows dependent overdetermination, Roche's argument still fails. For suppose that Roche is right. Then the following line of reasoning is in order:

- (1)  $M$  supervenes on  $P$  (by Supervenience).
- (2) then,  $M$  is either dependent or independent of  $P$ .
- (3) if independent, then genuine overdetermination holds. So Roche fails to prove that the overdetermination in mental causation is not a problematic sort.
- (4) If dependent, then non-genuine overdetermination holds. So Roche succeeds in proving that the overdetermination in mental causation is not a problematic sort.

I argue that (4) fails due to its negligence of the causal inheritance principle endorsed by Kim. Whatever overdetermination we have in mind, the idea of two different things causing the same effect makes sense only if each has some distinct causal role to play in the occurrence of the effect. However, this doesn't apply to the issue at hand given that  $M$  supervenes on  $P$ . In other words, just like the dormitivity of a sleeping pill whose causal power is nothing more than the causal power of its chemical reactions,  $M$ 's causal power is nothing more than  $P$ 's causal power. The point seems clear: whatever dependent overdetermination is, it holds only if each dependent cause has some distinct causal power. But in the current case,  $M$  has no new causal power other than the one inherited from  $P$ , so the argument fails.

Now I would like to move on to the main argument of this chapter—the counterfactual solutions to the exclusion argument. As I said earlier, there are different versions of the counterfactual solutions by different nonreductive physicalists. In this chapter I intend to lay out some of the well-known versions of it and to extract the common core that is prevalent in those versions. In the next chapter

I will argue that this common core is subject to various criticisms, making the counterfactual solutions not suitable as a solution to the exclusion argument.

## 5.2 Versions of the counterfactual solution

### 5.2.1 Innocent or innocuous overdetermination

The first version of the counterfactualist solutions I want to present is that of Karen Bennett (2003 and 2008). I believe her version is the most straightforward and thus a good way to start the discussion. But as I will argue later, her version or any other versions of the counterfactual solutions that will appear in this thesis argue in more or less the same fashion, so simplicity won't do harm.

I said earlier that there is a difference between *genuine* and *non-genuine* overdetermination—the difference in independency or dependency in the two overdetermining causes. Bennett takes this point further and proposes the two necessary conditions for genuine overdetermination. For her, causes overdetermine the effect iff the following are non-vacuously true (Bennett, 2008, p. 288):

$$(OC1) (P_1 \wedge \sim M_1) \Box \rightarrow P_2.$$

$$(OC2) (M_1 \wedge \sim P_1) \Box \rightarrow P_2.$$

The first thing to notice is her use of the term related with vacuity. The talk of vacuity is prevalent in Lewisian counterfactual analysis, and this clue hints at the importance of Lewisian counterfactuals for counterfactualists. Accordingly, we would need to look at what Lewisian counterfactual analysis is, and this issue will take some considerable portion of my discussion in a later chapter. In any case, non-vacuity here means a simple rule in logic: on the one hand,  $P \rightarrow Q$  is vacuously true iff  $P$  is always false so that  $P \rightarrow Q$  can always be true regardless of the truth

value of the antecedent  $P$ ;  $P \rightarrow Q$ , on the other hand, is non-vacuous iff the antecedent  $P$  is possible to be either true or false and consequently affects the truth value of  $P \rightarrow Q$ .

Bennett argues that overdetermination of a problematic sort is the one that meets those two necessary conditions. If an alleged case of overdetermination doesn't meet those conditions, then that case is "innocent", causing no ontological or metaphysical problem that bothers our intuition. Bennett argues further that the case of overdetermination between a mental and a physical property in the exclusion argument is one of those innocent cases. Consequently, her point undermines Kim's whole project in that Kim's exclusion argument derives its force from the claim that the case of a mental and a physical property involves overdetermination of a problematic sort.

To see how the necessary conditions work, imagine again two bullets hitting one single target, each of which is sufficient to cause his or her death. Let the first bullet be  $P_1$  and the second  $P_2$ . Also let  $Q$  denote the death of the target of the bullets. We may then say that the overdetermination involved in this case is genuine given that (OC1) and (OC2) are non-vacuously true. For suppose a situation in which the second bullet was missing—that is, suppose a  $(P_1 \wedge \sim P_2)$ -world. Although  $P_2$  is absent in that world,  $P_1$  is sufficient to cause  $Q$ . And the same goes for (OC2); a  $(\sim P_1 \wedge P_2)$ -world also guarantees the death of the target. Given that each antecedent can be true in some world and affect the truth value of the consequent, the firing squad example turns out to be a genuine case of overdetermination.

However, the case of mental causation doesn't involve genuine overdetermination; rather, it involves non-genuine, thus innocent overdetermination.



To see why, let us evaluate the causal competition between  $P_1$  and  $M_1$  vis-à-vis  $P_2$  along with the necessary conditions. Starting with (OC2),  $(M_1 \wedge \sim P_1) \square \rightarrow P_2$  seems non-vacuous. For imagine a possible world in which  $M_1$  is present but  $P_1$  is not. It is easy to see that by multiple realization, in that world there would be a different realizer of  $M_1$ . Then in the exclusion argument,  $M_1$  is supposed to be a sufficient cause of  $P_2$ , so (OC2) is non-vacuously true. The problem rather lies in (OC1). Just like (OC2), the non-vacuity of (OC1) depends on its antecedent. And for that to be the case, we should be able to imagine a possible world in which  $M_1$  is not present but  $P_1$  is. However, such a world seems impossible. For supervenience stipulates that it is metaphysically necessary that if the supervenient base is present, then the supervening property is present as well. In other words, because of the supervenience relation between  $M_1$  and  $P_1$ , and of the asymmetric nature between the supervening and subvenient properties, any  $P_1$ -worlds are  $M_1$ -worlds (or any non- $M_1$  worlds are non- $P_1$  worlds) *tout court*.<sup>9</sup> This consequently shows that (OC1) is vacuously-true; the antecedent is always false so the whole conditional can be true or false regardless of the antecedent. Consequently, the structure involving mental and physical properties tied by supervenience doesn't give rise to the problem of overdetermination since the alleged causal overdetermination between  $M_1$  and  $P_1$  as Kim wants doesn't arise according to Bennett's account of overdetermination.

To summarize, Bennett argues against Kim's implicit assumption on overdetermination between  $M_1$  and  $P_1$  that the overdetermination involved there is

---

<sup>9</sup> Of course, there could be possible worlds in which the supervenience relation between  $P_1$  and  $M_1$  fails to hold. But what is the point of talking about such worlds in the first place? Those worlds would be quite remote from the actual world in the sense that the nomological relation doesn't apply to the case at issue. Talking about such worlds would be neither productive nor meaningful. See Crisp and Warfield (2001) for a similar remark.

a problematic sort—a genuine, non-vacuous overdetermination. However, Bennett shows that Kim’s assumption is unwarranted provided that the overdetermination at issue doesn’t meet her two necessary conditions for a genuine, non-vacuous overdetermination. Since the overdetermination between  $M_1$  and  $P_1$  doesn’t meet the necessary conditions, it is an innocent overdetermination.

Additionally, before closing this section, I’d also like to mention Jesper Kallestrup’s work (2009) on this issue, since virtually the same version of the counterfactual solution can be found in it. His formulation of necessary and sufficient conditions for a genuine overdetermination is almost identical to Bennett’s

(Kallestrup, 2009, p. 471): E is overdetermined by C and C\* iff,

- (i) C is sufficient for E,
- (ii) C\* is sufficient for E,
- (iii) if C occurred without C\*, E would have occurred, and
- (iv) if C\* occurred without C, E would have occurred.

It is easily seen that the (iii) and (iv) are nothing but (OC1) and (OC2) in Bennett’s account. The parallel in the two conditions for a genuine overdetermination leads to the parallel in the consequence as well: just like Bennett’s argument, Kallestrup also points out that since (iii) or (iv) is vacuously true in the mental causation case, it doesn’t pose any serious metaphysical problem. Kallestrup (2009) thus concludes that “.. this commitment [to the overdetermination conditions mentioned above] is innocuous if either (iii) or (iv) is vacuously true. The nonreductive physicalist need therefore only endorse an acceptable form of overdetermination (p. 472).” Since Bennett’s and Kallestrup’s argument are almost identical, it will be shown that both are subject to the criticisms that will follow up later in this thesis.

### 5.2.2 List and Menzie's causation as difference-making

The second version of the counterfactual solutions is found in Christian List and Peter Menzie's causation as difference-making (2009). According to their view, the view of causation implicit in the exclusion argument leads to an unwelcoming result—that is, the cause becomes too specific and disproportionate to the effect.

To see what they mean by the proportionality between the cause and the effect, think of Stephen Yablo's red-color-pecking bird example (1992). Suppose a bird is trained to peck at only red objects. Suppose also that after the training, the bird pecks at a red object nearby, which happens to be a crimson object. The question now is: what caused the bird to peck at that object—being red or being crimson? The exclusion principle states that if a subvenient cause (or property) *P* is sufficient to cause the effect *E*, then any supervening cause (or property) *Q* of *P* is excluded from being the cause of *E* (List and Menzies, 2009, p. 478). Provided that being red supervenes on being crimson, the exclusion principle implies that the bird pecked at the object because of its being crimson, instead of being red. However, “the target's being crimson is too specific to count as the cause: citing it as the cause of the pecking might give the erroneous impression that the pigeon would not peck at anything non-crimson (List and Menzies, 2009, p. 480).” In other words, the exclusion principle applied to any supervenience relation yields an unwelcoming consequence: a cause gets too specific and as a consequence it ignores the multiple realizability of the effect.

They argue that the reason why a cause must be proportional to its effect derives from the dictum that “changing the causal property from being absent to

being present (or vice versa) changes the effect property from being absent to being present (or vice versa) (List and Menzies, 2009, p. 480).” For we have seen that the bird example shows that if the causal property (being crimson) had been changed from being present to being absent, and instead some other causal property similar to the original one, like being scarlet, had been present the effect property would not have changed—it would have been still present. On the other hand, if we invoke being red as the cause of the bird’s pecking, there arises no problem: changing being red from being present to being absent would prevent the bird’s pecking at the object from being present.

They openly admit that this way of viewing causation is prevalent throughout different conceptions of causation—counterfactual, probabilistic, interventionist and contrastive ones (List and Menzies, 2009, p. 480). However, the counterfactual solutions being the main target of this thesis, I will zero in on the counterfactual view of causation. The discussion so far shows that for List and Menzies specificity determines a real cause. And how specific is enough is determined by questioning if an alleged cause makes a difference to its effect; if it makes a difference as to the effect while keeping multiple realizability intact, then it is a real cause; if it doesn’t make a difference as to the effect and ignores multiple realizability, then it is not.

Formally, List and Menzies (2009) put the following as the truth conditions for making a difference:

Truth conditions for making a difference: The presence of  $F$  makes a difference to the presence of  $G$  in the actual situation just in case (i) if any relevantly similar possible situation instantiates  $F$  it instantiates  $G$ ; and (ii) if any relevantly similar possible situation instantiates  $\sim F$ , it instantiates  $\sim G$ . (p. 481)

In my interpretation, their conception of a “proportional difference-making” cause means the following:  $P$  is a difference-making cause of  $E$  iff (1)  $P \square \rightarrow E$  and (2)  $\sim P$

$\square \rightarrow \sim E$ . As to the proportionality, (3)  $P$  is a proportional cause of  $E$  iff taking  $P$  to be the cause of  $E$  doesn't violate multiple realizability (call it  $(PE)$  for brevity). For instance, crimson is not a proportional cause of the bird's pecking because it doesn't allow being non-crimson-but-other-shade-of-red to be the cause of the bird's pecking. We can easily expand it such that in the case of mental overdetermination, a difference-making cause of the effect is that which meets (1)  $P \square \rightarrow E$ , (2)  $\sim P \square \rightarrow \sim E$ , and  $(PE)$ .

To see why, apply difference-making conditions to the two supposed causes of  $P_2$ — $P_1$  and  $M_1$ . Then  $P_1$  is a difference-making cause of  $P_2$  iff  $P_1 \square \rightarrow P_2$  and  $\sim P_1 \square \rightarrow \sim P_2$ . The same is true of  $M_1$ :  $M_1$  is a difference-making cause of  $P_2$  iff  $M_1 \square \rightarrow P_2$  and  $\sim M_1 \square \rightarrow \sim P_2$ . Before evaluating these conditions, there needs to be some modification of the formula. For given that we are dealing with the case that involves supervenience,  $P_1$  and  $M_1$  shouldn't be evaluated independently, which means that I should also consider  $M_1$  when evaluating  $P_1$  and vice versa; to the effect that when I consider if  $P_1$  is a proportional difference-making cause of  $P_2$ , I suppose that if  $P_1$  without  $M_1$  can cause  $P_2$  and vice versa for  $M_1$ , which is equivalent of Bennett's overdetermination conditions. So the causal evaluation of  $P_1$  as to  $P_2$  should be the following:<sup>10</sup>

$$(1) (P_1 \wedge \sim M_1) \square \rightarrow P_2$$

$$(2) (\sim P_1 \wedge \sim M_1) \square \rightarrow \sim P_2$$

The second condition seems to have no problem; if none of the alleged causes occur,

---

<sup>10</sup> And it is easy to see that the second condition of Bennett's genuine overdetermination follows from evaluating  $M_1$ :  $(M_1 \wedge \sim P_1) \square \rightarrow P_2$ . But I skip this simply because the falsification of genuine overdetermination in the current case only requires that one of the necessary conditions for genuine overdetermination fails. So showing that  $(P_1 \wedge \sim M_1) \square \rightarrow P_2$  is vacuously true would be enough for the present purpose.

then their effect doesn't occur either. However, the same reasoning found in Bennett can be invoked for the first condition here; recall that in Bennett's account, a case of overdetermination is innocent if it's not the case that two necessary conditions for genuine overdetermination hold. It is easy to see that in a proportional difference-making view of causation, one of the conditions, viz.,  $(P_1 \wedge \sim M_1) \square \rightarrow P_2$ , is not met; this conditional is vacuously true given that the antecedent is metaphysically impossible given the supervenience between  $P_1$  and  $M_1$ .<sup>11</sup>

In summary, I have dealt with a proportional difference-making view of causation found in List and Menzies and have shown that their strategy can ultimately boil down to that of Bennett.

---

<sup>11</sup> One can at this point raise a question if there are some possible worlds in which  $(P_1 \wedge \sim M_1) \square \rightarrow P_2$  is true. This is reasonable but as will be shown in the next chapter, Lewisian counterfactual semantics will reveal that possible worlds in which  $(P_1 \wedge \sim M_1) \square \rightarrow P_2$  is true aren't worth discussing and thus negligible given that those worlds are too far-fetched from the actual world in terms of laws and regularities.

## CHAPTER 6

### OBJECTIONS TO THE COUNTERFACTUAL SOLUTIONS

In this chapter, I will present three criticisms that are brought against the counterfactual solutions:

- (1) Counterfactual analysis of the supervenience relation between the mental and the physical gives rise to semantic emptiness as to the antecedent of the counterfactual conditional of the supervenience relation (Harbecke, 2014).
- (2) Counterfactual causation actually implies the exclusion argument (Zhong, 2011, 2014, and 2015).
- (3) Counterfactual causation mistakenly views supervenience as a causal relation.

I argue that each criticism independently undermines the credibility of the counterfactual solutions.

#### 6.1 Counterfactuals and semantics

So far when we talked about overdetermination conditions, we have assumed all along that  $(\sim P_1 \wedge M_1) \rightarrow P_2$  is non-vacuously true because of the multiple realizability of  $M_1$ ; even if  $\sim P_1$ , there would be some other physical realizer of  $M_1$  such that there can still be  $M_1$  and thus it would anyway result in the occurrence of  $P_2$ . The first criticism will take up this issue and examine whether such an assumption is unproblematic.

First of all, the counterfactual evaluation of supervenience relation has so far allowed that the occurrence or nonoccurrence of  $P_1$  doesn't bear any significant

impact on the occurrence of  $P_2$  because  $P_2$  can still occur by the help of some other physical realizer, say,  $P_3$ . The main motivation behind this “replacement strategy” is the idea that mental properties are multiply physically realizable. And I agree with the claim that multiple realizability is probably true both metaphysically and empirically. However, once approached by a counterfactual perspective, multiple realizability raises some issues related with semantics. To see why, I need to deal with David Lewis’s model of counterfactuals (1973). According to the standard Lewisian counterfactuals, the actual world  $w$  contains an ordered set of worlds such that each world is nested around  $w$  either closely or remotely, depending on the similarities it bears to  $w$ . And a possible world  $x_1$  is closer or more similar to  $w$  than a possible world  $x_2$  just in case  $x_1$  matches more with  $w$  than  $x_2$  in terms of the facts and the natural laws holding in it. In this system, the truth or falsity of a counterfactual statement in  $w$  is decided by the evaluation of the closest possible worlds that are similar to  $w$ : for instance  $P \Box \rightarrow Q$  is true iff in all the closest possible worlds to the actual world  $w$ , if  $P$  holds then  $Q$  holds as well; in other words, if all the  $P$ -worlds that are accessible from  $w$  are also  $Q$ -worlds,  $P \Box \rightarrow Q$  is true.

Now we may give an independent reason why  $(P_1 \wedge \sim M_1) \rightarrow P_2$  is vacuously true and thus doesn’t hold. For given the actual world,  $(P_1 \wedge \sim M_1)$ -worlds are more remote than  $(P_1 \wedge M_1)$ -worlds. Provided that there is a nomological law holding between  $P_1$  and  $M_1$  in the actual world  $w$ , the closest possible worlds would naturally be the ones in which the same nomological law holds. For the same reason, all  $(\sim P_1 \wedge M_1)$ -worlds are less close to the actual world than  $(Q_n \wedge M_1)$ -worlds where  $Q_n$  is some other physical realizer of  $M_1$ . So in the overdetermination case,  $\sim P_1 \Box \rightarrow \sim P_2$  would always be false given that the possible



world in which  $P_1$  is replaced by some other realizer of  $M_1$  that causes  $P_2$  is closer to the actual world than one where in the absence of  $P_1$ ,  $P_1$  isn't replaced, for that would mean that in that world the nomological supervenience fails.

However, as Jens Harbecke rightly points out, this replacement strategy causes the “emptiness problem”. Causation based on counterfactuals is built on the notion of causation as difference-making (List and Menzies, 2009, p. 476); for A to cause B is tantamount to saying that had A not happened, B wouldn't have happened either. And surely, we invoke counterfactuals in this way to distinguish real causes from faulty ones to the effect that we can *meaningfully* and *truly* say that B counterfactually depends on A. But the above reasoning on counterfactuals shows that a meaningful discussion of counterfactual dependence seems to be jeopardized in mental causation. Recall that according to the current model of counterfactuals,  $\sim P_1 \Box \rightarrow \sim P_2$  is false given that there may be  $Q_n$  that realizes  $M_1$ , which then results in  $P_2$ . In other words,  $(\sim P_1 \wedge M_1) \Box \rightarrow P_2$  would always be true as far as the supervenience is considered. (And this should be considered important because in the current context, this supervenience differentiates accessible worlds from inaccessible, and close worlds from remote worlds).  $M_1$  would be realized by some other realizer in the closest possible world to the actual world, and this causes the emptiness problem. For there will be no point in saying “had  $P_1$  not occurred.. and so on” because it wouldn't affect a bit about the truth of the consequent  $P_2$ ; whatever we talk about,  $P_2$  will happen anyway. Counterfactual semantics is supposed to help us talk meaningfully about causes and effects, but counterfactual semantics plus supervenience seems to cause semantic vacuity. So the problem is about the compatibility between the semantics of counterfactuals and the counterfactual

evaluation of supervenience. Given that the counterfactual evaluation of supervenience is the default set up for the counterfactualists, we should examine if there is any way out for them.

It seems to me that the only way for the counterfactual evaluation of supervenience to be meaningful is to lower or expand the accessibility condition, i.e., to allow possible worlds in which the supervenience between  $P_1$  and  $M_1$  and multiple realizability fail. However, doing so would make it the case that  $(P_1 \wedge \sim M_1) \Box \rightarrow P_2$  is non-vacuous because some possible worlds in which  $M_1$  fails to supervene on  $P_1$  or to have an alternative physical realizer are allowed as the result of lowering or expanding the accessibility condition. But this move is flawed because this will result in the overdetermination at issue becoming *guilty*, which is against the initial claim of the counterfactualists.

The initial argument for the counterfactual solution lies in the claim that overdetermination in mental causation is innocent because it doesn't meet one of the necessary conditions for a genuine overdetermination—that is, (OC1), viz.,  $(P_1 \wedge \sim M_1) \Box \rightarrow P_2$ , is vacuously true due to the fact that the occurrence of  $P_1$  without  $M_1$  is metaphysically impossible when there is a nomological law regarding the supervenience of the latter on the former. And the reason why the overdetermination at issue doesn't meet the condition turns on supervenience and multiple realizability; because of the theses of supervenience and multiple realizability, a  $(P_1 \wedge \sim M_1)$ -world is more remote than  $(Q_n \wedge M_1)$ -worlds, where  $Q_n$  stands for some other alternative physical realization base of  $M_1$ . I believe this consequently points out that there is no way out; the semantic vacuity is still present.

## 6.2 Counterfactuals and downward causation

I said earlier that there are several versions of the exclusion argument, each of which invokes different assumptions. As these versions make use of different premises, some versions are more vulnerable to the counterfactual solutions. However, there is still one version of the argument that survives the counterfactual solutions—a version of the exclusion argument with the principle of downward causation. What's more, it rather points out that the counterfactual approach *implies* the exclusion argument (Zhong, 2011 and 2014). Here I elaborate on Zhong's argument and try to further analyze the implication.

To begin with, there are essential premises that underlie all versions of the exclusion argument, so let me spell out those so that we can concentrate rather on the problematic parts. Although these premises are equivalent to the ones found in Kim, for the ease of the argument I follow Zhong's words (2011):

(S) *Supervenience*: Mental properties (among other higher-order properties) supervene on physical properties. That is, if any system *s* instantiates a mental property *M* at *t*, there necessarily exists a physical property *P* such that *s* instantiates *P* at *t*, and necessarily anything instantiating *P* at any time instantiates *M* at that time.

(NO) *Non-overdetermination*: No single event can have more than one sufficient cause occurring at any given time—unless it is a genuine case of causal overdetermination. And there is no systematic overdetermination in cases of mental causation.

(I) *Irreducibility*: Mental properties are not identical with physical properties. (pp. 131-132)

All the premises above should be by now familiar for I have explained multiple times throughout the discussion. I'll simply skip them assuming that we all know what those mean and imply. These three premises being essential ingredients in all three versions of the exclusion argument, those versions differ in that different additional premises are added to these essential premises so as to arrive at the same conclusion: *M* is excluded by *P*. So let me dig into the real trouble. It will be convenient first to

introduce a figure that help depict the trouble lurking in the exclusion argument.

Figure 1 depicts what kind of causal and supervenience relations the physical and mental properties in the exclusion argument are bounded by.

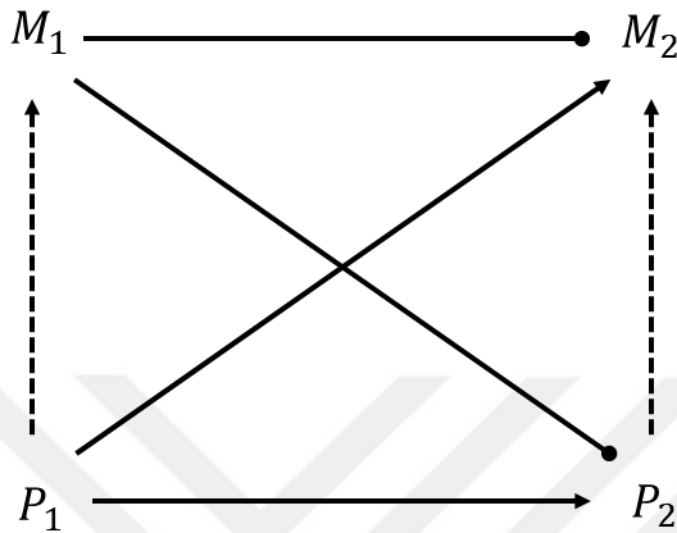


Figure 1 Depiction of causal and supervenience relations in the exclusion argument.

First of all, the dotted arrows reflect the supervenience relations between  $P_1$  and  $M_1$ , and between  $P_2$  and  $M_2$ . Next, a round-headed normal line stands for the alleged mental causation of  $M_1$  as to  $M_2$  and  $P_2$ . Lastly, arrow-headed normal lines are to show the causal relation between  $P_1$  and  $P_2$ , and between  $P_1$  and  $M_2$  respectively. Zhong claims that among three versions of the exclusion argument the first two are vulnerable to the counterfactual approach whereas the third one resists it. Let us first review the first two versions. Although Zhong separates these two versions and deals with them one by one, I will deal with them at once for I believe they diverge only because the viewpoint from which Fig. 1. is viewed is different.

Firstly, the first version of the exclusion argument invokes *Causal Inheritance Principle* (CI) according to which the causal power of any supervenient

properties is derived from their base properties.<sup>12</sup> In this version  $M_1$ 's causal power is excluded by  $P_1$  for the following line of reasoning: if  $M_1$  causes  $M_2$ , then its causal power is derived from  $P_1$  by (CI). Then given (NO) and (I), there is no room for  $M_1$  to play any causal role as to the occurrence of  $M_2$ ; all the work can be done and explained by  $P_1$ . However, Zhong argues that the exclusion argument with (CI) is faced with a counterexample involving multiple realizability and the closeness of possible worlds to the actual world. To see how the first version of the exclusion argument fails, Zhong puts some detailed example to elaborate on this point, but I will simply skip that because I believe there is a shorter and more intuitive explanation: a simple reflection on multiple realizability and the closeness of various possible worlds to the actual world already tells us why the first version of the exclusion argument doesn't work.

We know from supervenience and multiple realizability that other than  $P_1$ ,  $M_1$  can have various physical realizers—call it  $Q_n$ . Then this simple implication under counterfactual analysis shows that the first version that invokes (CI) is false. Counterfactually speaking, suppose  $M_2$  occurs in the actual world. Then any possible worlds in which  $P_1$  doesn't occur and  $M_2$  doesn't occur as well are more remote from the actual world than possible worlds where  $M_2$  occurs in virtue of some other physical realizer  $Q_n$ . In other words,  $((Q_n \wedge M_1) \rightarrow M_2)$ -worlds are closer to the actual world than  $((\sim P_1 \wedge \sim M_1) \rightarrow \sim M_2)$ -worlds. And  $(\sim M_1 \rightarrow \sim M_2)$ -worlds are closer to the actual world than  $(\sim M_1 \rightarrow M_2)$ -worlds. For the nonoccurrence of  $M_1$  means that there is no other physical realizer such as  $P_1$  and  $Q_n$ ; that is, there is no physical base for  $M_2$ . Lastly, possible worlds in which the

---

<sup>12</sup> For a detailed discussion of *Causal Inheritance Principle*, see Kim (1995).

occurrence/nonoccurrence of  $M_1$  has causal bearing on  $M_2$  are closer to the actual world than possible worlds in which the occurrence/nonoccurrence of  $P_1$  has causal bearing on  $M_2$ . This in turn shows that in the current case,  $M_1$  doesn't inherit its causal power from  $P_1$  and is a more appropriate cause of  $M_2$ .

The second version of the exclusion argument which invokes upward causation faces a similar problem. According to Zhong (2011), the second version, instead of (CI), adds two additional premises:

- (UC) *Upward Causation*: If property  $A$  causes property  $B$ , then  $A$  would cause any supervenient property of  $B$  instantiated on this occasion.
- (CCP) *Causal Completeness of Physics*: If a physical event has a cause that occurs at  $t$ , it has a (sufficient) physical cause that occurs at  $t$ . (p. 135)

Recall Fig. 1. We see that  $M_2$  has its physical realizer  $P_2$ . Then (CCP) stipulates that if  $P_2$  has a cause, then that cause must be a sufficient physical cause— $P_1$ . And if  $P_1$  is a cause of  $P_2$ , then by (UC),  $P_1$  is also a cause of  $P_2$ 's supervenient property, viz.,  $M_2$ . So there is no room for  $M_1$  as a cause whatsoever; given the aforementioned (NO), according to which except genuine overdetermination no event has two distinct and sufficient causes,  $M_1$  can't participate in a causal transaction because  $P_1$  is already at work.

However, the second version is also vulnerable to the criticism based on counterfactual analysis. This time too, there is a case in which  $P_1$  isn't present, but  $M_2$  is, given multiple realizability. If there is  $M_2$ , then by supervenience, there is  $P_2$  that realizes  $M_2$ . (CCP) then comes in to point out that  $P_2$  has a physical cause  $P_1$ . Lastly, by (UC), if  $P_1$  causes  $P_2$ , then  $P_1$  causes any of  $P_2$  supervenient properties, including  $M_2$ . However, does this line of reasoning entail that  $P_1$  is a cause of  $M_2$ ? The answer is no. In order for  $P_1$  to be a cause of  $M_2$  under counterfactual analysis, it should be the case that if  $P_1$  hadn't occurred,  $M_2$

wouldn't have occurred either. But the nonoccurrence of  $P_1$  doesn't guarantee the nonoccurrence of  $M_2$ . Given that multiple realizability holds,  $M_2$  can have  $Q_n$  instead of  $P_1$  as a cause; if  $P_1$  had not been present,  $M_1$  would still have been present. In this sense,  $P_1$  is not a cause of  $M_2$ .

Lastly, there is a version of the exclusion argument whose premises invoke what's called the principle of *Downward Causation* (DC), according to which "if property  $A$  causes property  $B$ , then  $A$  must cause any base property of  $B$  instantiated on this occasion (Zhong, 2011, p. 138)." This is the version that is immune to the counterfactual approach. But I want to link this version with Kim's latest version of the exclusion argument. Let me first put Zhong's explanation of (DC) with a slight modification (2011):

A higher-order property  $M_2$  is supposed to be caused by a mental property  $M_1$ . From the Supervenience Principle, we can know that  $M_2$  has some physical property  $P_2$  as its supervenient property on this occasion. Then according to the Downward Causation Principle, if  $M_1$  is a genuine cause of  $M_2$ ,  $M_1$  must be a cause of  $P_2$ —that is,  $M_1$  causes  $M_2$  by causing  $M_2$ 's base property  $P_2$ . (p. 139)

That  $M_1$  causes  $M_2$  by causing  $M_2$ 's base property  $P_2$  is the key to the linkage between the third version and Kim's latest version of the exclusion argument. For exactly the same principle of downward causation appears in Kim as he says "in order to cause a supervenient property to be instantiated, you must cause one of its base properties to be instantiated (Kim, 2005, p. 20)."

Now, suppose that  $M_1$  causes  $M_2$ . Then under the counterfactual approach,  $M_1$ 's causing  $M_2$  is identical to the claim that  $M_2$  counterfactually depends on  $M_1$ . Next, given  $M_2$  is instantiated in this case, supervenience (S) puts that  $M_2$  has its physical realizer, viz.,  $P_2$ . That being said, it trivially follows from (S) that the following two conditionals hold: (i)  $\sim M_2 \square \rightarrow \sim P_2$ , and (ii)  $P_2 \square \rightarrow M_2$ . From

this, we may say that  $P_2$  is also counterfactually dependent on  $M_1$ . For if  $\sim M_1 \square \rightarrow \sim M_2$ , and  $\sim M_2 \square \rightarrow \sim P_2$ , then by transitivity,  $\sim M_1 \square \rightarrow \sim P_2$ . From this we infer the Downward Causation Principle—that  $M_1$  causes  $P_2$  whereby it causes  $M_2$  as well. However,  $P_2$  has a sufficient physical cause  $P_1$ , so  $M_1$  is excluded, i.e.,  $M_1$  doesn't cause  $P_2$ . And if  $M_1$  doesn't cause  $M_2$ 's physical base property  $P_2$ , then  $M_1$  doesn't cause  $M_2$ . In a logical form, the argument can be summarized as follows (Zhong, 2011, p. 141) (slight modification as to the letters.):

- (1) If  $M_1$  causes  $M_2$ , then  $M_2$  counterfactually depends upon  $M_1$ ;
- (2)  $M_2$  is realized by  $P_2$ ;
- (3) So, any  $(\sim M_2)$ -world is also a  $(\sim P_2)$ -world, and any  $(P_2)$ -world is also an  $(M_2)$ -world;
- (4) So, if  $M_2$  counterfactually depends upon  $M_1$ , then  $P_2$  also counterfactually depends upon  $M_1$ ;
- (5) So, if  $M_2$  counterfactually depends upon  $M_1$ , then  $M_1$  causes  $P_2$ ;
- (6) (From (1) and (5)), if  $M_1$  causes  $M_2$ , then  $M_1$  causes  $P_2$  (i.e., the *Downward Causation Principle*);
- (7) (From the Completeness Principle),  $P_2$  is caused by  $P_1$ ;
- (8) (From (7), the Non-overdetermination Principle and the Irreducibility Principle),  $M_1$  doesn't cause  $P_2$ ;
- (9) (From (6) and (8)), therefore,  $M_1$  doesn't cause  $M_2$ .

### 6.3 Counterfactuals and supervenience

The final criticism concerns the implication the counterfactual approach bears to mental and physical properties grounded in supervenience. The interim conclusion is



that counterfactual analysis of causation *isn't* applicable to mental causation. This will be shown by the analysis of the two views on causation—production and dependence views (Hall, 2004). I argue that a simple analysis on what causation is in the counterfactual approach reveals that counterfactualists mistakenly, and more importantly, unknowingly interpret supervenience as a causal relation.

Start with the production and dependence views of causation. Hall (2004) defines each in the following passage:

.. [Dependence] is simply that: counterfactual dependence between wholly distinct events. In this sense, event *c* is a cause of (distinct) event *e* just in case *e* depends on *c*; that is, just in case, had *c* not occurred, *e* would not have occurred. The second variety [production] is rather more difficult to characterize, but we evoke it when we say of an event *c* that it helps to *generate* or *bring about* or *produce* another event *e*, and for that reason I call it “production”. (p. 225)

Leaving the production view aside for later discussion, the dependence view is our current target. The first thing to consider is what he means by distinct events because the analysandum of the counterfactual analysis at hand is grounded in a special sort of relation, that is, supervenience. Thus, attention is required as to what makes mental and physical events distinct. A detailed discussion of the distinctness of events will unfold in the next section as this issue has to do with the conditions for causal and noncausal counterfactual dependence. But let's say for now that mental and physical events are (qualitatively) distinct in that one event instantiates a mental property while the other instantiates a physical property. At any rate, the next sentence can be taken at face value: *c* is a cause of *e* iff  $c \square \rightarrow e$  and  $\sim c \square \rightarrow \sim e$ . More generally, the following thesis holds (Hall, 2004, p. 225):

Dependence: Counterfactual dependence between wholly distinct events is sufficient for causation.

The idea that such counterfactual dependence defines causation is “the cornerstone of

every counterfactual analysis” (Hall, 2004, p. 225). In any case, we’ve already got enough ingredients to launch the first criticism.

### 6.3.1 Supervenience as a causal relation?

When a mental property  $M$  supervenes on a physical property  $P$ , what can be said about this relation between the two? One thing that immediately comes to my mind is that  $M$ ’s spatiotemporality is identical with  $P$ ’s; if  $P$  occurs,  $Q$  occurs simultaneously in the same spatiotemporal region. On the contrary, if anything stands in some causal relation to another such that one thing is a cause and the other is an effect, there is a spatiotemporal *difference* between the two—e.g., a white billiard ball hits another ball at  $t_1$  and another ball consequently moves at  $t_2$  ( $t_1 < t_2$ ). Generally speaking, such spatiotemporal difference between a cause and an effect is taken as a necessary condition for causality. However, such spatiotemporal difference is not found in supervenience; subvenient and supervenient properties occur at the same time. Thus, the beauty of a painting supervenes simultaneously on the painter’s finishing his last brushing at  $t_1$ . The fact that the supervenience between mental and physical properties doesn’t exhibit a spatiotemporal difference supports the claim that supervenience is not a causal relation.

Recall now why nonreductive physicalists think that (OC1), viz.,  $(P_1 \wedge \sim M_1) \Box \rightarrow P_2$  is vacuously true: because of supervenience, any instance of  $P_1$  *guarantees, implies, or indicates* an instance of  $M_1$ , i.e.,  $\Box(P_1 \rightarrow M_1)$ . So the antecedent  $(P_1 \wedge \sim M_1)$  of (OC1) is false in all the closest possible worlds. I argue that in thinking so, nonreductive physicalists are confused about supervenience in that they take it as a causal relation. More specifically, when they say  $P_1$  guarantees

$M_1$ , it seems to me that they are taking the following to hold:

- If  $P_1$  had occurred,  $M_1$  would have occurred.
- If  $P_1$  hadn't occurred,  $M_1$  wouldn't have occurred either.

For what else could counterfactualists mean via counterfactual talk when they say  $P_1$  guarantees  $M_1$ ? Nothing other than the ones above. And when they are committed to this, they are committed to causation as counterfactual dependence; for A to cause B is for B to counterfactually depend on A. What the above construction implicates is precisely this counterfactual dependence relation between  $P_1$  and  $M_1$ . So far, so good. The supervenience relation between  $P_1$  and  $M_1$  can be interpreted in the form of counterfactual dependence. But there is a subsequent question to follow: Is the counterfactual dependence between  $P_1$  and  $M_1$  causal? Recall the thesis of Dependence; it only requires that counterfactual dependence be *sufficient*, not necessary, for causation. So we can say that there are causal counterfactual dependence and noncausal counterfactual dependence. What I want to do is ask if the counterfactual dependence between base and supervening properties is causal.

According to the Lewisian analysis of counterfactuals, B *causally depends* on A iff if A were not to occur, B would not occur either (Menzies, 2017). But this leaves room for some cases in which noncausal counterfactual dependence is wrongly interpreted as causal. In order to avoid such an error, there need to be qualifications. More specifically, the above definition of causal counterfactual dependence works only for those cases that meet the following: (a) the relata are events, (b) the related events are distinct events, and (c) the related events are interpreted standardly (Menzies, 2017). I argue that the supervenience case satisfies all three conditions, and thus that the counterfactual dependence of the supervenience

case falls under causal counterfactual dependence.

The first qualification is easy to digest. Although I have so far been using expressions like “a mental property  $M_1$  causes another mental property  $M_2$ ”, it doesn’t mean that I was talking about property causation. As I have said earlier in this thesis, the reason why I use such locutions is for the ease of the argument; when I say “a mental property causes  $M_1$  another mental property  $M_2$ ”, what I mean is that an event consisting of  $M_1$  causes another event consisting of  $M_2$ .

The second qualification needs more attention. I said earlier that the criterion of the distinctness of events is causal power of events. However, this isn’t a complete definition of distinct events for causal power alone gives rise to some counterintuitive cases. For instance, imagine an event consisting of writing “University” and an event consisting of writing “Univers” as part of “University”. If “University” hadn’t been written, “Univers” wouldn’t have been written either; does it mean that “University” is a cause of “Univers”? Our intuition says it doesn’t. In order to avoid such misinterpretation, we need to introduce a more restricted sense of distinctness by stipulating additional conditions: events are distinct iff (a) events are not identical, (b) neither is part of the other, (c) and neither implies the other (Menzies, 2017). According to these conditions, the university example above is not a case of distinct events because (b) is violated; the combination of alphabets “Univers” is part of the word “University”. Now let’s evaluate the supervenience case along with these conditions so as to see if it passes to be genuinely distinct events, and thus to be a case of causal counterfactual dependence. Firstly, (a) is easily vindicated given that an event of  $P_1$  and an event  $M_1$  instantiate different properties, that is,  $M_1$  is a mental event and  $P_1$  is a physical event. Next,  $P_1$  and

$M_1$  are not part of the other in the sense that  $M_1$  isn't a mereological part of  $P_1$ .

While the university case is a case in which one thing constitutes the other, the supervenience case isn't like that;  $P_1$  and  $M_1$  are correlated, not constituted by one another. There is a sense in which  $M_1$  is physically grounded in  $P_1$ , but this doesn't have to mean that  $M_1$  thereby is part of  $P_1$ . Lastly, (c) stipulates that for two events  $c$  and  $e$  to be distinct,  $\sim\Box(c \leftrightarrow e)$  should obtain. Let  $c$  be an event instantiating the occurrence of  $P_1$  and  $e$  be an event instantiating the occurrence of  $M_1$ . Then  $c \rightarrow e$  holds provided that the occurrence of  $P_1$  metaphysically necessitates the occurrence of  $M_1$ . However,  $e \rightarrow c$  isn't true because of multiple realizability; the occurrence of  $M_1$  doesn't imply or necessitate a particular token occurrence of  $P_1$ . So the biconditional requirement fails to hold, which means that the supervenience relation passes (c) as well.

The last qualification is that events are to be interpreted standardly. The standard interpretation of counterfactual dependence of events means that events are analyzed under no extreme or bizarre circumstances. For instance, backtracking interpretation is one such bizarre circumstance. To illustrate, suppose  $P_1$  is a cause of two distinct events  $M_1$  and  $P_2$  (never mind questioning whether  $P_1$  actually causes  $M_1$ . This is just to give an example). Then we can say by virtue of counterfactual dependence that if  $P_2$  weren't to occur,  $P_1$  wouldn't occur. Moreover, if  $P_1$  were not to occur,  $M_1$  wouldn't occur either. Then we get a counterintuitive result that if  $P_2$  were not to occur,  $M_1$  wouldn't occur, i.e.,  $P_2$  is a cause of  $M_1$ . The standard interpretation puts that backtracking and other extreme ways of interpreting counterfactual dependence are excluded. And I think that the supervenience relation can be interpreted without such obstacles; if  $P_1$  were to

occur,  $M_1$  wouldn't occur as well, period.

I think the discussion so far gives a good enough reason to think that counterfactual analysis implies that  $P_1$  and  $M_1$  stand in a causal relation. However, this result is unacceptable since we have seen that supervenience relation is not a causal relation.

### 6.3.2 Supervenience and backward causation

Given supervenience, the relation between  $P_1$  and  $M_1$  can be interpreted in two ways; the first way was discussed in the previous section where I interpreted it as  $P_1 \rightarrow M_1$ . In this section I examine the second way, which is more problematic. The second interpretation is the following:  $\sim M_1 \rightarrow \sim P_1$ . Provided that there is a correlation between  $P_1$  and  $M_1$  such that if  $M_1$  doesn't hold, then  $P_1$  doesn't hold either, we can legitimately construct a conditional  $\sim M_1 \rightarrow \sim P_1$ .<sup>13</sup> And just like the previous section, the conditional wrongly implies that  $P_1$  and  $M_1$  stand in a causal relation given that the counterfactual interpretation of the conditional is "if  $M_1$  hadn't occurred,  $P_1$  wouldn't have occurred either." This is nothing but to say that  $M_1$  is a cause of  $P_1$ . Two implications can be drawn from this.

The first is that this interpretation seems to allow backward causation. Backward causation refers to a case where a temporally preceding cause is caused by its effect. In the current case, although there is no temporal difference in the occurrence of  $M_1$  vis-à-vis  $P_1$  in one sense, there is another sense in which  $P_1$  is temporally prior to  $M_1$ : because of the supervenience of  $M_1$  on  $P_1$  and of the

---

<sup>13</sup> This also follows given that  $P_1 \rightarrow M_1$  logically implies  $\sim M_1 \rightarrow \sim P_1$ .

ontological priority of the physical over the mental,  $P_1$  comes first.<sup>14</sup> The ontological priority of  $P_1$  over  $M_1$  thus points to the fact that the second interpretation is non-sense; for  $\sim M_1 \rightarrow \sim P_1$  under counterfactual analysis seems to suggest that  $M_1$  is a cause of  $P_1$  that is ontologically prior to  $M_1$ .

The second has to do with Physical Closure, according to which a physical effect can be fully explicated by a sufficient physical cause alone. To see how this goes, I invite readers to the following line of reasoning: suppose that  $P_1$  causes  $P_2$  such that  $P_1 \rightarrow P_2$ . Currently, any  $P_1$ -worlds that are accessible from the actual world  $w$  are  $M_1$ -supervening worlds—that is,  $\sim M_1$ -worlds are *ipso facto*  $\sim P_1$ -worlds given that supervenience holds between the two.<sup>15</sup> Thus we may reasonably say that  $\sim M_1 \rightarrow \sim P_1$ . This naturally leads to the following consequence:

given  $\sim M_1 \square \rightarrow \sim P_1$  and  $\sim P_1 \square \rightarrow \sim P_2$ , it follows from these by transitivity that  $\sim M_1 \square \rightarrow \sim P_2$ .

What this shows, I believe, is that Physical Closure fails in this case. Physical Closure puts that a physical effect must have a physical cause that is sufficient alone. However, the counterfactual approach points out that a physical cause as such is not possible; there is a causal chain that results in the effect both *qua* physical and *qua* mental. In other words, the reasoning above seems to suggest that  $P_1$  alone can't cause  $P_2$  without the help of  $M_1$ . This consequence should be avoided given that Physical Closure is one of the basic premises of physicalism. At this point, one may ask what the problem really is: isn't it supervenience that really is problematic? The

---

<sup>14</sup> The physical is ontologically prior to the mental because the physical is the blocks of which the world we live in are made of. And this reasoning was invoked so as to support why  $P_1$  is the real cause of  $P_2$  in the exclusion argument.

<sup>15</sup> For the ease of the argument I invoke only a particular token of many physical realizers of  $M_1$  since invoking a set of all physical realizers of  $M_1$  has the same result.

supervenience between the mental and physical causes has created so much trouble for counterfactual analysis of causation. But I argue that it is the counterfactual notion of causation that actually is problematic. For the alleged causal dependency of  $P_1$  on  $M_1$  seems to be driven from the counterfactual view of causation; a different approach can avoid it. Note again that any counterfactual analysis of causation one way or another takes the following point: for A to cause B is for B to counterfactually depend on A. I believe this is the main culprit of why the counterfactual analysis vis-à-vis supervenience causes the causal dependency problem:

- (1)  $P_1$  causes  $P_2$  means  $P_2$  counterfactually depends on  $P_1$ , i.e.,  $P_1$  causes  $P_2$  iff  $\sim P_1 \square \rightarrow \sim P_2$ .
- (2)  $\sim M_1 \square \rightarrow \sim P_1$  (by supervenience).
- (3)  $\sim M_1 \square \rightarrow \sim P_2$  (by transitivity from (1) and (2)).

A series of counterfactual dependence between causes and effect yields that a physical cause alone isn't able to cause a physical effect.

Before closing this section, I would like to mention one independent, but related point to be made as to the causal status of supervenience. So far, the discussion in this section has relied on the claim that the physical-mental supervenience under counterfactual analysis mistakenly leads to supervenience being a causal relation. However, counterfactualists may argue that such a result doesn't have to be implied, hoping that saving supervenience from a causal interpretation might vindicate the applicability of the counterfactual approach to the exclusion argument. For instance, one may argue that in the conditional  $\sim P_1 \rightarrow \sim M_1$ , the subjunctive or counterfactual operator " $\rightarrow$ " doesn't need to be interpreted in causal terms. But I have already elaborated on this issue in the previous section, the result of



which is that such a noncausal interpretation in the current text is not possible. For the argument's sake, however, let us suppose that one has an argument for a noncausal interpretation of supervenience relation under counterfactual analysis.

If the picture about a causal competition between  $M_1$  and  $P_1$  tells anything, it is that we can't take  $M_1$  (and  $P_1$ ) single-handedly causing an effect. Indeed, this seems to be a trivial truth provided that if any properties stand in a supervenience relation, then even though each property distinctly and independently exists, there is a sense in which they go hand in hand. Thus, when we consider an effect  $P_2$ , there seems to be necessarily a structure consisting of  $P_1$ - $M_1$ - $P_2$ . If this is the case, there arises what I would call the problem of "causal link". Jeff Engelhardt (2015) makes the same point. According to him, it is hard to make sense of the causal or noncausal nature of the  $P_1$ - $M_1$ - $P_2$  structure. For we assumed that  $P_1$ - $M_1$  isn't causal while  $P_2$  should stand in a causal relation to its cause. So the question is: how can we make sense of the claim that the half causal, half noncausal process is a causal process after all? If something is a part of a causal "chain" or "link" that ends up with an effect, that thing itself must be causal as to the effect. But the assumption that the supervenience relation that holds between  $P_1$  and  $M_1$  isn't causal resists such a conclusion. Thus, supervenience seems to be in jeopardy when evaluated under counterfactual analysis; both of the two possible interpretations of supervenience under counterfactual analysis give rise to unwelcome consequences.

## CHAPTER 7

### CONCLUSION

The hitherto discussions can be summarized as follows: as I subscribe to reductive physicalism, I have tried to defend the exclusion argument against the so-called counterfactual solutions by nonreductive physicalists whose aim is to falsify one of the core premises of the exclusion argument. The motivation behind attacking the exclusion argument was that the exclusion argument has been thought to raise some serious challenges to the rival of reductive physicalism, that is, nonreductive physicalism. And on the way the reflections on alternative views as to mental causation and the basic concepts involved in the debate were made and investigated.

After that, the counterfactual approach vis-à-vis the exclusion argument was discussed: I first put forward versions of the counterfactual solutions to the exclusion argument and showed that they have common cores that are subject to the criticisms such as: that the replacement strategy found in the counterfactual solutions is not warranted as the strategy gives rise to the semantic emptiness to counterfactual talks; that a certain version of the exclusion argument is actually implied by the counterfactual approach itself; that it confuses supervenience relation as a causal one.

What I have done so far, however, doesn't imply that the exclusion problem survives and therefore reductive physicalism wins once and for all. There are criticisms to reductive physicalism and the exclusion argument from different perspectives that are still open to dispute. For instance, some philosophers make use of determinable-determinate relation to solve mental causation (e.g., Yablo, 1992,

Wilson, 2009), or use dual-explanandum strategy to give causal explanations of one event at two different levels—the mental and the physical level (e.g., Stueber, 2005, Kroedel, 2015). Another remaining question is if counterfactual notion of causation isn't appropriate for the evaluation of supervenience relations, what other notion of causation can do the job. It seems to me that if we follow Ned Hall's distinction between causation as (counterfactual) dependence and as production, the only remaining option is the production view. And as I have mentioned earlier, Kim also espouses the production view. However, there isn't a full-blown account of the production view; it's still at a suggestive step (Hall, 2004)—let alone its application to evaluating supervenience relations. List and Menzies (2009) also make a similar remark when they say that “.. unless a better explication can be given of causation as production, this notion can hardly play a significant role in the debate about mental causation (pp. 489-490).” So there are still a lot of works to be done in making sense of mental causation along with causation as production. However, what I have shown in this thesis is that *at least*, the counterfactual approach doesn't offer us a plausible solution to the exclusion argument, and that counterfactual analysis doesn't seem to be an appropriate medium through which supervenience can be understood. It is an open question if the exclusion problem can be defeated after all by some other means. But those who want to refute the exclusion argument would need something other than the counterfactual solutions.

## REFERENCES

- Alexander, S. (1927). *Space, time, and deity*, vol. 2. London: Macmillan.
- Armstrong, D. (1961). The causal theory of mind. In Chalmers, D. (Ed.), *The nature of mind and other essays*. Ithaca: Cornell University Press.
- Armstrong, D. (1968). *A materialist theory of the mind*. London: Routledge.
- Bennett, K. (2003). Why the exclusion problem seems intractable, and how, just maybe, to tract it. *Noûs*, 37(3), 417-497.
- Bennett, K. (2004). Spatio-temporal coincidence and the grounding problem. *Philosophical Studies* 118(3), 339-371.
- Bennett, K. (2008). Exclusion again. In Hohwy, J., & Kallestrup J. (Eds.), *Being reduced: New essays on reduction, explanation, and causation*. Oxford: Oxford University Press.
- Christian, L., & Menzies, P. (2009). Nonreductive physicalism and the limits of the exclusion principle. *The Journal of Philosophy* 106(9), 475-502.
- Churchland, P. (1981). Eliminative materialism and the propositional attitudes. *The Journal of Philosophy* 78(2), 67-90.
- Davidson, D. (1970). Mental events. In Davidson, D. (Ed.), *Essays on actions and events*. Oxford: Clarendon Press.
- Dennett, D. (1971). Intentional systems. *The Journal of Philosophy* 68(4), 87-106.
- Descartes, R. (1998). The principles of philosophy. In Clarke, D. (Ed.), *Meditations and other metaphysical writings*. London: Penguin Books.
- Descartes, R. (2009). Meditations on first philosophy. In Ariew, R. & Watkins, E. (Eds.), *Modern philosophy*. Indianapolis: Hackett.

- Engelhardt, J. (2015). Mental causation is not just downward causation. *Ratio* 30(1), 31-46.
- Feigl, H. (1958). The 'mental' and the 'physical'. In Feigl, H., Maxwell, G., & Scriven, M. (Eds.), *Minnesota studies in the philosophy of science*, vol. 2. Minneapolis: University of Minnesota Press.
- Fodor, J. (1989). Making mind matter more. *Philosophical Topics* 17, 59-79.
- Hall, N. (2004). Two concepts of causation. In Collins, J., Hall, N., & Paul, L. A. (Eds.), *Counterfactuals and causation*. Cambridge: the MIT Press.
- Harbecke, J. (2014). Counterfactual causation and mental causation. *Philosophia* 42, 363-385.
- Hare, R. M. (1952). *The language of morals*. Oxford: Oxford University Press.
- Huxley, T. H. (1874). On the hypothesis that animals are automata, and its history. *The Fortnightly Review* 16 (New Series). 555–580. Reprinted in Huxley, T. H. *Method and results: Essays by Thomas H. Huxley*. New York: D. Appleton and Company.
- James, W. (1879). Are we automata? *Mind* 4, 1-22.
- Kallestrup, J. (2006). The causal exclusion argument. *Philosophical Studies* 131(2), 459-485.
- Kim, J. (1984). Concepts of supervenience. *Philosophy and Phenomenological Research* 45. 153-176. Reprinted in Kim, J. *Supervenience and mind*. Cambridge: Cambridge University Press.
- Kim, J. (1987). 'Strong' and 'global' supervenience revisited." *Philosophy and Phenomenological Research* 48. 315-326. Reprinted in Kim, J. *Supervenience and mind*. Cambridge: Cambridge University Press..
- Kim, J. (1988). Supervenience for multiple domains. *Philosophical Topics* 16. 129-150. Reprinted in Kim, J. *Supervenience and mind*. Cambridge: Cambridge University Press.
- Kim, J. (1989a). Mechanism, purpose, and explanatory exclusion. *Philosophical*

*Perspectives 3, Philosophy of Mind and Action Theory*. 77-108. Reprinted in Kim, J. *Supervenience and mind*. Cambridge: Cambridge University Press.

Kim, J. (1989b). The myth of nonreductive materialism. *Proceedings and Addresses of the American Philosophical Association* 63. 31-47. Reprinted in Kim, J. *Supervenience and mind*. Cambridge: Cambridge University Press.

Kim, J. (1990). Supervenience as a philosophical concept. *Metaphilosophy* 21. 1-27. Reprinted in Kim, J. *Supervenience and mind*. Cambridge: Cambridge University Press.

Kim, J. (1993). *Supervenience and mind*. Cambridge: Cambridge University Press.

Kim, J. (1995). The non-reductivist's troubles with mental causation. In Heil J., & Mele A. (Eds.), *Mental causation*. Oxford: Clarendon.

Kim, J. (1998). *Mind in a physical world*. Cambridge: the MIT Press.

Kim, J. (2005). *Physicalism or something near enough*. Princeton: Princeton University Press.

Kroedel, T. (2015). Dualist mental causation and the exclusion problem. *Noûs* 49(2), 357-375.

LePore, E., & Loewer, B. (1987). Mind matters. *The Journal of Philosophy* 93, 630-642.

Lewis, D. (1973). *Counterfactuals*. Oxford: Basil Blackwell.

Leuenberger, S. (2009). What is global supervenience? *Synthese* 170(1), 115-129.

Menzies, P. (2017, December 21). Counterfactual theories of causation. *The Stanford Encyclopedia of Philosophy*. Retrieved from <https://plato.stanford.edu/entries/causation-counterfactual/>.

Moore, G. E. (1922). *Philosophical studies*. London: Kegan Paul, Trench, Trubner & Co.

Moore, D. (2012). Causal exclusion and dependent overdetermination. *Erkenntnis*

76(3), 319-335.

Nagel, E. (1961). *The structure of science*. New York: Harcourt, Brace & World.

Place, U. T. (1956). Is consciousness a brain process? *British Journal of Psychology* 47(1), 44-50.

Putnam, H. (1973). Psychological predicates. In Capitan W. H., & Merrill, D. D. (Eds.), *Art, mind, and religion*. Pittsburgh: University of Pittsburgh Press. Reprinted in Chalmers, D. *Philosophy of Mind: Classical and Contemporary Readings*. Oxford: Oxford University Press.

Russo, A. (2016). Kim's dilemma: Why mental causation is not productive. *Synthese* 193, 2185-2203.

Roche, M. (2014). Causal overdetermination and Kim's exclusion argument. *Philosophia* 42, 809-826.

Smart, J. J. C. (1959). Sensations and brain processes. *Philosophical Review* 68, 141-156.

Stueber, K. (2005). Mental causation and the paradoxes of explanation. *Philosophical Studies* 122, 243-277.

Wilson, J. (2009). Determination, realization, and mental causation. *Philosophical Studies* 145, 149-169.

Yablo, S. (1992). Mental causation. *Philosophical Review* 101(2), 245-280.

Zhong, L. (2011). Can counterfactuals solve the exclusion problem? *Philosophy and Phenomenological Research* 83(1), 129-147.

Zhong, L. (2014). Sophisticated exclusion and sophisticated causation. *The Journal of Philosophy* 111(7), 341-360.

Zhong, L. (2015). Why the counterfactualist should still worry about downward causation. *Erkenntnis* 80, 159-171.