

9-25-2015

The exclusion problem and counterfactual theories of causation

Yee Hang SZE

Follow this and additional works at: http://commons.ln.edu.hk/philo_etd



Part of the [Philosophy Commons](#)

Recommended Citation

Sze, Y. H. (2015). The exclusion problem and counterfactual theories of causation (Master's thesis, Lingnan University, Hong Kong). Retrieved from http://commons.ln.edu.hk/philo_etd/13

This Thesis is brought to you for free and open access by the Department of Philosophy at Digital Commons @ Lingnan University. It has been accepted for inclusion in Theses & Dissertations by an authorized administrator of Digital Commons @ Lingnan University.

Terms of Use

The copyright of this thesis is owned by its author. Any reproduction, adaptation, distribution or dissemination of this thesis without express authorization is strictly prohibited.

All rights reserved.

THE EXCLUSION PROBLEM AND
COUNTERFACTUAL THEORIES OF CAUSATION

SZE YEE HANG

MPHIL

LINGNAN UNIVERSITY

2015

THE EXCLUSION PROBLEM AND
COUNTERFACTUAL THEORIES OF CAUSATION

by
SZE Yee Hang

A thesis
submitted in partial fulfillment
of the requirements for the Degree of
Master of Philosophy in Philosophy

Lingnan University

2015

ABSTRACT

The Exclusion Problem and Counterfactual Theories of Causation

by

SZE Yee Hang

Master of Philosophy

Jaegwon Kim famously challenged non-reductive materialism/physicalism (NRM), a popular stance in the philosophy of mind, with the so-called exclusion argument. The argument is alleged to show that if NRM is true, then mental properties cannot be causes. In recent years, there have been many reactions to the exclusion problem based on counterfactual accounts of causation. In particular, List and Menzies gave an interesting response based on a Lewis-style counterfactual theory of causation, with some modifications made to Lewis's semantics of counterfactuals. In this thesis, I first argue that their central thesis regarding the issue of exclusion can actually be established with Lewis's original semantics, without modifications. I then clarify the real consequence of List and Menzies' modification by showing that List and Menzies' modified counterfactual theory, but not Lewis's original theory, can satisfy Zhong (2014)'s requirement of causal autonomy.

My analysis reveals that, despite Zhong (2014)'s employment of an interventionist theory of causation, it is not any peculiarity with the interventionism that is essential for establishing the possibility of causal autonomy. On the contrary, I argue that the existing semantics for interventionist counterfactuals do not seem to serve Zhong's purpose particularly well. These considerations suggest that once one decides to use a dependence notion of causation to counter the exclusion argument, a Lewis-style theory of causation is good enough.

DECLARATION

I declare that this is an original work based primarily on my own research, and I warrant that all citations of previous research, published or unpublished, have been duly acknowledged.

Sze Yee Hang

(Sze Yee Hang)

Date: 8 October 2015

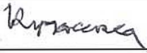
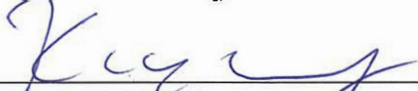


CERTIFICATE OF APPROVAL OF THESIS

THE EXCLUSION PROBLEM AND
COUNTERFACTUAL THEORIES OF CAUSATION

by
SZE Yee Hang

Master of Philosophy

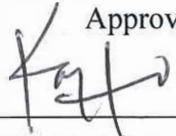
Panel of Examiners:

 _____	(Chairman)
(Prof. Rafael De Clercq)	
 _____	(External Member)
(Prof. Wong Kai Yee)	
 _____	(Internal Member)
(Prof. Zhang Jiji)	
 _____	(Internal Member)
(Prof. Mikael Pettersson)	

Supervisor:

Prof. Zhang Jiji

Approved for the Senate:



Prof. Mok Ka Ho Joshua
Chairman, Postgraduate Studies Committee

25 SEP 2015

Date

Contents

List of Figures	ii
Chapter I - Introduction to the Exclusion Problem	1
1.1 The Problem of Mental Causation.....	2
1.2 The Exclusion Problem	4
1.3 Yablo's Determinate/Determinable Solution.....	8
1.4 Mental Causation from a Counterfactual Point of View	12
1.5 Different Conceptions of Causation and Their Significances	15
1.6 Plan of the Thesis	19
Chapter II - Using Counterfactual Theories of Causation to Solve the Exclusion Problem	22
2.1 Introduction	23
2.2 List and Menzies' Theory of Causation	25
2.3 Establishing List and Menzies' Scenarios with the Centering Requirement	30
2.4 Why and Where Do We Need to Employ the Weak-Centering Requirement ...	39
2.5 The Causal Autonomy Picture Established in a Lewisian framework of Counterfactual Theory of Causation with Weak-Centering.....	40
Chapter III - Further Considerations	49
3.1 The Causal Autonomy Picture and Overdetermination.....	50
3.1.1 Problems with the Causal Autonomy Picture	50
3.1.2 Overdetermination?	53
3.2 Considerations on adopting the Interventionist Theory	57
3.2.1 A Sketch of the Interventionist Theory of Causation	57
3.2.2 Pearl's Semantics and the Centering Requirement	61
3.2.3 Brigg's Revised Semantics and Violation of Weak Centering	63
3.3 Interventionism and the Alleged Problems of Causal Systems with Supervenience	64
3.3.1 Baumgartner's Critique.....	64
3.3.2 Woodward's Response and Its Problems	67
3.3.3 Considerations about the Woodward-Baumgartner Debate	70
Concluding Remarks	72
References	76

List of Figures

Figure 1: Downwards Exclusion.....	33
Figure 2: Upwards Exclusion.....	35
Figure 3: Overdetermination / Compatibility.....	38
Figure 4: Causal Autonomy (with Weak-Centering Requirement)	43

Chapter I - Introduction to the Exclusion Problem

1.1 The Problem of Mental Causation

The problem of mental causation - concerns about whether and how the mental can have causal effects on the affairs in the physical realm - has been one of the central issues in philosophy of mind and metaphysics since the beginning of the modern age. Descartes' famous discussions in his *Meditations on First Philosophy* about material substance and mental substance, body and soul and their interactions, probably provided grounds for the subsequent discussions - although philosophers today might not be using the very same ideas and vocabularies that Descartes had endorsed, the scope of discussion still remains more or less the same. The core issues are: what is the nature of mind? What is its relationship with the material body? How do these two seemingly different "things" interact? Descartes himself proposed a dualist theory, which is kind of central to his philosophical system as a whole - but others need not agree with this stance. As Jaegwon Kim noted (Kim 2010), in a series of correspondence between Descartes and Princess Elisabeth of Bohemia (Kim 2007), it is evident that even in Descartes' age, there had already been voices that cast doubt on dualist theories about the mind. Princess Elisabeth's argument, as Kim speculated, might well be the very first rebuttal against dualist theories by pointing out that there exists the interaction problem, that if the mental and the physical are totally different things, it will be hard to properly address mental causation (p.243-244, 260-261).

Subsequent developments in philosophy of mind and metaphysics had largely debunked Descartes' substance dualism and it does not have a large number of followers in the contemporary scene of philosophy. Nevertheless, the dualist intuition - the idea that what is going on in the mind and what is going on in the

"outside", "material" world has fundamental and seemingly intractable differences - that drove Descartes and other dualists to form dualist theories, lingers on. But the problem of mental causation still poses difficulties for philosophers who want to build a satisfactory account about the mind, especially for those who do not want to completely "reduce away" the mind.

Although adhering to a strongly reductive theory of the mind might be expedient in the sense that we can completely eliminate the problem of interaction and mental causation, for there is no special status of the mental that philosophers need to specifically account for, since all mental properties/event can be completely reduced to the physical, it cannot evade other more distinct but still related problems like free will, agency and ethical responsibilities, which seemingly require that the mental is independent and distinct to a certain extent - these topics and assumptions being debatable, but intuitively speaking, it is hard to attribute responsibility without assuming agency, choice and free will, which are not very compatible with a highly deterministic world, where macro-physical level is realized¹ and determined only by micro-physical facts and events, and everything that happens adheres strictly to natural/physical laws. This explains why, despite the fact that scientific discoveries probably point towards monist materialism (that the mind is not something over, distinct and above the material body), there are still plenty of philosophers who resist complete reduction and explaining away the mind.

¹ The mental-physical relationship is often regarded as supervenience in contemporary literature. For A to be supervenient on B means that there can be no change in A without changes in B. There are debates on whether supervenience is the kind of realization we wanted for explaining the mind-body relationship, as it is a purely modal concept, and for some philosophers it seems not strong enough, as they see the mind-body relationship as more tightly connected than supervenience can cover. Some say that the mental emerges from the physical, and some say that the mental is grounded in/by the physical. In this thesis I will often use the terms "realization" and "realize" when depicting the mental-physical relation, which are commonly employed in the literature on the exclusion problem.

Rather, these philosophers would resort to "non-reductive materialism (NRM)" ("materialism" is loosely identical with "physicalism" - the subtle differences do not matter here and the two terms will be used interchangeably hereafter), which accepts that there is still only one kind of "substance", but the mental is not completely identical and reducible to the material. Generally speaking, for the non-reductive materialists the mental properties are not reducible into respective physical properties, and their relationship should rather be seen as other kinds of determination, for example supervenience, while there being only one kind of substance: the physical. NRM theories try to find an account for the relationship between the mental and the physical, but basically the idea is that the mental properties, although not identical to the physical ones, are "realized" by the physical level, and only by physical realizations that these mental properties can manifest causal power and other significations in the worldly operations, and they alone will not be able to cause differences in the world. Most physicalists would probably require that the causal power of the mental events is identical to that of the physical event that realizes it - no more, no less, otherwise it would seem to violate the physicalism part in non-reductive physicalism.

1.2 The Exclusion Problem

Kim is a stern critic of the non-reductive materialist theories of the mind. For many years he has argued that the NRM stance is unstable as there are critical flaws in the structure of this approach, and that a NRM theorist will ultimately be put into the situation of choosing between the substance dualism side and the completely reductive side, on the spectrum of theories about the relationship between the mental

and the physical. He thinks that NRM theories are just property dualism in disguise that base itself on a physicalist outlook ("Property dualism based on ontological physicalism is called non-reductive materialism (or physicalism). Emergentism, anomalous monism, and Putnam-Fodor functionalism are the best-known examples of non-reductive materialism." Kim 2005, p. 158.) Kim has formulated his arguments several times, and the one that is seen in his book "Physicalism, or something near enough" (Kim 2005) might best represent his view. In that book, Kim argues that it is impossible to establish a "stable" non-reductive materialism. He thinks that this is just an "intermediate halfway house" between substance dualism and complete reductivism. Kim argues that according to his exclusion arguments, it is not possible for the mental have causal powers on either the mental or the physical level, and non-reductive materialism would end up as "relegating mental phenomena to the status of epiphenomena". (p.158) Kim thinks that non-reductive materialism is just a version of property dualism, and he argues that, according to his conception of causality, "duality" of property and substance are both "not allowed" in the process of something causing something (ibid.) - Kim sees that for each event there can only be one and only one directly precedent cause, and if a physical event is already caused by a preceding physical event, it cannot be simultaneously caused by another event.

One of the central arguments that Kim uses to attack the NRM theories is the exclusion problem, which does seem to be posing a serious problem for anyone who wish to defend a version of NRM.

Kim and reductivists claim that the exclusion problem arises when we are trying to find a way to non-reductive materialism. The exclusion problem arises when

combining the requirement of physical closure and the intuition that events are generally not overdetermined. According to the principle of the closure of the physical realm, every physical property/event has a sufficient physical property/event as its cause. Mental events are different from physical events. It is generally unlikely for one event to have two causes, and if it has a cause it must be a physical one, and thus the Mental is excluded from having any causal power and influence at all. Kim believes that this is a fatal blow for all NRM theorists, and it has indeed caused much trouble. According to Loewer, the most recent version of Kim's exclusion argument goes like this: mental event m causes mental event m^* , and m^* has p^* as its supervenience base, and m caused m^* by causing p^* , because m^* is realized/instantiated by p^* . Also, m itself has a physical supervenience base p . With appeals to the principle of closure – “if a physical event has a cause that occurs at t , it has a (sufficient) physical cause that occurs at t ”, and the principle of exclusion – “no single event can have more than one sufficient cause occurring at any given time – unless it is a case of causal overdetermination”, and because for Kim “ p^* is not causally overdetermined by m and p ”, because the relationship between the two events are not the same as the case “in which there are two shooters and each of which kills the victim”. And with the non-overdetermination and the physical closure principle, as well as the prevalent idea among the NRM theorists that the mental properties are not the same as the physical properties, Kim argued that “putative mental cause m is excluded by the physical cause p ”. Therefore, Kim argues that in the NRM theories the causal power of distinct mental properties and events cannot be satisfactorily accounted for. (Loewer 2007, pp.250-251)

The Exclusion problem poses a serious question to many philosophers who try to find a way to formulate a satisfactory account of non-reductive materialism, and is often used by opposing camps to attack the non-reductivists, claiming that this is a fundamental flaw in the structure of non-reductive theories, as it seems conceptually impossible to reconcile the conflicting premises and intuitions. NRM theorists readily accept that mental properties are distinct and different from the purely physical ones (or else there is no need of a specifically "non-reductive" theory), and that as physicalists they would need to accept the completeness of physics, assuming that there is only one kind of "substance" that is the physical/material substance, and the causal relationship only occur at this level of substance, while there is no other kind of substances (e.g. the Cartesian mental substance) that can causally affect the physical substance. All events occur in the realm of the physical. (This would probably lead to another problem that concerns the definition of "the physical", though it is out of the scope of this thesis). In this regard, it is impossible for both the mental and the physical events to be a cause of a later event, and with the physicalist intuition that the mental manifests its causal power with the physical realizer, and the idea that the physical realm is closed in the sense that any cause of a physical event must itself be a physical one, the mental factor in the causal relationship tree will be an excluded one, with no effects on the physical realm whatsoever. The exclusion problem seems to give us a picture that is not very different from epiphenomenalism - and that is no good news for non-reductive materialists.

Karen Bennett (2008)'s presentation and analysis of the exclusion problem is precise and illuminating as she pointed out what kinds of claims and intuitions drive us into the apparent cul de sac. Bennett concludes that the exclusion problem occurs when

we put together five claims that are not consistent as a group. Bennett thinks that we will be in trouble if we maintain that, (1) mental properties/events are distinct from physical ones, that (2) the physical is a closed and complete realm for causation – for every physical occurrence has a sufficient physical cause, that (3) mental events/properties are able to cause physical events/properties in virtue of their mental-ness, that (4) the mental-physical causation relationship is not a systematic overdetermination one, unlike the firing squad example, and that (5) no effect has more than one sufficient cause if it is not a (rare) overdetermination case, i.e. the exclusion principle. Because mental causation, as commonly perceived, is not a systematic overdetermination scenario (4), and for every physical occurrence there is a sufficient physical cause (2), and with the mental and physical properties/events being distinct and different (1), we cannot have (3) established because of the exclusion principle (5). (Bennett 2008, pp.280-281)

It seems that Kim's fierce critique of the non-reductivist conceptions of the mind stands on firm grounds, as the exclusion argument(s) exposes the inconsistency between the various intuitions and claims that a non-reductive materialist often has.

1.3 Yablo's Determinate/Determinable Solution

Yet not every philosopher is convinced by the exclusion argument and Kim's attack on the non-reductive materialist theories. Stephen Yablo is one of the prominent defenders of the NRM approaches to the philosophy of mind, and his idea of proportionality and difference making is one of the more famous defenses of mental causation in such approaches. Yablo (1992) observes that it is possible to adapt the

traditional paradigm of one-way necessitation to the discussion of mental causation. He thinks that it is possible to treat the mental and the physical as a determinable-determinate relation, like redness and crimson. Yablo writes that “since a determinate cannot preempt its own determinable, mental events and properties lose nothing in causal relevance to their physical bases... If anything, it is the other way around.” (p.225) It seems that this is a fitting analogy, and satisfies one of the most common intuitions - the idea of multiple realizability, that the mental phenomena can be realized by a number of neuro-physical formations, just like redness can be “realized” by a bunch of similar but different colors that exists in the spectrum of red, like crimson, scarlet, maroon, etc.

His idea is that when facing multiple candidates of causes, "When all of the conditions are met—that is, y is contingent on x, and requires it, and x is adequate, and enough, for y—x will be called proportional to y. Without claiming that proportionality is strictly necessary for causation, it seems clear that faced with a choice between two candidate causes, normally the more proportional candidate is to be preferred." (Yablo 1992, p.244) He argued that the relationship between a mental property and the corresponding physical property is analogous to the relationship between a determinable property and a corresponding determinate property. To provide a simplified picture of Yablo's arguments - he used a pigeon which is trained to peck at red objects as example: a pigeon which is trained to peck at red objects would peck at crimson/scarlet objects, as the latter instantiate the property of red. But although the crimson/scarlet properties are the realizer of the property red and they are the property in role of making the trained pigeon peck, we will still attribute the cause to the property of redness, instead of the lower level properties of

crimson/scarlet, because redness is apparently the more proportional choice in the said situation - the pigeon is trained to peck at red objects, and will peck at anything that realizes the color red. Yablo (1992) argues that in such cases, the determinable property would be a more suitable candidate to be attributed as the cause. (p.261) Since it is a determinable-determinate relationship, there are no causal competition that can occur: the spatial-temporal properties of the occurrence of the color red and the microproperties are completely identical and, cannot be separated.

And it seems that the determinables - macro-properties - are more proportional, in the sense that the pigeon would still have pecked, even if the determinate micro-properties are different, so long as the macro ones stay the same. If we train the pigeon to peck at redness, it would peck whenever there is redness, regardless whether that redness is realized by scarlet or crimson. Maybe the pigeon pecked at crimson, but it would still have pecked whenever scarlet is present, even when there is no crimson. It would not have pecked when there is no red object (suppose this pigeon is with military grade discipline), however. So to say that scarlet (or crimson) is the real cause seems a little bit off, because it cannot catch the idea that the pigeon is trained to peck at redness, and it is the redness that causes the pigeon to peck. Therefore, redness, the macro property, the determinable, would be “better placed to play the role of cause.”, as it is more proportional to be called the cause than the lower level property. (ibid.)

But we should notice that Yablo's conception of causation may be quite different from what Kim and the philosophers who embrace the exclusion argument have in mind. Here it is helpful to borrow Ned Hall (2004)'s idea that there are two prevalent

view of causation - one dependency/counterfactual view, another being the production view. Kim (2007) explicitly endorses the production view, while the exclusion skeptics would often lean towards the dependency/counterfactual view. In Yablo's paper, the distinction between the two kinds of causation does not play an explicit role, but it is arguable that his conception about the proportionality of cause, causal relevance and the idea of difference making is more friendly to the dependency view than the production view (the List and Menzies paper supports this observation), as the production view would probably not allow the notion and the existence of multiple "causally sufficient candidates" - the "oomph" and "juice" view of causal power. But in general, Yablo's argument against the exclusion principle is neutral in regard of what kind of conception about the nature of causation we endorse. He even thinks that "any credible reconstruction of the exclusion principle must respect the truism that determinates do not contend with their determinables for causal influence." (p.232) Yablo does not think that his discussions would need to presuppose a theory of causality that is different from Kim's conception. In fact, Yablo thinks that even Kim would need to accept that properties that stand in a determinate-determinable relation are not competitive for their causal status. Yablo thinks that this is simple metaphysics, and he even goes on to argue that "determinates are tolerant, indeed supportive, of the causal aspirations of their determinables", and when we approach the issues of mental causation and its relationship with causal exclusion, the exclusion problem is "neutralized" and the solution is "anticlimactic" (ibid.)

But still, it is debatable whether the relationship between mental properties and the physical properties resembles the determinable/determinate relationship. The

difference-making and proportionality account that Yablo proposed are later formalized and used by List and Menzies, who explicitly endorse a counterfactual theory of causation.

1.4 Mental Causation from a Counterfactual Point of View

In the early literature on the issues of exclusion argument, philosophers did not often explicitly point out which conception of causation that they actually have in mind when talking about mental causation. List and Menzies (2009) explicitly state that they try to look at the mental causation problems from the difference-making point of view, which is similar to Yablo's proportionality account. (Yablo says "the motivation for imposing a proportionality constraint on causes is the dictum that causes make a difference to their effects. This dictum underlies many different theories of causation: counterfactual, probabilistic, interventionist, and contrastive ones." p.481) Under this conception, the non-reductive materialist theories will not be vulnerable to Kim's attacks. And although they formulated another version of the Exclusion argument that targets at their difference-making conception of causation, the paper has shown that whether non-reductive materialism is tenable "depends on the empirical characteristics of each causal system in question"(p.500), as there are different kinds of exclusions that may happen in the light of difference making, including cases in which the causal power of higher-level properties excludes that of the lower-level ones. List and Menzies (2009) claim that their analysis vindicates non-reductive physicalism "at least minimally" (p.500), since they are able to establish the possibility of three kinds of scenarios, that it is possible for causal systems to exhibit instances of upwards exclusion, downwards exclusion and

compatibility. While scenarios of upwards exclusions would render non-reductive materialism false, the remaining two kinds of systems allows non-reductive materialism, for in these two systems, the mental/higher-level events/properties are causally apt – in the downward exclusion scenario the higher level even excludes the lower level and is able to be the sole cause of the effect. (p.500)

One reason why the exclusion argument seems so powerful is that it is "taken to be an analytic truth"(p.476), as we have seen previously that the conflict between the different claims that NRM theories presumed are apparently obvious and intuitive. The significance of their findings is that the exclusion problem does not seem to be a conceptual one like how it was originally presented in Kim's arguments. By changing the problem to an empirical one, using cases in the real life and adopting the difference-making view of causation, List and Menzies' findings can effectively put Kim's claim that non-reductive materialism is just property dualism in disguise into question. Kim would probably maintain his idea that the counterfactual/dependency theory is not the best account of causation, but he did not give any knock-down arguments for a production view either, and criticisms from the dependency view of causation would still be significant for those who want to find out what the exclusion problem really means for mental causation.

Barry Loewer (2007) explicitly states that Kim's conception of Exclusion argument is strongly based on the production view. He argues that Kim automatically wins the debate if we are to understand causation as a production process, as under this conception exclusion and non-overdetermination would become "virtually analytic", because it is impossible for an event Q that is said to be produced by the event P, to

be produced by a distinct event M. With the production view, it is just not possible for the mental to cause something when this something already has a sufficient physical cause (since mental causation is not a classical firing squad overdetermination case). (p.253) According to Loewer, "Kim seems to think of causal production as an intrinsic relation between relatively local events" (ibid.)

Having such a commitment does need explicit justification, but Kim did not give such arguments in detail in his earlier discussions of the exclusion problem. Loewer continues to write, "The intuitive force of Kim's argument derives from the fact that we tend to think, mistakenly, that causation is a fundamental relation of production that connects relatively local events" (pp.254-5).

Loewer also points out that with the adoption of the dependence view, the most important part of the exclusion argument - the no overdetermination requirement - is "defanged", because under such conception of causation, it is natural to have an effect overdetermined. Loewer writes, "There is no problem of overdetermination if causation is understood as dependence. On Lewis's account of counterfactuals a particular event (or the value of a range of possible events) can depend on many co-occurring events. The motions of one's body, for example, the motions of a person's arms and hands when reaching into the refrigerator, depend counterfactually both on her mental states (which snacks she wants) and on her brain (and other bodily) states and on a myriad of other states and events." (pp.255-6).

And for Loewer there are actually plausible reasons why we should adopt the dependency view instead of the production one. According to Loewer, the

dependency view might be more complying to how the neural mechanisms of our brains work. Loewer thinks that the neural/cognitive structure of the brain “grounds Lewis’s account of counterfactuals”, as “different decisions that one might make are realized in differential brain phenomena that can result via the laws from tiny microscopic immediately prior physical differences. If the laws are deterministic then these small differences from actuality involve small localized violations of law”. (Loewer 2007, p.259).

Similar observations can be found in Woodward (2003) too.

1.5 Different Conceptions of Causation and Their Significances

Causation is an important concept, and like all other prominent philosophical concepts, it is vague, deep, and full of contentions. Numerous attempts have been made to properly account for causation, and yet there is still not a single most satisfactory account of this concept that has been made yet. Hume is often seen as the creator of the regularity theory, and yet the addition of a second sentence that paved the way for later developments of counterfactual theories had confused readers and interpreters. The previous paragraphs show that Kim thinks that the production account is generally more favorable, while many other tried to use the counterfactual account to look into problems concerning mental causation and exclusion. A helpful paper that explicitly distinguishes the two conceptions is written by Ned Hall (2004). He thinks that these two conceptions of causation are "fundamentally different", and naturally they are not compatible with each other. Hall concisely introduced the difference by saying that the counterfactual view, a species of what he call the

dependence view, means that causation is "counterfactual dependence between wholly distinct events... 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", and the production view is more sophisticated and difficult to characterize, but roughly speaking it means that causal relation is that "an event c helps to generate or bring about or produce another event e". (p.225) The two conceptions seems equally fundamental in our unexamined rough intuitions about what causation is meant to be, but as Hall shows that they are fundamentally incompatible with each other, and choosing one view over the other when dealing with issues in mental causation would lead to different outcomes and effects. In fact, it is possible that the exclusion problem is not a serious threat when we adopt the counterfactual view, but it is critical for NRM theorists if they are more inclined to accept the production/generation view of causality.

Kim is aware of this side of arguments and he is not moved - "The tide now seems to have turned in favor of a broadly sine qua non conception of causation whose most influential modern version is due to David Lewis's seminal account of causation in terms of counterfactual dependence.... there appear to be numerous ongoing research projects attempting to develop a satisfactory version based on Lewis's basic insights. The idea of counterfactual dependence is this: e is counterfactually dependent on c just in case if c had not occurred, e would not have occurred." (2010, pp.251-2) And he thinks that the reason why he is not adopting this point of view is that causation understood as counterfactual dependency is still immature and there are many problems that the dependency view will have to fix in order to provide a satisfactory result - "There are numerous outstanding difficulties with the counterfactual

approach, among them the problems of overdetermination and preemption— problems that seem highly resistant to solution. The current literature is rife with increasingly complex and clever counterexamples and equally complex and ingenious remedies to evade them. The impression one gets from looking in from outside is that we are still very far from achieving the desired end, and that a reasonably simple and intuitively well-motivated counterfactual account of causation is not yet in sight." (p.252)

These criticisms are not unfair, but in the article Kim still did not give a knock down argument why it is mandatory, or at least more reasonable, to adopt a production conception of causation. Kim wrote that "Why should we resort to this "thick" variety of causation in thinking about mental causation? My answer is pretty simple: We care about mental causation because we care about human agency, and agency requires the productive/generative conception of causation."(Kim 2007, p.236) But, as a superficial evaluation, another conception about causation that involves agency and intervention - the interventionist view of causation - actually finds more affinity to the counterfactual approach than the production approach. One of the more serious flaws about the production approach is that the concept of "causal juice" is simply too mysterious and it does not quite fit with the findings of the sciences; Loewer quoted that on the fundamental level of physics it is doubtful whether talks about causality are meaningful or useful. Loewer argues that causation and cause-effect attribution is more of a supervenience thing (p.253). Such thoughts are not intended as knock-down arguments that dissuade us from adopting a production view, but they serve to show that Kim's arguments are not strong enough to favor the production view.

Chiwook Won (2014) also criticizes Loewer and other philosopher's attempt at solving the exclusion problem by adopting the counterfactual theory which should render the exclusion problems trivial. But similar to Kim's tactic, Won focused on criticizing the counterfactual dependency view of causation itself. Nevertheless, Won's attempted solution is more eclectic than Kim's, shown in the attempt of adoption a "hybrid" view of causation. Won thinks that it is possible to combine the dependency view and the production view of causation in dealing with mental causation, that the mental cause is a dependence cause of the effect while the underlying physical cause is a production cause (p.227). Won argues that, in the vein of Hall's distinction, if we think that different kinds of causation are involved when talking about mental causation, the exclusion problem might be able to be solved, since the problem arises when "the mental and physical causes are not properly construed as jointly causing the effect", and it is not a classical overdetermination case.

Both the story of productionists like Kim and that of the more neutral hybrid theories like Won tell us that although doubts are casted on the counterfactual theory of causation and the move of adopting it to look at mental causation, the exclusion problem is more or less still defanged or generated harmless if we are not strictly sticking to the generative/production view. The point of this thesis is not to provide arguments for or against a certain camp of theories of causation, but even the for-exclusion camp admits that there are rooms for discussion, and once we abandon the production view, the Exclusion problem will not seem so "near-analytic" and formidable.

1.6 Plan of the Thesis

The present thesis is mainly concerned with some recent attempts to apply counterfactual theories of causation to respond to the exclusion argument.

Chapter 2 introduces Christian List and Peter Menzies' response to the exclusion argument. List and Menzies (2009) employ a counterfactual theory of causation, in contrast with Kim's production/process theory. They argue that Kim's conception of causation is flawed as it cannot address the issues that arise when the proportionality constraint is brought into mind, and would fail to find out the real cause in some cases. List and Menzies proved, in a modified Lewisian possible world semantics, that the upward exclusion (where the causal power of the physical excludes that of the mental), downward exclusion (where the causal power of the mental excludes that of the physical), and the compatibility scenario (where both the higher and lower levels are suitable to be seen as causes), are all logically possible scenarios.

Therefore, the exclusion argument that is presented by Kim's, as a piece of a priori reasoning, cannot possibly establish its conclusion, for whether the upward exclusion or any exclusion takes place is an empirical rather than an a priori matter.

List and Menzies modified Lewis's semantics for counterfactuals to establish their main thesis about exclusion. They replaced Lewis's requirement of "centering" by that of "weak centering". In chapter 2 I will show that even without this modification, their main thesis can still be established. In other words, for the sake of their response to the exclusion argument, Lewis's original theory of causation suffices.

I then explore the real consequence of replacing “centering” with “weak centering” in regard to the exclusion argument. It turns out to be closely related to Lei Zhong (2014)’s response to the exclusion argument. Zhong argued that a proper response to the exclusion argument should lead to a structure of “causal autonomy” where, roughly speaking, mental causation and physical causation operate at different levels and do not, so to speak, cross levels. Zhong used an interventionist account of causation to argue for the possibility of causal autonomy. I will show, however, that List and Menzies’ account of causation, which replaces Lewis’s “centering” with “weak centering”, entails this possibility. Unlike Zhong’s argument that largely relied on intuitions to assess various interventionist counterfactuals, my argument will be a rigorous, model-theoretic proof. On the other hand, Lewis’s original semantics for counterfactuals does not allow the possibility of causal autonomy (Zhong 2011). Therefore, the real effect with respect to the exclusion problem, of replacing “centering” with “weak centering” in Lewis’s counterfactual theory of causation, is to license a picture of causal autonomy like Zhong’s.

However, in chapter 3, I will critically examine the picture of causal autonomy. I shall argue that the requirement of causal autonomy is on the one hand not well motivated, and on the other hand unable to fully accommodate the intuitions about mental causation that give the exclusion problem its teeth. Moreover, I critically examine the shift to interventionist theories of causation to account for the exclusion problem, attempted by Zhong (2014) and others. As shown in chapter 2, even if one aims at the structure of causal autonomy, List and Menzies’ modified counterfactual account of causation is sufficient. Bringing in the interventionist framework, while

perhaps enabling us to look at more intricacies concerning other issues about mental causation, does not seem to add anything new in countering the exclusion argument. On the contrary, the very notion of intervention, as commonly understood in the literature, does not seem to be straightforwardly applicable to systems with supervenience relations, as illustrated by some discussions in the literature.

Putting these considerations together, I will suggest in my concluding remarks that once one decides to use a dependence notion of causation to counter the exclusion argument, Lewis's original theory of causation seems good enough.

Chapter II - Using Counterfactual Theories of Causation to Solve the Exclusion Problem

2.1 Introduction

In chapter 1 of the thesis I have discussed the background materials of the debate about Kim's exclusion argument, and have briefly introduced the arguments from different sides of the participants. In this chapter I would like to further explore some of the points presented in List and Menzies (2009), as well as Zhong (2011, 2014). My main points will be that Lewis's original counterfactual theory of causation is good enough to meet List and Menzies' purpose, and the modification proposed in List and Menzies (2009) is needed only to license Zhong (2014)'s picture of causal autonomy.

Kim's exclusion argument cannot satisfactorily respond to scenarios where proportionality comes into consideration, as presented by Stephen Yablo (1992). According to List and Menzies, it is true that in some causal systems, certain forms of exclusion would happen, but under their revision, the exclusion principle as presented by Kim cannot be regarded as an a priori truth, and the revised exclusion principle comes in two different formulations - one downwards and one upwards - and together with a compatibility scenario, they are all possible scenarios, and one need to look into the detailed characteristics of specific, actual world causal systems, to find out which property is the real cause. After revision, existence of exclusion scenarios does not necessarily imply the rejection of non-reductive materialism anymore. The existence of downward exclusions refutes Kim's exclusion argument, and the existence of compatibility scenarios is one of the reasons that lead us to question whether we should uphold the non-overdetermination principle or not. List

and Menzies successfully show that exclusion is a contingent, empirical matter, rather than an a priori truth, as Kim seemed to claim.

According to List and Menzies, discussions about the exclusion problem are often ill-grounded for the lack of a clearly articulated theory of causation as a background for discussion. They observe that both proponents and critics of the exclusion principle "usually assume that its truth or falsity can be settled by an investigation of the concept of causation" (p.476), so they provided a well-defined difference making theory of causation, incorporating Yablo's concept of difference-making proportionality constraint, and largely following Lewis's counterfactual theory of causation, using a similarity based semantics for counterfactual conditionals. List and Menzies differ from Lewis in one point: they employ a weak centering requirement for the similarity between worlds, rather than the original, stronger centering requirement. I will show, however, that List and Menzies' main thesis regarding exclusion can be established with Lewis's original theory, making the modification seemingly unnecessary.

List and Menzies' approach deviates slightly from Kim's conception of the mental-physical causal structure in the way that they don't separate the effect into two layers - mental and physical. For List and Menzies, the effect is behavioral in a broader sense that it encompasses both the mental M and the physical P (List and Menzies uses N as in Neural for the base property, which is equivalent to P of physical in the context of the following discussion; the subtle differences do not matter). Kim's favorite picture consists, however, of four properties - mental cause (M1), mental effect (M2), physical cause (P1), physical effect (P2), with M1 realized by P1 and

M2 realized by P2. Kim does not believe that M1 is able to cause anything. Zhong (2001) shows that Lewis's counterfactual theory cannot establish that M1 is the cause of M2 without also causing P2. Zhong (2014) argues that using an interventionist theory of causation, it is possible to establish a scenario in which M1 causes M2 without causing P2, and P1 causes P2 without causing M2. The second point I will make is that this picture of "causal autonomy", though inconsistent under Lewis's original theory, is compatible with List and Menzies' modified account.

A note about notations. List and Menzies use M to refer to a mental property, N to refer to a neural property, and B to refer to a behavioral property. Zhong (2011) used M to refer to a mental cause and M* to refer to a mental effect, P to a physical cause and P* a physical effect. In this thesis notations M and P will be used to denote a mental property and a physical property, instead of List and Menzies' M and N, and when the effect is not separated into two layers B would be used as a behavioral effect. And for the causal autonomy picture we will follow Zhong (2014) to use M1 as mental cause, M2 as a mental effect, P1 as a physical cause and P2 as a mental effect. But they are interchangeable with M/M* and P/P*.

2.2 List and Menzies' Theory of Causation

Kim's killer move against the non-reductive physicalists who sought to preserve the causal power of the mental is the exclusion problem. The exclusion problem, as Kim claims, is an inevitable outcome resulting from the combination of several doctrines that most physicalists would generally accept. In Kim's formulation of the critique, when you combine the doctrine of non-overdetermination and the doctrine of

physical completeness, and with the picture of mental property M1 supervening on physical property P1 causing mental property M2 supervening on physical property P2, the mental level would be rendered causally powerless. Therefore, for Kim, it is not possible to maintain a stance of non-reductive materialist with the causal power of the mental preserved - either one goes reductive materialist, or one heads for substance dualism. Though the argument is appealing at first sight, there are in fact many points that are not self-evident and would allow plentiful rooms for discussion. In the coming section of the thesis I want to argue that Kim's argument is based on a specific view of causation - the production view, and it is not the only game in town, and as discussed in Hall (2004), it is, at best, as credible as the counterfactual view of causation; and under the counterfactual view, the exclusion problem is not as powerful as Kim claims.

List and Menzies, in their 2009 paper, "Non-reductive Physicalism and the Limits of the Exclusion principle", argued that Kim's formulation of the exclusion principle is flawed when we use a counterfactual approach to causation instead of a production one. Kim's idea of mental causation cannot successfully deal with Stephen Yablo's cases of pigeons pecking, as it would fail to pick out which factor –crimsonness or redness - is more appropriate to be said as the cause of the pigeons' behaviors.

Here is Yablo's (1992) pigeon example. Suppose a pigeon is trained to peck at red objects. It would peck at any object with a color that realizes redness. Say, crimson, scarlet, or other colors inside the red spectrum. The pigeon pecks at crimson. But we intuitively think that this crimsonness causes the pecking in virtue of realizing red.

To call the crimsonness the cause would be "overly specific and involves extraneous

detail" (List and Menzies 2009, p. 480). List and Menzies argue that the original exclusion argument "would say that since being red supervenes on being crimson and being crimson is causally sufficient for the pigeon's pecking, the redness of the target is not the cause" (ibid.). Then something seems a little bit off. It seems that, just as List and Menzies point out, if we are going to sort something solid out of this confusion, we would need a more clearly crafted theory of causation.

List and Menzies adopted a counterfactual theory of causation with a modified Lewisian possible world semantics for counterfactuals. To present it in a simplified fashion, according to the counterfactual theory of causation, A is a cause of B if and only if (1) if A were to be present then B would be present, and (2) if A were to be absent then B would also be absent.² Now we can see how this applies to Yablo's example: if the pigeon is trained to peck at red objects, it would peck at redness realized by all kinds of more specific colors. Suppose it pecks at a crimson thing in our actual world. If it is trained to peck (B) at redness (A), it would be reasonable to think that even if crimsonness (C) were absent, so long as redness (A) were still present (for example, realized by the color scarlet), pecking would still have occurred; but if redness is absent ($\sim A$), then the pigeon would not peck ($\sim B$). Redness satisfies the two counterfactual conditionals while crimsonness does not, so, according to the simple theory, red is a cause of pecking, but not crimson.

It is easy to translate this into languages spoken when dealing with mental causation.

If a person acts (B) because of a mental state / property (M), he would still have

² This simple counterfactual account of causation is nobody's official account. Lewis, for example, regards counterfactual dependence between distinct events only sufficient but not necessary for causation, in view of cases of redundant causation. But the further complication does not matter for the present purpose.

acted without the realizer P in the nearest possible world where P is absent but M is still present, as intuitively we would think that worlds where this person has the mental state M realized by another physical state Q is closer to the actual world than worlds where this concerned person does not have the mental state M. For example, suppose the concerned person is sleepy in our actual world due to an S-fiber firing. It is quite intuitive to think that, other things being equal, worlds where this person is sleepy due to T-fiber or U-fiber firing is closer to the actual world than worlds where this person is not sleepy at all.

List and Menzies acknowledge that there are cases where causal exclusion happens. To use Yablo's example to show this, if a pigeon is trained to peck at Red things, redness would exclude its realizers as cause, hence downward exclusion; if it is trained to peck at a more specific color like scarlet, scarlet-ness would have excluded the redness as the cause, and hence upward exclusion. But after taking Yablo's proportionality constraint into consideration, we know that Kim's exclusion principle will render us unable to pick out the real or at least the more proportional cause, if the cause is a higher-level property, for Kim's principle would always lead us to choose the physical base realizer as the cause. Therefore, revising the exclusion principle would render two formulations, where causal exclusion not only can happen in an upwards direction, but it can also happen in a downwards direction. The revised exclusion principle no longer restricts the direction of exclusion, and so it can be compatible with non-reductive materialism.

List and Menzies also argue that, unlike how Kim took his exclusion principle as an a priori truth and used it to rule out all possibilities of successfully establishing a

non-reductive theory of the mental, their new principle is sensitive to the empirical conditions, and can be false in some causal systems. List and Menzies' revision, therefore, provides rooms for establishing reductive materialism, non-reductive materialism, as well as other kinds of theories in the empirical world. The important thing is that their downwards exclusion principle is actually friendly to non-reductive materialism, saying that in the respective causal system where the higher-level property excludes the lower-one to be the cause of the respective behaviors, non-reductive materialism would be established.

List and Menzies loosened Lewis's original centering requirement in their semantics for counterfactuals. For Lewis, no world other than the actual world can be as close to the actual world as the actual world itself. List and Menzies have a slightly weaker requirement, that in the closest sphere of the possible worlds, there can exist some worlds other than the actual world that are as close to the actual world as the actual world itself.

A main motivation behind List and Menzies' modification of Lewis's centering requirement is to capture Yablo's notion of proportionality. Consider the pigeon example again. This time suppose the pigeon is trained to peck at precisely scarlet. Intuitively then scarlet but not red is the proportional cause of pecking. However, under Lewis's semantics, it is true that $\text{Red} \Box \rightarrow \text{Pecking}$, since they both obtain in the actual world. Moreover, it is also true that $\sim \text{Red} \Box \rightarrow \sim \text{Pecking}$, for in the closest $\sim \text{Red}$ -worlds there are no pecking due to the absence of scarlet. So Lewis's original theory will announce red as also a cause of pecking. By contrast, by weak centering, worlds with Red (say Crimson) but no Pecking can be as close to the actual world as

the actual world is, which could then falsify the conditional Red $\square \rightarrow$ Pecking. This way, only the intuitively proportional cause in this case, i.e. scarlet, will be picked out.

It is worth noting, however, that it is far from consensus that proportionality should be considered a semantic component of the concept of causation (Shapiro and Sober 2012). I do not intend to enter that debate here. Instead I will focus on List and Menzies' application of their theory to the exclusion problem. I will show that, regardless of whether they are justified in adopting the weak-centering requirement, for their main thesis regarding exclusion, this modification of Lewis is unnecessary.

2.3 Establishing List and Menzies' Scenarios with the Centering Requirement

List and Menzies established three scenarios after revising the exclusion principles and proposed that upwards exclusion, downwards exclusion and compatibility scenarios are all possible in the actual world, that exclusion is a contingent matter, and that the downwards exclusion and compatibility scenarios would show that non-reductive materialism is not refuted by the exclusion argument, because in the former case the mental property M serves as the cause of the effect B while the physical property P does not; and in the latter case both M and P are suitable to be attributed with the status of being a cause of effect B. In this section I will prove that the three scenarios are all compatible with Lewis's original theory, with the centering requirement.

Explaining the ideas of centering requirement, weak centering, and closeness of worlds, requires extensive discussions about Lewis's possible world semantics and the idea of similarity between worlds, and is out of scope of the thesis. For the purpose of our discussion a simple sketch of these ideas would suffice. The centering requirement means that the actual world is closer to itself than any other world is. It follows that if A and B are both true in the actual world, the counterfactual $A \Box \rightarrow B$ is true. With weak centering, there is no such consequence. In other words, even if A and B are both actually true, $A \Box \rightarrow B$ may well be false. This is because under the condition of weak centering, some other world can be as close to the actual world as the actual world itself. The main intuitive support of the condition of centering is that any other world is different from the actual world in some way; no matter how slight the difference is, it makes the world less similar to the actual world than the actual world itself is. Proponents of weak centering, on the other hand, typically respond that not every difference matters for overall similarity. The details of this metaphysical debate do not matter for the present purpose, and I will leave them aside.

For List and Menzies, the downwards exclusion principle reads like this:

Necessary and Sufficient conditions for downwards exclusion: an instance of downwards exclusion occurs if and only if M is a difference-making cause of B and B is present in some closest $\sim N$ -worlds that are M-worlds. (p.495)

And their conditions for upward exclusion are:

Necessary and sufficient conditions for upwards exclusion: An instance of upwards exclusion occurs if and only if N is a difference-making cause of B and either (i) B is absent in some closest M-worlds that are $\sim N$ -worlds or (ii) B is present in some closest $\sim M$ -worlds outside the smallest $\sim N$ -permitting sphere. (p.493)

(In the following discussion P (for physical) is used in N (for neural)'s place.)

One of the examples of the scenario where neither non-reductive stance nor reductive stance can be ruled out is that both the lower and the higher level are appropriate to be attributed as causes. All three scenarios will be explained in the coming paragraphs in adherence to Lewis's own centering requirements.

The first scenario where downwards exclusion happens is easy to conceive: that M is the cause of B while P is not. The idea of multiple-realization makes such scenarios possible. Figure 1 shows a situation or model in which it is true that $M \square \rightarrow B$ and $\sim M \square \rightarrow \sim B$, but it is not true that $\sim P \square \rightarrow \sim B$, meaning that M is a cause while P is not.

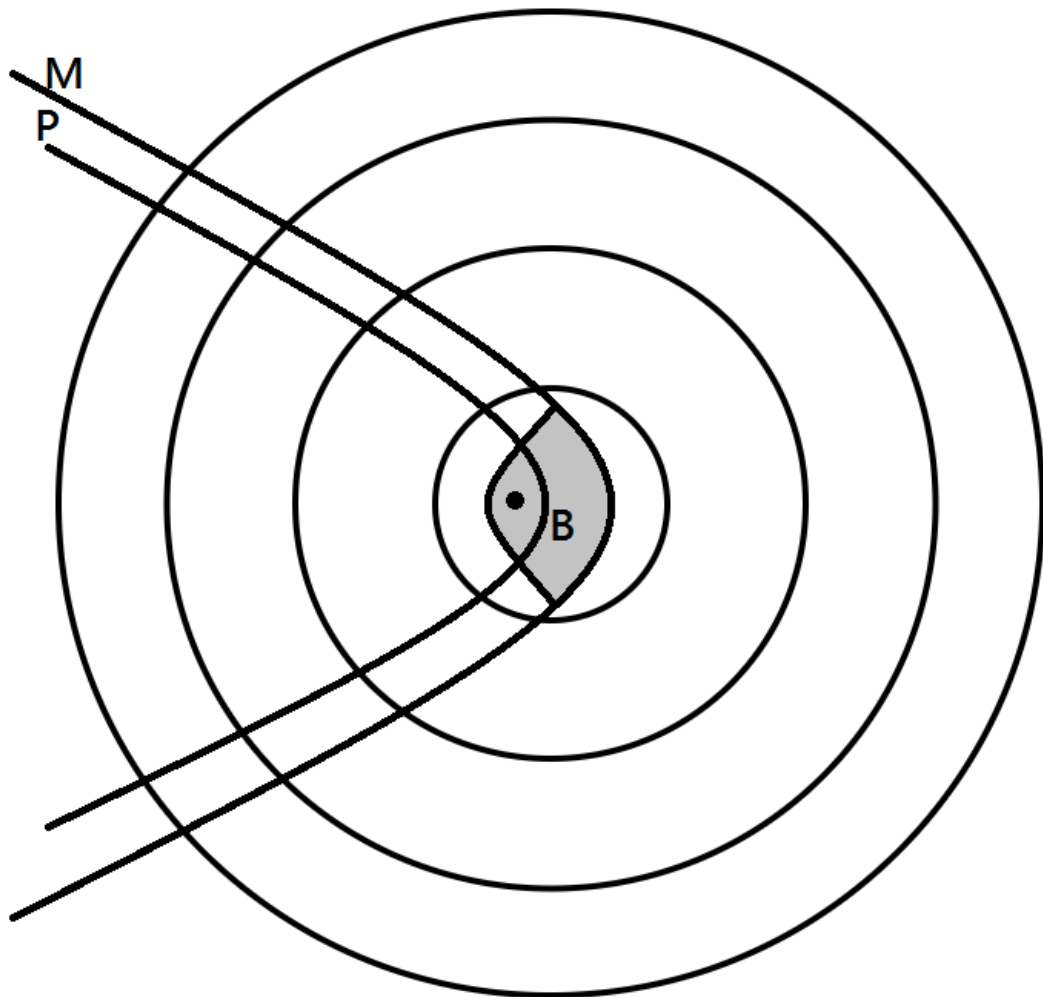


Figure 1. Downwards Exclusion

Figure 1: Downwards Exclusion

It precisely describes the scenario of a mental state (M), which can be realized by various different physical states, causes behavior B. For example, a person A is an alcoholic, he pines for vodka and goes to the convenience store to buy a bottle. The mental state of pining is realized by a neural state P. It is reasonable to imagine that in the nearest possible non-P world, A is still an alcoholic, still pines for beer, and still goes to buy vodka because of such pining, since the mental state of pining M might be realized by a different neural state other than P. But in the nearest non-M

world, A is not an Alcoholic and does not have desires for vodka and would not have gone out to purchase. Therefore, in such a scenario, M causes B while P does not.

The second scenario, Upwards Exclusion, where the lower level excludes the higher level, though being intuitive when put into the context of mental causation, is a little more trickier to explain under the Lewisian picture. Figure 2 illustrates such a scenario: M is realized by P, P Causes B (i.e., $P \Box \rightarrow B$, $\sim P \Box \rightarrow \sim B$), and M does not cause B because it is not true that $\sim M \Box \rightarrow \sim B$. This means that the physical is the cause while the mental is not. We can see from figure 2 that in all nearest worlds where there is no P, there is no B, while in some nearest non-M world, there is still B, which means it is not the case in the actual world that if there were no M then B would not be. Figure 2 satisfies List and Menzies' condition that "either (i) B is absent in some closest M-worlds that are $\sim N$ -worlds or (ii) B is present in some closest $\sim M$ -worlds outside the smallest $\sim N$ -permitting sphere."(p.493) In fact, it satisfies both conditions.

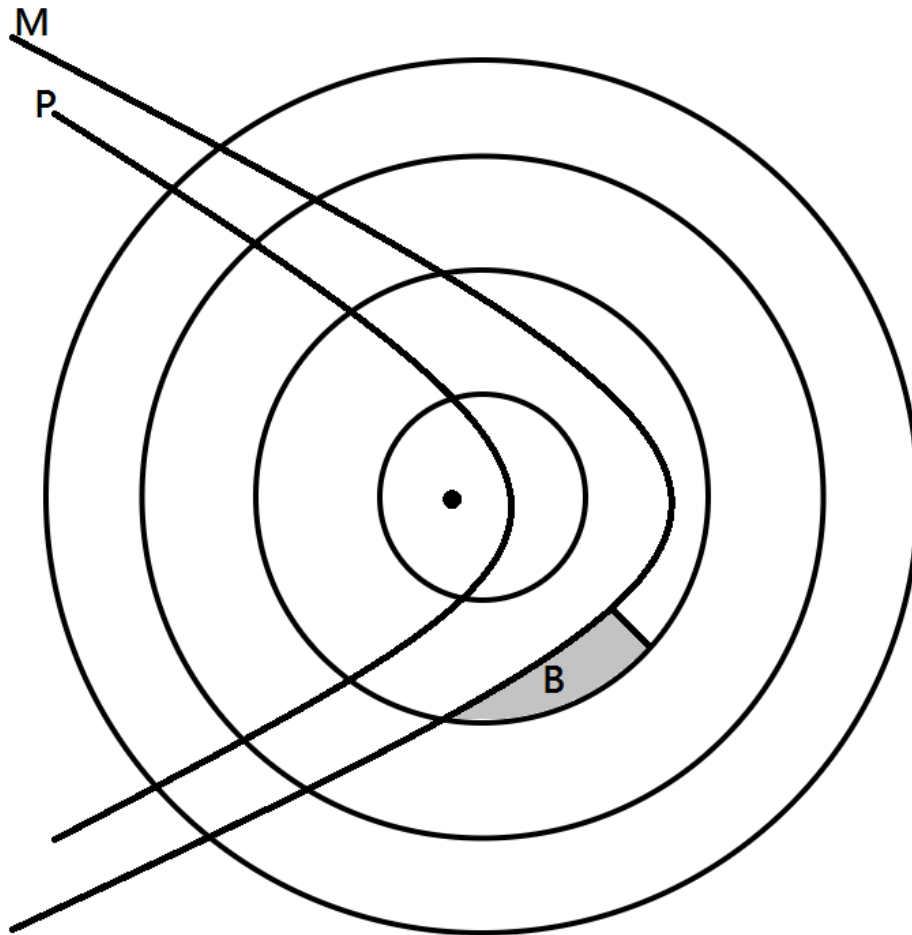


Figure 2 - Upwards Exclusion

Figure 2: Upwards Exclusion

A real world example that illustrates this scenario is tricky to find, but not impossible. Imagine a sleeping-disorder patient who has severe insomnia, and he can only sleep by taking a special kind of sleeping pill X that is very "strong". The M here refers to the qualitative feeling of drowsiness (M) that is realized by the neurological state under the effect of sleeping drugs, the P refers to the neurological state of being under the effect of the sleeping drug X. Because our patient took the right kind of medicine, he can fall asleep in the actual world.

While in the nearest possible worlds where P is absent, the patient took a different, slightly weaker sleeping pill Y for some reasons, which, while still being able to make the patient to have the same feeling of drowsiness (M), fails to affect the neurological system of the patient the same way drug X can ($\sim P$), and therefore the patient fails to fall asleep (hence $\sim B$). It is reasonable to think that the worlds where the patient still takes sleeping pills of other kinds is closer to the actual world than the worlds that the patient does not take any sleeping pills at all. In both worlds the patient has the qualitative feeling of strong drowsiness, but would fail to fall asleep without the use of a specific drug X , even if neural-state-affected-by-drug- X and neural-state-affected-by-drug- Y both realize drowsiness, hence it is quite reasonable to say that it is the underlying physical factor that is the real cause of the behavior of falling asleep, instead of the mental, since only neural-state-affected-by-drug- X can really cause the patient to fall asleep.

Whereas in the nearest non- M and non- P world, which is a little bit more far away than the worlds mentioned above, the patient might be using other more radical or non-orthodox methods to cure the insomnia. Suppose in such worlds the patient uses more radical methods to try to rest. Our patient might be using other kinds of drug (suppose he broke into a veterinary clinic and took a lot of strong sedatives purposed for veterinary surgery) that would knock him out in a matter of seconds, making him fall asleep without having to be drowsy at first, or he could have used some medical electronic appliances that can directly and abruptly change his brain functions, which enables him to tune himself into sleeping without the need of being drowsy before falling asleep. In either scenario mentioned above, our patient successfully goes into the state of sleeping (B) without having the mental state of drowsiness ($\sim M$) or the

specific neurological state of being affected by drug X ($\sim P$). Cases like this might be rare, but it is still empirically possible for such a case to happen, and it satisfies the causal system of upwards exclusion while sticking to Lewis's own requirement of centering.

The third scenario described by List and Menzies is a causal system where neither the lower level excludes the higher nor the higher excludes the lower. Figure 3 illustrates such a scenario, where $P \square \rightarrow B$, $M \square \rightarrow B$, and, incidentally, $\sim P \square \rightarrow \sim B$ and $\sim M \square \rightarrow \sim B$. In this case, both P and M can be said as causing the B to happen, since in both the nearest non-M worlds and nearest non-P worlds there would be no B. Imagine a case where a person A, after accidentally inhaling a certain amount of nitrous oxide ("laughing gas"), feels (strangely) jolly and laughs out loudly. The physical state of the brain being affected by the laughing gas (P) realizes the mental state of jolliness (M), and causes A to laugh (B). In the nearest possible world where there is no such accident and A does not inhale the gas ($\sim P$), he would not have felt jolly ($\sim M$) and hence would not have laughed ($\sim B$).

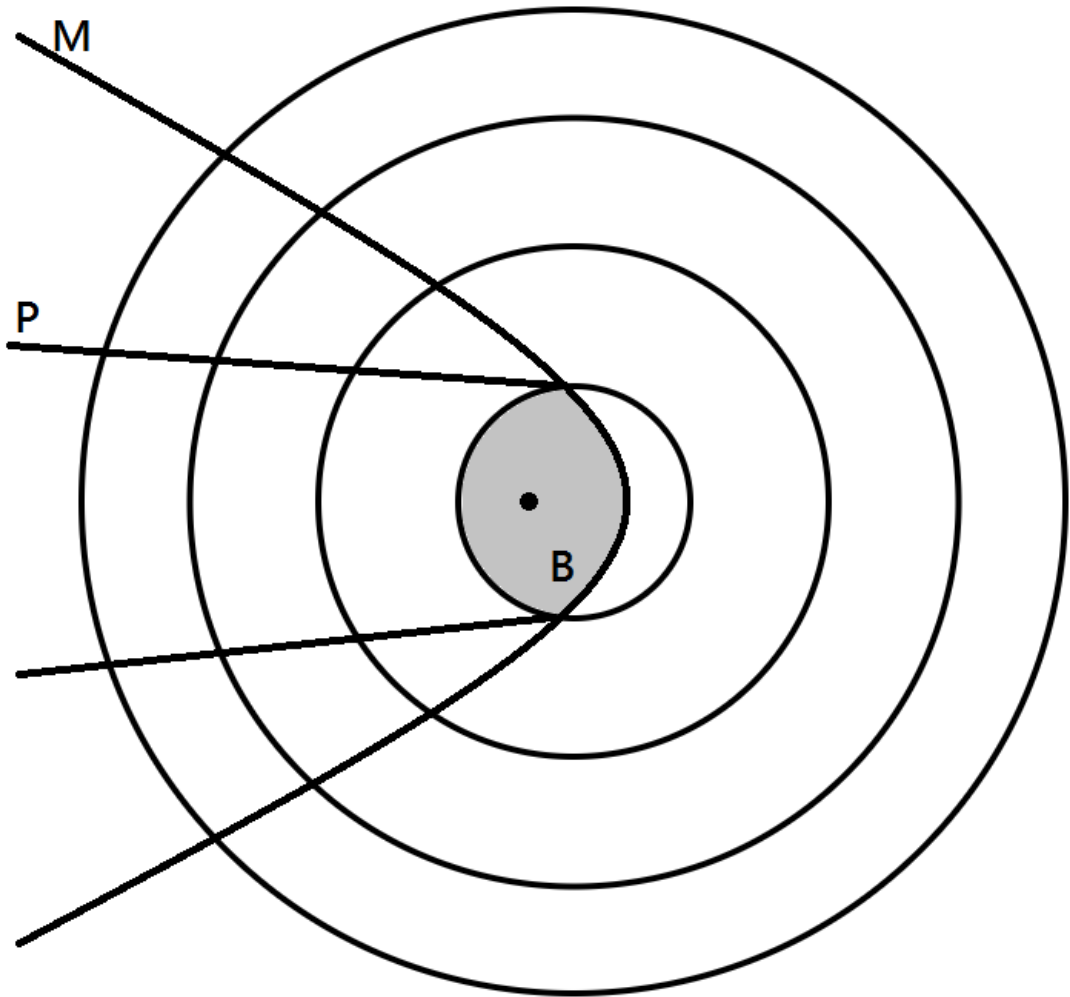


Figure 3. Overdetermination

Figure 3: Overdetermination / Compatibility

Figures 1, 2, and 3 above show that the three kinds of causal systems: 1. higher-level excluding lower-level, 2. lower-level excluding higher level, and 3. other systems, can all be constructed using Lewis's counterfactual theory of causation and without modifying the centering requirement. List and Menzies' main thesis regarding exclusion can be established under Lewis's original theory.

2.4 Why and Where Do We Need to Employ the Weak-Centering Requirement

The question now is: what difference does the modification of centering to weak centering make in regard to the exclusion problem? The answer turns out to be closely related to the picture of "causal autonomy" that Zhong (2014) argued for. In his 2011 paper, Zhong argued that Lewis's counterfactual theory has the following limitation in responding to the exclusion argument. His analysis reveals that if we adhere to Lewis's original centering requirement, then when M1 causes M2, it logically follows that M1 also causes P2, which would violate the non-overdetermination principle since P2 is caused by P1. However, if we adopt List and Menzies' approach, with the requirement of weak centering, we will be able to establish a picture of causal autonomy, where M1 causes M2 and P1 causes P2, but M1 does not cause P2 and P1 does not cause M2.

As mentioned in previous paragraphs, List and Menzies did not separate the behavioral effect (B) into two layers (M2 and P2), respectively parallel to the two levels of causes: M1 and P1. It is still debatable whether we need to separate it at all. We probably need such a picture if we want to take physical closure more seriously.

List and Menzies allow the mental to be the cause of the physical sometimes. Not all non-reductivists might agree upon this: they might want to limit the mental's causal power to the extent that it can only serve as a cause of effects that occurs on the higher-level, if we are to emphasize the physicalist part in non-reductive physicalism. For the non-reductivists who are not very comfortable with the idea that the mental is able to cause the physical base since they want to uphold the closure principle, which accommodates the physicalist intuition that on the fundamental level the physical is still caused by physical, the causal autonomy picture will be required, as it depicts the causal systems where both the physical and the mental have respective causal powers, but they would not interfere with each other, meaning that the physical closure is more effectively maintained, while the mental is not causally ineffective or epiphenomenal, in virtue of still being able to cause other mental properties. In this picture the mental (M1) causes the mental (M2), the physical (P1) causes the physical (P2), but the mental cause (M1) would not cause the physical (P2) and the physical cause (P1) would not cause the mental effect (M2). But this picture that depicts a causal autonomy, as Zhong (2011) shows, cannot be established with Lewis's original centering requirement. Nevertheless, it can be established, as shown in the next section, with List and Menzies' modification.

2.5 The Causal Autonomy Picture Established in a Lewisian framework of Counterfactual Theory of Causation with Weak-Centering

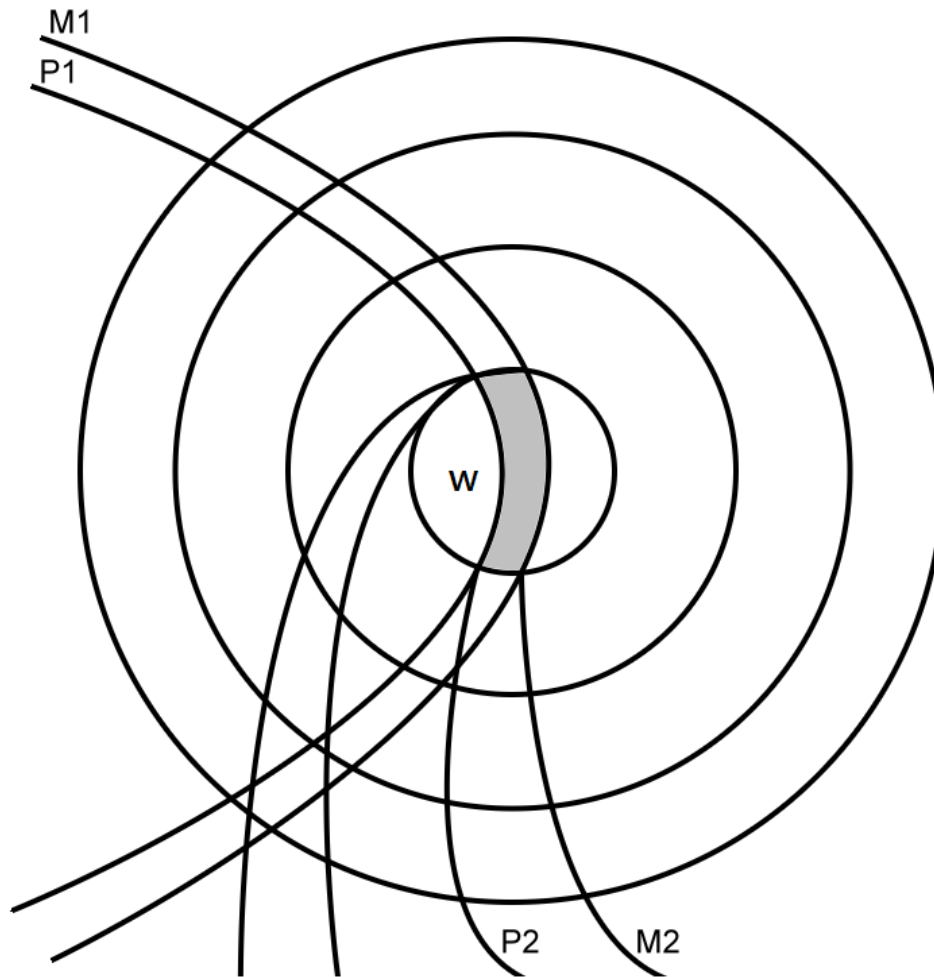
Zhong's critique of the counterfactual approach established on the basis of a disadvantage of the original centering argument. According to Zhong, Lewis's requirement would make it impossible to have M cause M* and P cause P* without

M also be a cause of P*. Since P* already has a sufficient physical cause P, if we cannot block M from causing P*, we would find ourselves in an overdetermination scenario, which Zhong thinks is not acceptable. But it is unable to block M's causal relation with P* under Lewis's requirement.

Suppose, in the actual world, an instance of mental causation happens, where M causes M* and P causes P*, with M realized by P and M* realized by P*. According to Lewis, it is automatically true that $M \Box \rightarrow P^*$, since M and P* both occur in the actual world. And, since P* realizes M*, any $\sim M^*$ world would also be a $\sim P^*$ world. For M to cause M*, according to the counterfactual theory of causation, two counterfactual conditions must be established: $M \Box \rightarrow M^*$ and $\sim M \Box \rightarrow \sim M^*$. With $\sim M \Box \rightarrow \sim M^*$, and any $\sim M^*$ world is also a $\sim P^*$ world, it follows that $\sim M \Box \rightarrow \sim P^*$. Given the two counterfactuals $M \Box \rightarrow P^*$ and $\sim M \Box \rightarrow \sim P^*$, it means that M also causes P*. (pp.141-2)

M1 causes P2 can be translated into two counterfactual conditionals: $(M1 \Box \rightarrow P2)$ & $(\sim M1 \Box \rightarrow \sim P2)$. Under the assumption that M1 causes M2, and so $\sim M1 \Box \rightarrow \sim M2$, and the assumption that P2 metaphysically necessitates M2, it logically follows that $\sim M1 \Box \rightarrow \sim P2$. However, while $M1 \Box \rightarrow P2$ is trivially true under Lewis's centering requirement and the presupposition that M1 and P2 are both actual, it is not the case anymore under the requirement of weak centering. According to the weak centering requirement, "even if F and G are both instantiated in the actual world, the smallest sphere around it also contains some other worlds instantiating F". Therefore, as seen in Figure 4, it can happen that a world where there is M1 but no P2 (shaded region: $M1 \& \sim P2$), is as similar to the actual world as the actual world itself, where there are

both M1 and P2. Therefore, even though M1 and P2 are both true in the actual world, the counterfactual $M1 \square \rightarrow P2$ can still be false, as is the case in Figure 4, and because of this, it is not the case that M1 causes P2. Also, according to Figure 4, some closest P1-worlds still have M2 meaning that it is false that $\sim P1 \square \rightarrow \sim M2$ and so P1 does not cause M2. Therefore, we achieve a picture where M1 causes M2, P1 causes P2, but M1 does not cause P2, and P1 does not cause M2.



$M1 \square \rightarrow M2$

$\sim M1 \square \rightarrow \sim M2$ - M1 Causes M2

$P1 \square \rightarrow P2$

$\sim P1 \square \rightarrow \sim P2$ - P1 Causes P2

$\sim (M1 \square \rightarrow P2)$ M1 Doesn't Cause P2

$\sim (\sim P1 \square \rightarrow \sim M2)$ P1 Doesn't Cause M2

Figure 4: Causal Autonomy (with only Weak-Centering Requirement)

Figure 4 illustrates a situation that satisfies what Zhong calls the “autonomy approach” to solve the exclusion problem. Zhong writes: “The exclusion problem

presents a mental property M and its physical realizer P as competing to be causally relevant to the same effect. But according to the autonomy solution, M may not be threatened with exclusion if M and P are causally relevant to different properties of the effect” (Zhong 2011, p. 140). Zhong argues, in his 2014 paper “Sophisticated Exclusion and Sophisticated Causation”, that by using an interventionist theory of causation, he “can reject both the upward causation principle and the downward causation principle, and hence support the autonomy option that M1 causes M2 and P1 causes P2, but M1 does not cause P2 and P1 does not cause M2.” (Zhong 2014, p. 359). While the interventionist theory is different from the counterfactual theory of causation and has its own merits, Figure 4 shows that the counterfactual theory can also achieve the causal autonomy option, by loosening Lewis’s centering requirement and adopting List and Menzies’ approach.

The causal autonomy picture cannot be achieved by following the original centering requirement, but it can be satisfied if we adopt the weak centering requirement, and the reason is similar to the problems of applying centering to deal with pictures of proportionality, which is mentioned above. The original centering requirement cannot deal with the causal relationships involving property realizations. If both the proportionality constraint and causal autonomy are required, adopting the centering requirement would lead to failures when we try to sort out the real cause when the action is caused by the determinate rather than the determinable, since both are true in the actual world and it would make it impossible to deny that the determinable also stands in the set of counterfactual relations to the effect that satisfy the requirements to be a cause. In the causal autonomy picture, since the higher level property is realized by the lower, and in any worlds where the higher level property

does not exist, the lower would not exist either. According to the original centering requirement, this would make both the higher level and the lower level property causes. It would be appalling if we apply such requirements on worlds where there is multiple realizability, that some higher level properties can actually be realized by more than one kind of lower level property, e.g. it sounds plausible and intuitive that a kind of mental property can be realized by several different formations of neural firing patterns. If such mental property is said to be the cause, we must find a way to forbid the lower level to automatically stand in the same counterfactual relationship as the higher level property, which is not possible in the Lewisian picture, but possible by adopting the weak-centering requirement. For example, for if M1 causes M2, any closest M1 world would also be a M2 world and any closest not M1 world would also be a not M2 world. Since M2 is realized by P2, $P2 \rightarrow M2$ and $\sim M2 \rightarrow \sim P2$, so any closest $\sim M2$ world would also be a $\sim P2$ world, establishing that any closest $\sim M1$ worlds would also be a not $\sim P2$ world, establishing $\sim M1 \square \rightarrow \sim P2$. But intuitively speaking, $M1 \square \rightarrow M2$ does not entail $M1 \square \rightarrow P2$ if the mental property is multiply realizable - as M2 may also be realized by other physical properties like P3 or P4. But in the picture with Lewis's original centering requirement, $M1 \square \rightarrow P2$ is automatically established since in our actual world M1 and P2 are both present, and it is not possible to say the worlds where M1 and not P2 is just as close to our actual world, the causal relation M1 causes P2 would mandatorily be established under the Lewisian counterfactual theory of causation, which is not allowed if principles of non-overdetermination and physical completeness are observed which only allows one cause to be the cause of P2 and that cause being the physical cause P1. But the weak centering requirement would fit more with the multiple-realizability intuition that allows the cases where M1 causes M2 but does not cause P2, since worlds where

P3 realizes M2 and worlds where P4 realizes M2 are just as close to the actual world where P2 realizes M2 as the actual world itself.

The picture also shows that while the mental cannot cause the physical, the physical will not be able to cause the mental as well. It is established that when there is P1 there will be M2, since P1 is the realizer of M1, meaning that, when there is P1 there is M1, and since when there is M1 there will be M2, it is true that when there is P1 there will be M2. But if M1 is multiply realizable, then even if there is no P1, in the closest possible world there would be another realizer for M1, meaning that there will still be an M2, therefore providing counterexamples to $\sim P1 \square \rightarrow \sim M2$, meaning that the P1-cause-M2 relationship cannot be established. Since both M1 causes M2 and P1 causes P2 are established, and both M1 causes P2 and P1 causes M2 fail to establish, causal autonomy is achieved.

As shown above, List and Menzies' formulation of counterfactual theory with weak centering requirement suffices to give a picture of causal autonomy even if we are to concede more to the requirements of Kim. It is actually questioning Kim's conclusion, by showing that even when we accept non-overdetermination and physical closure, it is still possible to have a world picture that achieves the general objective of the non-reductive materialists who think that it is important for us that the mental properties are causally empowered. List and Menzies' did not think it is necessary to separate the two levels so rigorously, and they do not have problems with the mental factors being able to cause behaviors without clearly separating the behavior into two layers - the mental side and the physical side. Nevertheless, the approach they proposed are strong and satisfactory enough for us to make a mental-friendly causal system even

when abiding to Kim's restrictive laws. Their formulation of the counterfactual account of causation actually fits better with our intuitions, since it can deal with cases if we take the proportionality constraint into account, as well as following the common intuition of multiple realizability, which are problematic when we adopt the original centering requirement. List and Menzies have also given formal proofs for their approach, attached as an appendix at the end of their paper.

Though the interventionist account for causation certainly has some advantages over the classic counterfactualist causation theory, it remains unclear whether it is the best move to take rather than List and Menzie's approach when dealing with the exclusion problem. Although Zhong in his 2014 paper had made convincing arguments for the interventionist approach and achieved satisfactory results of successfully making a picture of causal autonomy, there are still some points of considerations. In the paper, when it came to giving the causal autonomy picture Zhong did not use approaches that are unique to the interventionist theories, and most of the steps can be translated into the language of counterfactuals without losing something. Given that List and Menzies has given out a rigorous and more complete illustration of a counterfactual approach to the problem, it would require more reasons for adopting an interventionist approach instead of a revised counterfactual approach. The interventionist approach of causation and comparisons between it and the counterfactual approach on their applications on issues concerning mental causation would be discussed in chapter 3 of the thesis.

It is also worth noting that in another article of List and Menzies (2010), they called their approach in the 2009 paper an interventionist one that follows Woodward's

(2003, 2008b) proposal, albeit a simplified one. But in the 2009 paper they have not used the interventionist theory of causation at all; an orthodox Lewisian way of finding out proofs for counterfactual is used, namely the similarity semantics, as developed in Lewis (1973). Therefore, with the expositions and discussions above, in this thesis the List and Menzies' way would still be regarded as an example of using standard counterfactual theory of causation to deal with the exclusion problem and to establish non-reductive materialism.

Chapter III - Further Considerations

3.1 The Causal Autonomy Picture and Overdetermination

3.1.1 Problems with the Causal Autonomy Picture

Zhong (2011, 2014) envisages an orthodox Kim-ian picture of mental-physical relationship, that there are two layers of properties / events, one physical and one mental, standing in a supervenience relationship in which the upper layer supervenes on the lower layer (“Kim’s favorite diagram” as Loewer (2007) humorously put, p.257). Zhong would like to achieve a causal picture in which the mental causes only the mental and the physical only causes the physical, and the arrows of causation never go diagonal. Kim’s exclusion argument argues that the upper mental layer would never be able to cause anything, with the premises of non-overdetermination and physical closure, while the lower physical layer would be able to cause another physical instance, as well as the mental. Zhong (2011) argued that the Lewisian counterfactual theory cannot be used to resist the exclusion argument when adhering to Kim’s requirements, since the mental cause M1 inevitably causes physical effect P2 according to the counterfactual account of causation and the Lewisian semantics of counterfactuals, if M1 is assumed to be a cause of mental effect M2. This, according to Kim (and Zhong), would violate the premise of non-overdetermination, as the physical property P2, according to the physical closure clause, already has a sufficient physical cause P1. In his 2014 paper Zhong argued that an interventionist theory of causation, by contrast, allows a structure of causal autonomy in which P1 but not M1 is a cause of P2 and M1 but not P1 is a cause of M2, thus avoiding overdetermination.

But it seems that the causal autonomy picture is not adequate to account for some intuitions of mental causation. Take voluntary actions of primates as an example. From the cases of Anderson's monkeys as discussed by List and Menzies (2009), we can know that biological science has accepted that primates, including chimpanzees, gorillas and Homo sapiens (us), are capable of advanced and abstract psychological activities. It means that our primate brains are capable of generating abstract thoughts and high-level volitions to guide our actions and interactions with surroundings. In addition, in the realm of neural science the debate about how the mental thoughts are realized by biological organisms such as brains is still undergoing. Is the brain modular, or does it function like distributed systems, or maybe the picture proposed by advocates of connectionism bears the most resemblance to the true picture? There are no definitive answers yet. But brains are so malleable that we periodically see interesting stories in the news, that there are people who survived with rods penetrating their brains, that people with only half a brain left still manage to function normally just like those with complete ones, or that a patient with severe hydrocephalus that steamrolled his brain into a sheet somewhat managed to think and control the body like a normal person, etc. It seems that unlike building a personal computer, there are no schematics for us to precisely and accurately pinpoint the specific pattern of neural firings that indicates a specific higher-order thought, though rough directions and mappings of functional areas can be found. And accordingly, it seems that rather than being controlled by the more fundamental neural patterns, our actions may actually arise from higher-level volitions and such volitions are often capable to be attributed as causes of our actions.

List and Menzies' response to the exclusion argument does not face such problems because they allow the scenario of downwards exclusion, in which the higher level can be a cause, not just of effects of the same level, but also effects at lower levels.

List and Menzies argued that whether exclusion takes place and which kind of exclusion takes place depends on empirical researches into different causal systems. For example, if a primate voluntarily raises its left hand, it seems more intuitive and natural to say that its thought and volition cause both the physical realizers, like the muscle movements controlled by electro-chemical signals communicated through the neural system, and the higher-level behavioral event of raising the left hand. The causal autonomy picture, although elegant, is probably false. The intention of raising hand is the best candidate to be a cause of the behavior in such a scenario, because without it the primate would not have raised the hand. But with multiple realizability, and suppose that worlds with the primate having such an intention is closer to the actual world than those at which it does not have one, it is probably not the case that without that specific physical realizer of the intention, the hand would not be raised, since there would be other physical realizers that has taken its place and serve in the causal relationship. But with the causal autonomy requirement, we would need to strictly and clearly separate two layers, and would need to say that the intention causes the higher-level hand raising, while the neural states causes the muscles' contraction. Again, not necessarily false, but does not sound as intuitively promising as saying that the intention causes both layers. It seems that List and Menzies' line of thought would bear more resemblance to the actual picture happening in our world. Alternatively, maybe we can think of the causal autonomy picture as yet another possible scenario, where there is no exclusion but the two layers would never interact,

standing only in a realizer-realized relationship. Such a picture can also be subjected to empirical test as causal systems of this kind may exist, but it seems dubious that we should take this picture as true a priori, in the spirit of List and Menzies' doubt on the original Kim-ian exclusion argument.

3.1.2 Overdetermination?

Zhong's picture of causal autonomy is primarily motivated by the non-overdetermination principle. But it is dubious whether the relationship between a mental property and its physical realizer property should be regarded as a competing one. Often when philosophers talk about overdetermination, they envisage a firing squad scenario: a firing squad that consists of several gunmen who are ready to execute a prisoner on demand. When the order arrives, they all fire at once. Overdetermination takes place when two bullets arrive at the prisoner's heart and kill him at the very same time (the time should be precisely the same or else the scenario would be one of preemption). The simultaneous arrival implies that, in the nearest possible world where bullet A fails to kill (moisture, bad gun maintenance, Coriolis effect, etc.), bullet B would still have killed, and vice versa. This is a genuine case of overdetermination. It is at best doubtful whether the relationship between the mental and the physical is analogous to that between the two bullets. If we envisage the relationship between the mental and the physical as one of realization, then the firing squad example seems clearly different. After all, in Davidson's language, the mental and physical are token-identical, that they are not two events of the same kind/category/sort happening at once.

Bennett (2003) goes to length to argue that such “overdetermination” is not like the paradigmatic case, and it is possible to argue that the mental/physical relationship does not (always) involve overdetermination. In fact, there are philosophers who do not adhere to the non-overdetermination principle at all (e.g. List and Menzies 2009, 2010; Loewer 2007). According to Bennett, when we think of overdetermination, we think of cases like the death of the prisoner who got fatally shot by two executioners at the same time. Or, in the case of human behavior, cases like when “I want to raise my arm, and do so—and that, at the same time, somebody else grabs my arm and lifts it up” (Bennett 2003, p.476). But it seems that the relationship between the mental and the physical is not analogous to the aforementioned scenario. The physical and mental causes that are tightly related and are responsible for voluntary human behaviors, according to Bennett, do not seem to compete with each other when causing a consequent behavior. (ibid.) Bennett tries to justify such a view by arguing for a pair of necessary conditions for overdetermination, and showing that cases of mental causation fail to meet the conditions.

Bennett presents her formulation of the necessary conditions of overdetermination as such:

- e is overdetermined by m and p only if
- (O1) if m had happened without p, e would still have happened, i.e., $(m \ \& \ \sim p) \ \square \rightarrow e$, and
- (O2) if p had happened without m, e would still have happened, i.e., $(p \ \& \ \sim m) \ \square \rightarrow e$.

Bennett wants to evaluate the two counterfactuals to see if the cases of mental causation really satisfy the said conditions for being counted as overdetermination. She thinks that the compatibilist (non-reductive materialists) can deny that both of

them are non-vacuously true. In particular, she argues that the clause (O2) is either vacuous or false.

It is vacuous under the usual assumption that p is a sufficiently broad physical state/property such that it by itself realizes m in the sense that it is impossible for m to be absent in the presence of p. The vacuity is a salient signal of the fact that the mental/physical case is very much unlike the paradigmatic cases of overdetermination, where the overdetermining causes are independent, or at any rate, do not stand in any relation nearly as strong as that of metaphysical necessitation.

On the other hand, according to Bennett, one may reject the usual assumption and maintain that the typical examples of p and m cited in the literature do not, strictly speaking, bear the relation of supervenience. That is, p is not a sufficiently broad physical state/property that on its own is sufficient to realize m. It only realizes m in some proper environments or in combination with other physical states/properties. If so, it is possible for p to obtain without m, in which case (O2) is not vacuous.

However, argues Bennett, (O2) would probably be false in such a case. For the conditions that make p realize m are also conditions that make e happen in the presence of p. If p were to be present without m, those key conditions would not be present and hence e would probably be absent as well. At any rate, there is no reason to think that e would still obtain if p were to obtain without m.

The non-overdetermination principle in Kim's exclusion argument is based upon the intuition that it is unlikely that a mass range of events in the real world is overdetermined as in the firing squad case. However, it seems that the firing squad

case is quite far-fetched when being used as an analogy to the mental-physical relationship. It seems that there are some kind of “special relation” - supervenience, determinate-determinables, or other kinds of realization - that attain between the mental and the physical. The causal autonomy picture is needed only when we hold on to Kim’s premises; if these premises are false, there is no reason to follow the causal autonomy picture.

As discussed above, the causal autonomy picture only guarantees that the mental has causal power over other mental effects, but not the physical effects. Non-reductive materialism theorists who are willing to grant the mental more power would find such a picture not satisfactory, for they would like to establish that the mental has a stronger power of making things happen, be it mental or physical. It seems sometimes more suitable to say that it is our intention that makes our muscle moves and raises our hands - the behavior of moving a limb does not seem to be separable into two layers naturally: it seems strange to say that the difference between our muscle cells moving in certain way and our hands are raised is the same as the difference between mental states and physical realizers. In such cases, List and Menzie’s picture would be more realistic. It is more reasonable to follow List and Menzies to argue that in different causal systems there exist different kinds of causal paths, and it is an empirical matter to decide which kind of scenario a causal system exemplifies. It is possible that in some causal systems the causal autonomy picture would obtain, but we should be careful to claim that it is a priori true to every system with mental causation.

3.2 Considerations on adopting the Interventionist Theory

3.2.1 A Sketch of the Interventionist Theory of Causation

In chapter 2 it is argued that Zhong (2014)'s move to an interventionist account of causation might be unnecessary, for his causal autonomy picture can be established with List and Menzies' approach, and so the standard counterfactual theory is not, as Zhong argued (2011), a failed cause. The later paragraphs will briefly introduce the interventionist theory of causation as related to dealing with issues of mental causation, and will address some of the problems that these theories might face. These problems raise the question about whether moving to interventionism is a good move for dealing with the exclusion problem.

The interventionist theories of causation are based on an intuition about the nature of causation: we often think that if C is the cause of E, changing C in some way, and with other factors remain unchanged, would lead to a change of E. It is easily conceivable that, if C is the cause of E, then it follows that, with other things remain unchanged, changes in C would lead to changes in E, and the removal of C would result in the absence of E. We can see the similarity between this intuition and the basic intuition for the counterfactual approaches. From a counterfactual point of view, to say that C is the cause of E, we would need a pair of counterfactual conditionals: that if C were to be present then E would be present, and if C were to be absent then E would be absent. The interventionist intuition is a variation on the counterfactualist idea: if C is a cause of E, with other things remain unchanged, we can see a change in E when a change in C occurs.

Such a theory is also called the manipulability theory, for it is possible to control the outcome E with manipulations on C. It is one of the guiding thought of designing and conducting experiments that aim at discovering causal relations. When designing comparative experiments we will have a group as the treatment group that receives the designated treatment, and a control group which receive no treatment or an alternative treatment, with two groups comparable with respect to other potentially relevant variables remain unchanged. Conductors observe the results of the experiment, and try to gain from it the information about the relation between the manipulated variable and the effect variable of interest. As Woodward (2008a) remarked, the interventionist and manipulability theories of causation “according to which causes are to be regarded as handles or devices for manipulating effects, have considerable intuitive appeal and are popular among social scientists and statisticians.” Although the manipulationist idea is basic to scientific researches and experiments, Woodward (2008a) also remarked that it is not very welcomed among philosophers: “recent philosophical discussion has been unsympathetic to manipulability theories: it is claimed both that they are unilluminatingly circular and that they lead to a conception of causation that is unacceptably anthropocentric or at least insufficiently general in the sense that it is linked much too closely to the practical possibility of human manipulation”. But in recent times, after Pearl (2000, 2009)’s formulation of the interventionist theory of causation using structural equations and causal models, and Woodward’s (2003) “different way of characterizing the notion of an intervention which does not make reference to the relationship between the variable intervened on and its effects”, there has been an increase of interest in using this approach to investigate on issues of causation in

philosophy. Besides causation and philosophy of science, applications of the interventionist theory of causation are also seen in other areas of philosophy, for example on issues about mental causation and the exclusion problem.

To illustrate a simplified picture of what interventionist causation is about, we here quote Woodward (2003, p. 98) (The following is Woodward's definition of what it means for I to be an intervention variable for X with respect to Y.):

1. I causes X.
2. I acts as a switch for all the other variables that cause X. That is, certain values of I are such that when I attains those values, X ceases to depend on the values of other variables that cause X and instead depends only on the value taken by I.
3. Any directed path from I to Y goes through X. That is, I does not directly cause Y and is not a cause of any causes of Y that are distinct from X except, of course, for those causes of Y, if any, that are built into the I-X-Y connection itself; that is, except for (a) any causes of Y that are effects of X (i.e., variables that are causally between X and Y) and (b) any causes of Y that are between I and X and have no effect on Y independently of X.
4. I is (statistically) independent of any variable Z that causes Y and that is on a directed path that does not go through X.

("Cause" in this characterization always means "contributing cause" rather than "total cause.")

Given the notion of an intervention variable, an intervention may be defined as follows:

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

In more plain words, the basic idea of an intervention on X with respect to Y is that it should ensure that any remaining covariation between X and Y must be due to a causal influence of X on Y. Clause 2, for example, ensures that covariation between

X and Y under the intervention is not accountable by a common cause of X and Y other than the intervention; clause 3 then ensures that the intervention itself does not amount to a common cause of X and Y; similarly, clause 4 requires that no covariation of X and Y should arise as a result of some correlation between the intervention and some other cause of Y.

Zhong (2014) applies Woodward's interventionist theory on issues concerning non-reductive materialism and the exclusion problem. In his 2011 paper, he gives an argument exposing the failure of Lewis's original counterfactual theory of causation to deal with a two-layer causal autonomy picture. In his 2014 paper, the limitation is mentioned again. Lewis's account takes $C \square \rightarrow E$ as automatically true, when C and E are both present in the actual world. Zhong points out that "the presence condition is non-trivial in that it "rules out insufficiently specific causes", and in "dual-condition conception of causation" he does not presuppose a Lewisian criteria of world similarity. (p.345) It is reasonable for us to say that Zhong's approach requires a certain degree of loosening of the centering requirement (as has been extensively discussed in chapter II of this thesis), or else it would not be possible for the different layers to have causal autonomy.

In chapter 2 of the thesis, I have already argued that it is unnecessary to adopt an interventionist framework to establish the causal autonomy picture. The modified counterfactual theory as proposed by List and Menzies (2009) is adequate to give us a picture of causal autonomy, with the two layers both having causal powers within their own levels and do not causally interfere with each other. Besides this, there are also other reasons for us to question the move to interventionist theories.

3.2.2 Pearl's Semantics and the Centering Requirement

One of the questions that arises with the adoption of the interventionist theory on the exclusion problem is that: which semantics should we use when we evaluate the relevant counterfactual conditionals that are used to define causal relations? For the counterfactual theories, Lewis's similarity semantics (1973) seems to be an orthodox one for many to adopt. List and Menzies also have given a formal proof under their modified semantics in the appendix of their (2009) paper. By contrast, Zhong did not explicitly mention which kind of semantics that he used to evaluate the relevant conditional sentences, and it seems that James Woodward does not have an original formal semantics in his 2003 treatise. Briggs (2012) observes that Woodward is a proponent of causal modeling account of semantics when dealing with interventionist theory of causation (p.142). Briggs argues that though small differences exist between different philosophers' respective accounts, they agree in the basic outline (ibid.). Therefore, it leaves us rooms to discuss the adoption of semantics and its consequences.

There are different constructions of semantics developed by theorists for the interventionist theories of causation. Judea Pearl's (2000; 2009) structural model is a representative one. In contrast with the Lewisian possible-world semantics, Pearl constructed the semantics using the mathematical language of structural equation models.

For the present purpose, it is important to note that Pearl's semantics, despite its differences from Lewis's, in effect observes Lewis's centering principle. Pearl

proposes three axioms for the logic of counterfactuals - composition, effectiveness, and reversibility. The property of Composition entails Lewis's centering principle.

According to Pearl,

Property 1 (Composition)

For any three sets of endogenous variables X , Y , and W in a causal model, we have

$$Wx(u) = w \Rightarrow Y_{xm}(u) = Y_x(u) \quad (7.19)$$

Composition states that, if we force a variable (W) to a value w that it would have had without our intervention, then the intervention will have no effect on other variables in the system. That invariance holds in all fixed conditions $\text{do}(X = x)$.

Since composition allows for the removal of a subscript (i.e., reducing $Y_{xw}(u)$ to $Y_x(u)$), we need an interpretation for a variable with an empty set of subscripts, which (naturally) we identify with the variable under no interventions. (p.229)

Pearl writes that the centering requirement, that the axiom (6) $A \& B \Rightarrow A \square \rightarrow B$ --- if A and B are both true in the actual world, then it follows that if A were to be the case, then B would be the case --- follows from the property of Composition (pp. 240-1). So in this sense, Pearl's semantics observes the requirement of centering. As discussed in chapter 2 of the thesis, and as have been explained by Zhong (2011), this requirement is incompatible with the picture of causal autonomy, because according to this semantics, it is impossible to have $M1$ to be a cause of $M2$ but not of $P2$. It follows that, if we are to adopt Pearl's structural-based counterfactual semantics, we would face the same problem again. Whether Pearl's system is also suitable for modifications that loosen the centering requirement, like what List and Menzies did to Lewis's original semantics, is yet unknown. However, in the previous

section we argued that a modified counterfactual theory can already have the causal autonomy picture satisfactorily established.

3.2.3 Brigg's Revised Semantics and Violation of Weak Centering

Rachel Briggs (2012) discusses the inadequacies of existing semantics for interventionist theories of causation. She writes, “A number of recent authors ... advocate a causal modeling semantics for counterfactuals. But the precise logical significance of the causal modeling semantics remains murky... The causal modeling semantics is both an account of the truth conditions of counterfactuals, and an account of which inferences involving counterfactuals are valid. As an account of truth conditions, it is incomplete. While Lewis’s similarity semantics lets us evaluate counterfactuals with arbitrarily complex antecedents and consequences, the causal modeling semantics makes it hard to ascertain the truth conditions of all but a highly restricted class of counterfactuals.” (p.139) Briggs tried to “explain how to extend the causal modeling language to encompass a wider range of sentences, and provide a sound and complete axiomatization for the extended language”, and by doing this, she finds out that it results in “serious consequences concerning valid inference. The extended language is unlike any logic of Lewis’s: modus ponens is invalid, and classical logical equivalents cannot be freely substituted in the antecedents of conditionals.” (ibid.)

The details of Briggs’ work do not matter here, but the following point is relevant to our discussion. In her semantics for interventionist counterfactuals, even the

condition of weak centering is violated (p.152). It is thus not clear either that Zhong's (2014) argument can be formalized based on Briggs' semantics.

Therefore, there is an important open question if we are shift to an interventionist theory of causation to deal with the Exclusion problem. Not that it is not doable, but interventionist theory of causation is still under debate and there has not been a very mature consensus when in terms of semantics. Moreover, it might face the same problem as Lewis's version of counterfactual theory would face for the abidance to the centering requirement / composition property. If undergone radical changes like Briggs' semantics, the consequences on the concerned issues are yet unknown. This version of semantics would violate even the weak centering principle. It is probably safer to just use the counterfactual theory with weak centering requirement to deal with the exclusion problem, as List and Menzies used it to deal with a more general picture of mental causation, and as I have discussed in chapter 2 of the thesis, the causal autonomy scenario that requires the non-overdetermination and physical closure limitations. The modified counterfactual theory may not be perfect - possibly no philosophical findings would ever be - but it seems good enough.

3.3 Interventionism and the Alleged Problems of Causal Systems with Supervenience

3.3.1 Baumgartner's Critique

As discussed above, a causal system that involves a special, non-causal determination relationship is much more complicated than the usual scenarios. Karen Bennett has gone to length to argue that the situation where both a mental property and a physical property are causes is not like at all firing-squad overdetermination. What the relationship between the mental and the physical is has yet to be universally agreed, but many philosophers with an inclination to accept non-reductive materialism would settle on a supervenience relationship. To put it simply, to say that A supervenes on B is to say that no A changes are possible without some changes that occur to B (though changes in B do not necessarily lead to changes in A). This is, as most philosophers would agree, a non-causal determination relationship.

However, interventionism was initially developed to characterize causal systems that involve variables that can be independently manipulated. The notion of intervention often tacitly assumes that an (ideal) intervention on a variable X would not directly affect other variables in the system. Such a notion becomes problematic when supervenience relationship is present. One of the prominent critiques of applications of interventionism to the exclusion problem, made by Michael Baumgartner in his series of papers (Baumgartner 2009, 2010, 2013), is built on this observation.

Baumgartner claims that he does not aim to refute non-reductive physicalism, and he does not aim to attack on the interventionist approach, but he thinks that these two are incompatible with each other (Baumgartner 2013 p.8). He argues that attempts to (dis)solve the exclusion problem from an interventionist point of view are “bound to fail” (2013 p.25). In our discussion, we focus on his most recently published paper

(2013), as his most refined critique on interventionism as applied to the exclusion problem thus far.

Baumgartner (2013) points out the basic characteristics for non-reductive materialism: (NR1) for every physical event/property which has a cause, it has a complete sufficient physical cause, (NR2) the mental supervenes on the physical without being completely reducible to the physical, and (NR3) the mental is able to cause physical effects of their own supervenience bases, in virtue of their mentalness. (pp.3-4)

Baumgartner uses Woodward 2003's definition (as mentioned in part 3.2.1 of the thesis) as the classical definition for the interventionist theory of causation that he had in mind.

Baumgartner's complaint is straightforward. Given the supervenience relationship posited in (NR2), no change of a mental state/property is possible without a change in the underlying physical state/property. It follows that no intervention variable on the mental variable can be independent of the underlying physical variable. By (NR1), however, the underlying physical variables presumably on a causal path to the effect variable of interest without going through the mental variable. Therefore, by Woodward's definition of interventionism, any intervention variable on the mental variable with respect to the effect variable of interest should be independent of the underlying physical variable. It follows that there can be no intervention variable on the mental variable with respect to the effect variable of interest, and so that the mental variable cannot be a cause of the effect variable of interest.

Baumgartner believes that his argument shows that rather than being useful to solve the exclusion problem and providing grounds for establishing power of causation for the mental, “the interventionist theory of causation gives rise to a self-contained interventionist exclusion argument, which even rests on weaker premises than Kim’s arguments” (2013, p.2).

3.3.2 Woodward’s Response and Its Problems

Woodward in his 2014 paper (a largely completed draft version of this paper has been available since 2011, and that version has been quoted in Baumgartner’s discussion in 2013) has developed a response to Baumgartner’s 2009 and 2010 versions of arguments, by attributing a kind of special status to the relationship of mental-physical supervenience. He responded to the worries of Baumgartner and refined his interventionist theory to account for supervenient properties and supervenient base properties. Woodward argued against Baumgartner that “one can’t simply assume that because it is appropriate to control for ordinary confounders in cases in which no non-causal dependency relations are present, it must also be appropriate to control for factors like supervenience bases which do represent non-causal dependency relations.” (2014 p.34) And he writes that, “if it is “metaphysically impossible” to change the value of a supervening variable like M1 while holding P1 fixed, then the very fact that this is impossible is itself an indication that counterfactuals with this antecedent do not tell us about the causal effect (or the absence of such an effect) of M1 on other variables” (p.33)

Woodward went to length to articulate his modified definitions. For the sake of this brief discussion, I will quote Baumgartner's reformulated definitions here. According to Baumgartner (2013), the new Woodward theory mainly consists of two parts:

“(IV*) I is an intervention variable for X with respect to Y iff I satisfies (IV.i), (IV.ii), (IV.iii*), and (IV.iv*):
(IV.i) I causes X;
(IV.ii) I acts as a switch for all the other variables that cause X;
(IV.iii*) any directed path from I to Y goes through X or through a variable Z which is related to X in terms of supervenience (or definition);
(IV.iv*) I is (statistically) independent of every cause of Y which is neither located on a path through X nor on a path through a variable Z which is related to X in terms of supervenience (or definition)”

and

“(M*) X is a (type-level) direct cause of Y with respect to the variable set V iff there possibly exists an (IV*)-defined intervention on X with respect to Y such that all other variables in V that are not related in terms of supervenience (or definition) to X or Y are held fixed, and the value or the probability distribution of Y changes.

X is a (type-level) contributing cause of Y with respect to the variable set V iff

(i) there is a directed path from X to Y such that each link on this path is a direct causal relationship and
(ii) there possibly exists an (IV*)-defined intervention on X with respect to Y such that all other variables in V that are not located on a causal path from X to Y or on a path from a variable Z to Y, such that Z is related in terms of supervenience (or definition) to X or Y, are held fixed and the value or the probability distribution of Y changes.”

(2013, pp. 6, 13-14)

Baumgartner concedes that Woodward's new version of interventionism blocks his original critique, but he denies that this revision can offer great help for the interventionist quest on conquering the exclusion problem. He writes, “even though

the newest version presented in Woodward (2011), i.e., (M*)-(IV*)-interventionism, is not only compatible with, but in fact entails (NR3), the support non-reductive physicalists ... can at best hope to receive from (M*)-(IV*)-interventionism is extremely slim.” (2013 p.24)

Baumgartner gave two main arguments. First, he argued that Woodward’s new version of interventionist theory, though permitting scenarios of M1 causing P2, would also allow scenarios where the causal impact of the concerned mental property “collapses onto the causal impact of its supervenience base after the first link on a corresponding causal chain” (p.21), if there exists some physical intermediary factors on the causal path (e.g. P1 --- P’ --- P2). So Baumgartner writes, “Thus, the ‘causal autonomy’ of a mental property that (M*)-(IV*)-interventionism allows for is restricted to the first physical link on a causal chain out of that mental properties’ supervenience base. Undoubtedly, that is a consequence of (M*)-(IV*)-interventionism that does not square nicely with the requirements of non-reductive physicalists who subscribe to (CAM). (M*)-(IV*)-interventionism only leaves room for a very limited sort of causal autonomy of the mental.” (2013 p.21)

Second, according to Baumgartner, Woodward’s 11/14 version also allows an epiphenomenal causal structure to generate “the exact same difference-making relations or correlations under possible interventions” as a downward causation structure does. He criticizes the interventionist non-reductive materialists’ tendency to always prioritize the causal autonomy picture and think that the epiphenomenalist structure can be discarded even in the absence of empirical evidence (2013 p.22).

Baumgartner thinks that supports for this prioritizing are unbased, as the

metaphysical support (claiming that epiphenomenal structures does not exist) is too strong, and appealing to non-metaphysical supports (like claiming that such prioritizing is just a matter of representational convention) is too weak, and he argues that if one is to "introduce the convention that a causal structure should always be represented by a minimal graph, i.e. a graph with the least amount of edges, which adequately reproduces the empirical behavior of that structure", then epiphenomenal structures would be "universally favored" over the autonomy one. Hence Woodward's new version of interventionism cannot satisfactorily serve the purpose of establishing the causal power of the mental. (2013 p.23)

3.3.3 Considerations about the Woodward-Baumgartner Debate

One moral we can get from the debates between Woodward and Baumgartner is that the interventionist theory has its own problems when dealing with the exclusion problem. Apart from the issues with formal semantics that has been discussed earlier, we see that the interventionist theory is not yet fully developed to deal with issues related to supervenience (and other similar kinds of non-causal determination relationship). Woodward's 2011/2014 discussion is helpful, but it seems that this still cannot stop Baumgartner's worries. The real problem probably lies in the fact that there is yet to be an clear and articulated way to apply interventionism to a causal system that involves a more complicated realization structure: in its incubation interventionism supposedly only deals with causal relationships between "metaphysically independent" variables. How to extend the framework, especially the formal framework such as structural equation models, to handle non-causal dependencies as well as causal dependencies, remains to be fully worked out.

The “classical” counterfactual theories, as discussed extensively in chapter 2, however, seem to have no analogous worries. Compared to the interventionist theories, the classical counterfactual theories seem to deal with supervenience (mental properties with physical supervenience base) scenarios more easily, and have the formal semantics for counterfactuals that are more suitable when dealing with causal autonomy scenarios (as compared to Pearl’s semantics that entails centering and Brigg’s that violates weak-centering). Considering this, it seems that before the advocates can give us a more refined interventionist theory to rigorously deal with supervenience, it does not seem that there is an advantage to switch from a classical counterfactual theory to an interventionist one in order to deal with the exclusion problem.

Concluding Remarks

As demonstrated in chapter 2, List and Menzies successfully show that Kim's a priori exclusion is false, as there are cases where the higher-level properties exclude the lower level, and there are cases where exclusion does not happen at all. Therefore exclusion is a contingent matter, and whether it obtains in the causal system of interest needs empirical investigations. It is also shown that such a result can be achieved with Lewis's original counterfactual theory of causation, without loosening the centering requirement, as List and Menzies' work does not require that the effect to be split into two layers, and that the different levels of properties are forbidden from "cross-causing", and they can only cause on the same level that they are respectively on. If the causal autonomy picture (M1 supervenes on P1, M2 supervenes on P2, M1 causes M2, P1 causes P2, M1 does not cause P2, P1 does not cause M2) is not required, then Lewis's original account of counterfactual theory of causation would be enough. As shown in chapter 2, all three scenarios - downward exclusion, upward exclusion, and compatibility - can be proved to be possible within Lewis's framework. And as discussed in chapter 3, the necessity for a causal autonomy picture, which arises from the requirement of "non-overdetermination", is put under serious doubt.

Where weak centering is needed, is when we aim to allow Zhong's structure of causal autonomy. Zhong's argument is more adhesive to Kim's original diagram, that he does not want to allow the mental to be able to cause the physical, upholding Kim's non-overdetermination requirement, and try the best to not violate the physical closure requirement, maintaining that on the microphysical level the physical must be caused by the physical. But such a picture, as discussed at the start of chapter 3, has been argued by many (Bennett being a sophisticated representative) to be poorly

motivated. As Bennett discussed at length, it is doubtful the relationship between mental and physical in scenes of mental causation is an overdetermination relationship. Moreover, the kind of causal autonomy envisaged by Zhong does not seem to adequately accommodate the intuitions about mental causation that give rise to the challenge in the first place. It seems that the mental and the physical are not equal competitors on a causal picture, that it is not analogous to classical overdetermination/preemption scenarios, like members of the firing squads executing a prisoner.

Zhong (2014) turned to an interventionist theory of causation to respond to the exclusion argument and establish a causal autonomy picture. However, beside that such a picture, as shown in chapter 2, can be established with the weak-centering Lewisian framework, which question the need to turn to alternative theory of causation (as foreshadowed in Zhong 2011), it also seems that, as discussed in chapter 2, that it is unclear that interventionism is superior to standard counterfactual theories for the aforementioned purposes, as the available interventionist logics of counterfactuals either maintain centering (Pearl 2000,2009) or violate weak centering (Briggs 2012). Adding such complications, it seems to put further doubts on whether we should move to an interventionist theory of causation to deal with the exclusion problem.

The Interventionist theory of causation is an important theory as it provides us with insights that other theories of causation might lack, and it is a theory of causation that most scientists and statisticians would agree upon, as they use this in everyday practices, that scientific findings through doing experiments that sorts out correlation

and causal relations follows in such a manipulability and intervention guiding principle. It would also shed light on discussions about mental causation, especially when philosophers want to use the findings of neural science and want to have dialogs with neuroscientists, psychiatrist and psychologists alike. But it seems that, as discussed in the thesis, it might not be as helpful as the original counterfactual theory of causation, which, as demonstrated in chapter 2, has already given a comprehensive response to Kim's exclusion argument. Therefore, it seems that Lewis's counterfactual theory would be enough for non-reductive materialists to respond to exclusionist doubts and soundly establish the frameworks that allows the mental to have strong causal powers. And even if one has a strong desire to adopt a causal autonomy picture, which is probably unnecessary and undesirable as discussed, it is also doable with only a slightest adjustment to Lewis's counterfactual theory, by loosening the centering requirement to a weak-centering one. To make a simple conclusion, Lewis's counterfactual theory is sufficient enough for non-reductive materialists to respond to Kim's exclusionist attack and soundly establish versions of non-reductive materialism in which the mental is sometimes strong enough to not only cause other mental effects, but also physical effects. This should sound enough for most non-reductive materialists.

References

- Baumgartner, Michael (2009). Interventionist Causal Exclusion and Non-reductive Physicalism. *International Studies in the Philosophy of Science* 23 (2):161-178.
- Baumgartner, Michael (2010). Interventionism and Epiphenomenalism. *Canadian Journal of Philosophy* 40 (3):359-383.
- Baumgartner, Michael (2013). Rendering Interventionism and Non-Reductive Physicalism Compatible. *Dialectica* 67 (1):1-27.
- Beebe, Helen; Menzies, Peter & Hitchcock, Christopher (eds.) (2009). *The Oxford Handbook of Causation*. Oxford University Press.
- Bennett, Karen (2003). Why the exclusion problem seems intractable and how, just maybe, to tract it. *Noûs* 37 (3):471-97.
- Bennett, Karen (2008). Exclusion again. In Jakob Hohwy & Jesper Kallestrup (eds.), *Being Reduced: New Essays on Reduction, Explanation, and Causation*. Oxford University Press.
- Briggs, Rachael (2012). Interventionist counterfactuals. *Philosophical Studies* 160 (1):139-166.
- Davidson, Donald (1970). Mental events. In L. Foster & J. W. Swanson (eds.), *Experience and Theory*. Humanities Press. 79-101. Reprinted in Davidson 1980, pp. 207–25
- Davidson, Donald (1980). *Essays on Actions and Events*. Oxford University Press.
- Hall, Ned (2004). Two concepts of causation. In John Collins, Ned Hall & Laurie Paul (eds.), *Causation and Counterfactuals*. The MIT Press. 225-276.
- Hall, Ned; Paul, L. A. & Collins, John (eds.) (2004). *Causation and Counterfactuals*. Cambridge, Mass.: MIT Press.
- Hohwy, Jakob & Kallestrup, Jesper (eds.) (2008). *Being Reduced: New Essays on Reduction, Explanation, and Causation*. Oxford University Press.
- Kim, Jaegwon (1993). *Supervenience and Mind*. Cambridge University Press.
- Kim, Jaegwon (2005). *Physicalism, or Something Near Enough*. Princeton University Press.
- Kim, Jaegwon (2007). Causation and mental causation. In Brian P. McLaughlin & Jonathan D. Cohen (eds.), *Contemporary Debates in Philosophy of Mind*. Blackwell. 227--242.
- Kim, Jaegwon (2010). *Essays in the Metaphysics of Mind*. Oxford University Press.
- List, Christian & Menzies, Peter (2009). Nonreductive Physicalism and the Limits of the Exclusion Principle. *Journal of Philosophy* 106 (9):475-502.
- Loewer, Barry M. (2007). Mental causation, or something near enough. In Brian P. McLaughlin & Jonathan D. Cohen (eds.), *Contemporary Debates in Philosophy of Mind*. Blackwell.
- Macdonald, Graham & Macdonald, Cynthia (eds.) (2010). *Emergence in Mind*. Oxford University Press.

- McLaughlin, Brian P. & Cohen, Jonathan D. (eds.) (2007). *Contemporary Debates in Philosophy of Mind*. Blackwell Publishing.
- Menzies, Peter & List, Christian (2010). The Causal Autonomy of the Special Sciences. In Cynthia McDonald & Graham McDonald (eds.), *Emergence in Mind*. Oxford University Press.
- Paul, Laurie Ann (2009). Counterfactual theories. In Helen Beebe, Peter Menzies & Christopher Hitchcock (eds.), *The Oxford Handbook of Causation*. Oxford University Press.
- Pearl, Judea (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press. Second Edition (2009).
- Shapiro, L. & Sober, E. (2012). Against proportionality. *Analysis* 72 (1):89-93.
- Won, Chiwook (2014). Overdetermination, Counterfactuals, and Mental Causation. *Philosophical Review* 123 (2):205-229.
- Woodward, James (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.
- Woodward, James (2004). Counterfactuals and causal explanation. *International Studies in the Philosophy of Science* 18 (1):41 – 72.
- Woodward, James (2008a). Causation and manipulability. *Stanford Encyclopedia of Philosophy*.
- Woodward, James (2008b). Mental causation and neural mechanisms. In Jakob Hohwy & Jesper Kallestrup (eds.), *Being Reduced: New Essays on Reduction, Explanation, and Causation*. Oxford University Press.
- Woodward, James (2009). Agency and Interventionist Theories. In Helen Beebe, Christopher Hitchcock & Peter Menzies (eds.), *The Oxford Handbook of Causation*. OUP Oxford.
- Woodward, James (2014). Interventionism and Causal Exclusion. *Philosophy and Phenomenological Research* 91 (1).
- Yablo, Stephen (1992). Mental causation. *Philosophical Review* 101 (2):245-280.
- Yablo, Stephen (1997). Wide causation. *Philosophical Perspectives* 11 (11):251-281.
- Zhong, Lei (2011). Can Counterfactuals Solve the Exclusion Problem? *Philosophy and Phenomenological Research* 83 (1):129-147.
- Zhong, Lei (2014). Sophisticated Exclusion and Sophisticated Causation. *Journal of Philosophy* 111 (7):341-360.