Philosophy and
Phenomenological Research

Philosophy and Phenomenological Research
Vol. LXXXIII No. 1, July 2011

Can Counterfactuals Solve the Exclusion Problem?

LEI ZHONG
University of Michigan, Ann Arbor

A quite popular approach to solving the Causal Exclusion Problem is to adopt a counterfactual theory of causation. In this paper, I distinguish three versions of the Causal Exclusion Argument. I argue that the counterfactualist approach can block the first two exclusion arguments, because the Causal Inheritance Principle and the Upward Causation Principle upon which the two arguments are based respectively are problematic from the perspective of the counterfactual account of causation. However, I attempt to show that the counterfactualist approach is unable to refute a sophisticated version (i.e. the third version) of the exclusion argument in that the Downward Causation Principle, a premise of the third exclusion argument, is actually implied by the counterfactual theory of causation. Therefore, even if other theories of causation might help the non-reductive physicalist to solve the exclusion problem, the counterfactual theory of causation cannot.

Introduction

Non-reductive physicalism is afflicted with the so-called 'Causal Exclusion Problem', which asserts that if mental properties are not physical properties, the putative causal power of the mental would be excluded by the causal work of the physical (see Kim 1993, 1998, 2005).[1] Whenever any mental property *M* is instantiated, it is realized by some particular physical property *P*. This physical property *P* seems to be responsible for producing the effects that *M* is supposed to cause. But then what causal work is left for *M* to do? The causal power of mental properties appears to be 'screened off' by that of physical properties, if *M* is irreducible to *P*.

[1] The exclusion problem is not a problem regarding the causal power of mental *events*, but the causal efficacy of mental *properties*. What matters is not whether mental events can cause anything, but whether mental events can cause it *qua* (or in virtue of) mental properties.

Most non-reductive physicalists are anxious to solve the exclusion problem. They don't take seriously the option of epiphenomenalism, for the causal efficacy of the mental is so crucial to human life. As Jaegwon Kim points out, since the possibility of human agency and human knowledge presupposes the reality of mental causation, "for most philosophers the causal efficacy of the mental is something that absolutely cannot be given away." (Kim 1998)

Whether the causal power of mental properties is excluded by that of physical properties substantially depends on what it is for one property *causes* another.[2] Among different accounts of causation, the counterfactual theory of causation is very influential (Lewis 1973, 2000; Collins *et al* 2004b). On this theory, *A* causes *B* if *B counterfactually depends* upon *A*.[3] A quite popular approach to solving the exclusion problem is to adopt the counterfactual theory of causation (LePore and Loewer 1987; Horgan 1989, 1997; Block 1990; Baker 1993; Antony and Levine 1997; Loewer 2002, 2007). According to this counterfactualist solution, an event *e* can counterfactually depend upon the instantiation of a mental property *M* rather than the instantiation of its particular physical realizer *P* on that occasion. If this is true, the counterfactual theory of causation can in this way vindicate the causal efficacy of the mental, even if mental properties are not identical with physical properties.

Opponents usually attempt to attack the counterfactualist solution by denying the counterfactual theory of causation as such (see Kim 1973, 1998). Certainly, the counterfactual theory of causation is far from conclusive, and incurs many objections and problems. However, it would be too quick to say that the counterfactual account has already failed. As recent research shows, many philosophers are making efforts to develop and elaborate the counterfactual theory of causation and trying to solve objections and counterexamples (Collins *et al* 2004). A future upgraded version of the counterfactual theory may be able, for example, to account for causal asymmetry, accommodate the phenomena of causal preemption, and distinguish genuine counterfactual dependence that is essential to causation from spurious counterfactual dependence. So, it might be fair to say that the debate over the counterfactual account of causation has not yet been settled down. An

---

[2] For the sake of brevity, I use the term 'causation' in a broad sense, referring to causal relations not only between events but also between properties. When I say "property *A* causes property *B*", I always mean that an event that instantiates *A* causes another event that instantiates *B* in virtue of *A* and *B*—or in Horgan's terminology, *A quauses B* (Horgan 1989).

[3] Counterfactual dependence is standardly regarded as a sufficient condition rather than a necessary condition for causation. See Lewis 1973.

objection to the counterfactualist solution that proceeds by rejecting this theory of causation is thus less than compelling. Since the more serious criticisms of a position are usually internal rather than external ones, a more powerful objection to the counterfactualist solution is to indicate some inconsistency or tension within this proposal without challenging the counterfactual theory of causation itself. This is what I will do in this paper.

I wish to stress that there are several different versions of the Causal Exclusion Argument which assume different premises. When philosophers discuss the exclusion problem, they are not always aware of this. Either they talk about only one version of the exclusion argument, or they think (mistakenly) that these versions are roughly equivalent. So, it might be helpful to distinguish different arguments for the exclusion problem. This paper will discuss three versions of the exclusion argument. In Section 1, I will attempt to show that the Causal Inheritance Principle upon which the first argument is based is problematic at least from the perspective of the counterfactual theory of causation. In Section 2, I will argue that the Upward Causation Principle which the second argument assumes doesn't hold in the framework of the counterfactual theory of causation. However, as I will discuss in Section 3, although the counterfactualist solution can block the first two exclusion arguments, it is unable to refute a sophisticated version (i.e. the third version) of the exclusion argument, which claims that mental properties are causally impotent because they have no power of doing downward causation that is necessary for the causal efficacy of the mental. I will indicate that the Downward Causation Principle, a premise of the third exclusion argument, is actually implied by the counterfactual theory of causation. Therefore, even if other theories of causation *may* help the non-reductive physicalist to block the third version of the exclusion argument, the counterfactual account of causation cannot.

1. Exclusion and Causal Inheritance

The first version of the exclusion argument, Argument 1, goes as follows:

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

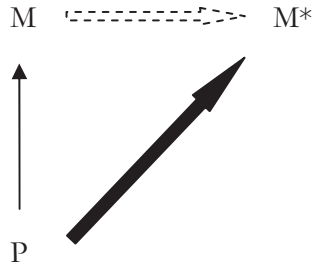
(CI) *Causal Inheritance*: If property *A* is instantiated on a given occasion by being realized by property *B,* then the causal power of *A* on this occasion inherits or derives from the causal power of *B*.

(NO) *Non-overdetermination*: No single event can have more than one sufficient cause occurring at any given time—unless it is a genuine case of causal overdetermination. And there is no systematic overdetermination in cases of mental causation.[4]

(I) *Irreducibility*: Mental properties are not identical with physical properties.

(Conclusion) The causal power of *M* is excluded by the causal power of *P*.

See the below diagram, Diagram 1, that illustrates the argument:



M is a mental property, and *M** represents a higher-order property, mental or behavioral, which is alleged to be caused by *M*.[5] *P* is a physical realizer of *M* on this occasion (*P* is a physical property in the supervenience base of *M*).[6] According to the Causal Inheritance Principle, the causal power of *M* derives from the causal power of *P*. And, given

[4] Causal overdetermination is when an effect has more than one cause, and each cause would have brought about the effect even if the other one had not done so. Consider the shootings of a tiger by two hunters. Suppose that the two shootings are set up in such a way that either would have killed the tiger if the other had failed. Then the death of the tiger is causally overdetermined by the two shootings. It would be *ad hoc* to suggest that *M* and *P* causally overdetermine the occurrence of *M** (see Kim 1998, 2005).

[5] We care about whether what we think can cause what we do and whether what we think earlier can cause what we think later. So I stipulate that *M** represents a mental or behavioral property.

[6] Here *P* is assumed to be a minimally sufficient set of subvenient base properties of *M*. *P* is a minimally sufficient set of subvenient base properties of *M* if and only if there is no proper subset P_1 of *P* such that *M* supervenes upon P_1 when *M* supervenes upon *P*.

the Non-overdetermination Principle and the Irreducibility Principle, it seems to follow that the causal power of *M* is excluded by that of *P*.

Let's consider the premises of the argument. The Supervenience Principle is widely-accepted, the denial of which would make mental-physical relationship extremely unintelligible. The Non-overdetermination Principle seems to be also pretty reasonable and is a standard view in contemporary philosophy of mind. Moreover, since I discuss whether non-reductive physicalism can save mental causation, I assume that the Irreducibility Principle is true for my purpose. Thus, the Causal Inheritance Principle is the only debatable thesis, which the counterfactualist solution attempts to attack.[7]

At first sight, the Causal Inheritance Principle is pretty intuitive. For example, someone dropped a glass cup on the wood floor, and then the cup was broken. Why was the cup broken? An apparent (partial) answer is: "Because the cup is fragile." But why is the cup fragile? It is because the cup has such and such molecular structure. Thus, the molecular structure that realizes fragility on this occasion appears to be a genuine cause of the cup's being broken. Granted that there is no overdetermination, if fragility is not identical with the molecular structure, then it seems to follow that the causal power of fragility would be excluded by the causal power of the molecular structure.

However, I want to indicate that the principle of causal inheritance is problematic from the perspective of the counterfactual theory of causation. Before I discuss this, let me say something about the counterfactual theory. On Lewis's original account, event *c* causes event *e* (where *c* and *e* are two *distinct* events[8]) if *e* counterfactually depends upon *c*—the causal relata are events rather than properties. But in our current discussion of mental causation, we need to modify the counterfactual theory of causation between events to be a theory of causation (i.e. quausation) between properties. So, on the modified counterfactual theory of causation, property *A* causes property *B* (where the instantiation of *A* and the instantiation of *B* are two distinct events or states) if *the instantiation of B counterfactually depends upon the instantiation of A* (i.e., *B would not have been instantiated on an occasion if A had not been instantiated on that occasion*). Moreover, according to the standard analysis of counterfactual conditionals, the conditional "if *A* had not been instantiated, then *B* would not have been instantiated" is understood as

---

7 (CI) is originally formulated by Kim. Kim writes: "If *M* is instantiated on a given occasion by being realized by *P*, then the causal powers of *this instance of M* are identical with (perhaps, a subset of) the causal powers of *P*." See Kim 1993, p. 208.

8 Two events are distinct from each other if there is no 'ontological' relationship between them: neither is realized by the other, neither is identical with the other, and neither is part of the other… See Lewis 1973; Kim 1973.

"*some* world where neither *A* nor *B* is instantiated (i.e., ($\sim A\&\sim B$)-world) is closer to the actual world than any world where *A* is not instantiated but *B* is instantiated (i.e., ($\sim A\&B$)-world)" (Lewis 1973).

To begin with, it is important to note that on the counterfactual theory of causation, *M*'s causing *M** (*if M had not been instantiated, then M* would not have been instantiated*) plus *M*'s being realized by *P* (*every ($\sim M$)-world is also a ($\sim P$)-world, and every P-world is also an M-world*) is logically compatible with the case that *P* does NOT cause *M** (*that if P had not been instantiated, M* would have been still instantiated*). In what follows, I will illustrate why the Causal Inheritance Principle is defeated by the counterfactual account of causation. In doing so, I will assume (reasonably) that mental properties are multiply realized by physical properties.

Now let me give an example. Consider a person, Ivan, who was in pain and hence walked to the hospital to see doctor. Suppose that the property of having such and such neurons firing in the anterior cortex (call this property '*P*') is the realizer of Ivan's pain in this case. Of course, pain can be realized in numerous different ways. Neuroscience suggests that other similar properties can also realize pain. For instance, when a person instantiates a property *Q* that is exactly like *P* except in which a particular firing neuron is different, she will also feel pain. Similarly, there could be countless such properties that also realize pain. In this example, which property, the property of being in pain, or the neural property *P*, is causally relevant to Ivan's going to the hospital?

It is reasonable to assume that any world where Ivan doesn't instantiate *P* and Ivan doesn't go to the hospital is more remote from the actual world than some world where Ivan instantiates *Q* rather than *P* (accordingly Ivan feels pain) and Ivan still goes to the hospital. So, *P* is not a good candidate for the cause of Ivan's going to the hospital. By contrast, a world where Ivan doesn't feel pain and he doesn't go to the hospital is probably closer to the actual world than any world where Ivan doesn't feel pain but still somehow goes to the hospital. Therefore, we should say that the property of being in pain is a better candidate for being the cause than property *P* in that Ivan's going to the hospital is counterfactually dependent upon Ivan's feeling pain, but not Ivan's being in *P*.[9] In this case, the causal power of pain can go

[9] Certainly, since counterfactual dependence is just a sufficient but unnecessary condition for causation, it doesn't follow from the *mere* fact that Ivan's going to the hospital doesn't counterfactually depend upon Ivan's being in *P* that the latter event is not the cause of the former event. But granted that there is no overdetermination in this case, since Ivan's walking to the hospital counterfactually depends upon another event (i.e., Ivan's being in pain) which occurred at the same time and place as Ivan's being in *P*, it is reasonable to rule out Ivan's being in *P* as the cause of Ivan's going to the hospital.

beyond, and hence doesn't inherit, the causal power of *P*. Thus, the Causal Inheritance Principle doesn't hold.

2. Exclusion and Upward Causation

Let's consider the second version of the exclusion argument, Argument 2:

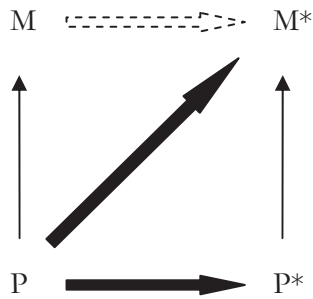
> (S) *Supervenience*: Mental properties (among other higher-order properties) supervene on physical properties. That is, if any system *s* instantiates a mental property *M* at *t*, there necessarily exists a physical property *P* such that *s* instantiates *P* at *t*, and necessarily anything instantiating *P* at any time instantiates *M* at that time.
>
> (UC) *Upward Causation*: If property *A* causes property *B*, then *A* would cause any supervenient property of *B* instantiated on this occasion.
>
> (CCP) *Causal Completeness of Physics*: If a physical event has a cause that occurs at *t*, it has a (sufficient) physical cause that occurs at *t*.
>
> (NO) *Non-overdetermination*: No single event can have more than one sufficient cause occurring at any given time—unless it is a genuine case of causal overdetermination. And there is no systematic overdetermination in cases of mental causation.
>
> (I) *Irreducibility*: Mental properties are not identical with physical properties.
>
> (Conclusion) The causal power of *M* is excluded by the causal power of *P*.

In Argument 2, the Causal Inheritance Principle is replaced by two new principles, the Causal Completeness of Physics Principle (the Completeness Principle, for short) and the Upward Causation Principle. Before we evaluate this argument, I want to say something about the Completeness Principle. First, we need to know what the term 'physical' means. Is the term 'physical' used in a broad sense or a narrow sense? Suppose that the term 'physical' here is used in the broad sense such that the physical not only include the fundamental

physical properties and entities (atoms, quark, electrons, etc.) which are the subjects of physics, but also include higher-order properties which supervene upon base physical properties, such as chemical properties, biological properties, physiological properties. Then mental properties can be physical properties in this sense, since mental properties also supervene upon and are realized by the base physical properties. The Completeness Principle thus fails to threaten the causal efficacy of mental properties. So, probably the term 'physical' should be understood in a narrow sense, which only denotes the fundamental physical properties, events, and entities.

Second, when the Completeness Principle says that "a physical event *e* that has a cause must have a physical cause", does it just mean that *e* must be caused by another physical event *c*? If the principle were understood in this sense, it would be too weak. Suppose that token physicalism is true: every mental event is a physical event. Even if there were a physical event that is caused by a mental event (also a physical event) solely in virtue of mental properties, this case is also compatible with the Completeness Principle understood in the weaker sense. But this is definitely odd. So, the principle should be read in a stronger sense, which says that a physical event *e* that has a cause must be caused by a physical event *c* in virtue of a physical property *P* that is instantiated by *c*.

Now consider Diagram 2 as below:



Given the Supervenience Principle, there must be a physical property *P** that realizes *M** on this occasion. And, given the Completeness Principle, *P** is caused by *P*. Then according to the Upward Causation Principle, *P* causes *M**. But since there is no overdetermination and *M* is not identical with *P*, the putative causal power of *M* would be excluded by that of *P*.

Moreover, since *M* supervenes upon *P* on this occasion, it cannot be suggested that there is a causal chain with *M* as an intermediate causal link: *P* causes *M*, and then *M* causes *M**. If this were the case, *M* would be a genuine cause of *M**, and hence the causal

efficacy of M would be preserved. But this option is not plausible. For it is standardly held that supervenience, which is typically a sort of synchronic or 'vertical' relationship, is different from causation, which is normally a kind of diachronic or 'horizontal' relationship (See Kim 1998).

As for the premises of the argument, the Completeness Principle is pretty reasonable and is also widely accepted by philosophers of mind. Moreover, as I said earlier, the Supervenience Principle, the Non-overdetermination Principle, and the Irreducibility Principle are all assumed to be true. Then only the Upward Causation Principle is left. This principle originated from Kim's discussion of what he calls the 'Causal Realization Principle':

> Given that an instance of M occurs by being realized by P, then 1) any cause of this instance of P must be a cause of this instance of M; and 2) any cause of this instance of M must be a cause of this instance of P (Kim 1993).

The Causal Realization principle consists of two sub-principles: the first is just the Upward Causation principle, and the second is what I call the 'Downward Causation Principle'. Kim seems to regard the two principles equivalent, but as I will show in the current and the next sections, they are not. I will discuss the second principle in the next section. Now let's focus on the Upward Causation Principle.

I will argue that a counterfactual theory of causation can reject the Upward Causation Principle. On the counterfactual theory of causation, even though M^* is realized by P^* (*every* $(\sim M^*)$*-world is also a* $(\sim P^*)$*-world, and every* P^**-world is also an* M^**-world*), and P^* is caused by P (*if* P *had not been instantiated, then* P^* *would not have been instantiated*), it doesn't logically follow that M^* is caused by P (*that if* P *had not been instantiated, then* M^* *would not have been instantiated*). For it could be shown that in the closest possible world where P is not instantiated (and hence P^* is not instantiated either), another physical realizer of M^*, Q^*, is instantiated so that M^* is also instantiated. That is, M^* would have still been instantiated even if P had not been instantiated.

Let's go back to the example of pain. On the actual occasion, Ivan's pain is realized by P (the property of having such and such neurons firing in the anterior cortex), and Ivan's walking to the hospital is realized by P^* (say, the property of having such and such fibers firing in the backbone). Suppose that Q, a physical property very similar to P, can also realize Ivan's pain, and Q^*, a physical

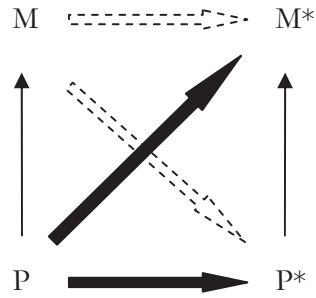
property very similar to P^*, can also realize Ivan's walking. Since P^* is caused by P, on a counterfactual theory of causation, in the closest possible world where P is not instantiated, P^* is not instantiated either. But since it is reasonable to assume that Q, a realizer of pain, and Q^*, a realizer of walking, would be also instantiated in this possible world, pain and walking would be thus instantiated in the same world. In a word, in the closest world where P is not instantiated, walking is still instantiated. That is to say, M^* doesn't counterfactually depend on P, even though M^* is realized by P^* and P^* is caused by P—M^* may counterfactually depend upon, and hence be caused by, M. So, the Upward Causation Principle is false.

3. Exclusion and Downward Causation

Now let's consider the third, and also the most sophisticated version of the exclusion argument, Argument 3:

> (S) *Supervenience*: Mental properties (among other higher-order properties) supervene on physical properties. That is, if any system s instantiates a mental property M at t, there necessarily exists a physical property P such that s instantiates P at t, and necessarily anything instantiating P at any time instantiates M at that time.
>
> (DC) *Downward Causation*: If property A causes property B, then A must cause any base property of B instantiated on this occasion.
>
> (CCP) *Causal Completeness of Physics*: If a physical event has a cause that occurs at t, it has a (sufficient) physical cause that occurs at t.
>
> (NO) *Non-overdetermination*: No single event can have more than one sufficient cause occurring at any given time—unless it is a genuine case of causal overdetermination. And there is no systematic overdetermination in cases of mental causation.
>
> (I) *Irreducibility*: Mental properties are not identical with physical properties.
>
> (Conclusion) The causal power of M is excluded by the causal power of P.

See Diagram 3 as below:



A higher-order property M^* is supposed to be caused by a mental property M. From the Supervenience Principle, we can know that M^* has some physical property P^* as its subvenient property on this occasion. Then according to the Downward Causation Principle, if M is a genuine cause of M^*, M must be a cause of P^*—that is, M causes M^* by causing M^*'s base property P^*.

Now let's turn to the (putative) M-to-P^* causation. From the mental-physical supervenience again, M has a physical property P as its supervenience base. According to the Completeness Principle, P is a genuine cause of P^*. From the Non-overdetermination Principle and the Irreducibility Principle, we must eliminate M as P^*'s cause. Since M is not a cause of P^* and since M can cause M^* only if M causes P^*, M cannot be a cause of M^* either. Therefore, the putative causal power of M is totally excluded by the causal power of P (Kim 1998, 2005).

Consider the first step of this argument, which assumes the Downward Causation Principle. Kim writes:

> To cause a supervenient property to be instantiated, you must cause its base property (or one of its base properties) to be instantiated. To relieve a headache, you take aspirin: that is, you causally intervene in the brain process on which the headache supervenes. That's the only way we can do anything about our headaches. To make your painting more beautiful, more expressive, or more dramatic, you must do physical work on the painting and thereby alter the physical supervenience base of the aesthetic properties you want to improve. There is no direct way of making your painting more beautiful or less beautiful; you must change it physically if you want to change it aesthetically—there is no other way (Kim 1998, pp. 42-3).

Kim maintains that such examples convincingly establish the thesis of downward causation. But some philosophers deny this. Consider below two slightly different and nonequivalent principles:

(R1) In order for X to cause a supervenient property to be instantiated, X must cause one of its base properties to be instantiated.

(R2) In order for X to cause a supervenient property to be instantiated, the instantiation of one of its base properties must be caused (Crisp and Warfield 2001).

(R1) is just the Downward Causation Principle. But some philosophers insist that the above examples Kim gives only favor (R2); it is unclear why they must support (R1). Crisp and Warfield put it this way:

> What's clear in the headache case is that if your headache is to go away, the brain processes on which it supervenes must be altered. But is it likewise clear that one and the same property must be causally responsible for both the disappearance of the headache and the alteration of the underlying brain processes? Why not suppose that (*a*) what causes the disappearance of my headache is my instantiation of the property being-such-as-to-have-just-taken-an-aspirin, a higher-order functional property realized by certain of my first-order physical properties, and that (*b*) what causes the alteration of my headache's underlying base properties is the instantiation of these first-order properties? This way of putting things looks to us to be coherent and so we find it far from obvious that Kim's examples specifically motivate (R1) over (R2) (Crisp and Warfield 2001, p. 311).

Crisp and Warfield's points actually suggest a general strategy to solve the exclusion problem, what I call 'the autonomy approach'. 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 (Yablo 1992,; Thomasson 1998; Marras 1998; Crisp and Warfield 2001; Gibbons 2006). On this approach, M causes M^* and P causes P^*, but M doesn't cause P^* and P doesn't cause M^*.[10] This approach is totally compatible with the Causal Completeness of Physics and the Non-overdetermination Principle. It seems to me that Kim doesn't provide any convincing arguments against this option.

But I don't plan to discuss the plausibility of the autonomy approach itself in this paper. The issue this paper is dealing with is not

[10] The autonomy solution doesn't have to hold the strong view that downward causation or upward causation *never* happens; in order to reject the exclusion thesis, the autonomy solution only needs to claim that cross-level causation doesn't *always* happen.

whether the exclusion problem can be solved, but whether it can be solved by the counterfactualist approach. In what follows, I will argue that the autonomy strategy is not available to the counterfactualist solution, as the counterfactual theory of causation implies the Downward Causation Principle. I put a tentative argument as follows:

(1) If M causes M^*, then M^* counterfactually depends upon M;

(2) M^* is realized by P^*;

(3) So, any ($\sim M^*$)-world is also a ($\sim P^*$)-world, and any P^*-world is also an M^*-world;

(4) So, if M^* counterfactually depends upon M, then P^* also counterfactually depends upon M;

(5) So, if M^* counterfactually depends upon M, then M causes P^*;

(6) (From (1) and (5)), if M causes M^*, then M causes P^* (i.e., *the Downward Causation Principle*);

(7) (From the Completeness Principle), P^* is caused by P;

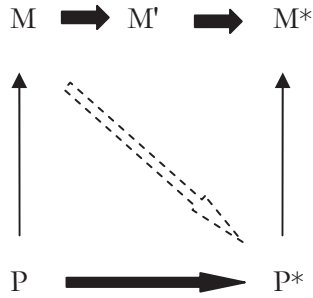
(8) (From (7), the Non-overdetermination Principle and the Irreducibility Principle), M doesn't cause P^*;

(9) (From (6) and (8)), therefore, M doesn't cause M^*.

The inference from (1) to (6) is supposed to establish the Downward Causation Principle. More specifically, it is important to notice the inference from (2) to (4), which is supposed to show that if M^* counterfactually depends upon M, then P^* would also counterfactually depend upon M. Why? Given that M^* is realized by P^*, any ($\sim M^*$)-world is also a ($\sim P^*$)-world, and any P^*-world is also an M^*-world. Then, a ($\sim M\&\sim M^*$)-world is at the same time a ($\sim M\&\sim P^*$)-world, and ($\sim M\&P^*$)-worlds are a subset of ($\sim M\&M^*$)-worlds (given multiple realizability). So, if it is the case that some ($\sim M\&\sim M^*$)-world, w, is closer to the actual world than any ($\sim M\&M^*$)-world, then it is also the case that w, as a ($\sim M\&\sim P^*$)-world, is closer to the actual world than any ($\sim M\&P^*$) world. So, if M^* counterfactually depends upon M, then P^* would also counterfactually depend upon M. Then it follows, together with other premises, that if M causes M^*, M must cause P^*. But since P causes P^*, M cannot cause P^* (given the Non-overdetermination Princi-

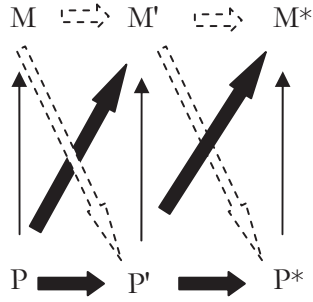
ple and the Irreducibility Principle). Therefore, *M* cannot cause *M** either.

Someone might be quick to point out that counterfactual dependence is just a sufficient but unnecessary condition for causation, so Premise (1) "if *M* causes *M**, then *M** counterfactually depends upon *M*" is false. In other words, it is possible that *M* causes *M** but *M** doesn't counterfactually depend upon *M*. See below Diagram 4:



M′ represents another higher-order property (mental or behavioral). Suppose that *M** doesn't counterfactually depend upon *M*, but *M** counterfactually depends upon *M′*, and *M′* counterfactually depends upon *M*.[11] Then *M* causes *M′* and *M′* causes *M**. Since causality is transitive, *M* causes *M** even though *M** doesn't counterfactually depend upon *M* (Lewis 1973). This is a case in which causation doesn't entail counterfactual dependence. And, since *M** doesn't counterfactually depend upon *M*, it doesn't follow that *P** counterfactually depends upon *M*. In this case, *M* causes *M**, but may not cause *P**. Thus the Downward Causation Principle seems to be problematic.

However, this suggestion doesn't really work, as it just pushes the problem one step back. Consider Diagram 5 as below:



[11] It is worth noticing that counterfactual dependence is not transitive. See Lewis 1973.

Given the Supervenience Principle, there must be a physical realizer of *M'*, *P'*. If *M'* counterfactually depends upon *M*, then *P'* would also counterfactually depend upon *M* (that is, *M* would cause *P'*). But since *P* causes *P'*, *P'* cannot be caused by *M* (given the Non-overdetermination Principle and the Irreducibility Principle). It follows that *M'* doesn't counterfactually depend upon *M* either (*M doesn't cause M'*). So, there would be no reasons left for regarding *M* as a cause of *M**. Therefore, I conclude that the counterfactualist approach fails to solve the third version of the exclusion problem. I put a formal argument as below:

(1) If *M* causes *M**, then either *M** counterfactually depends upon *M*, or there is a causal intermediary *M'* such that *M'* counterfactually depends upon *M* and *M** counterfactually depends upon *M'*;

(2) If *M** counterfactually depends upon *M*, then (as I discussed earlier on) *M* would cause *P**;

(3) If *M'* counterfactually depends upon *M*, then, by similar reasoning, *M* would causes *P'*;

(4) *P'* causes *P**;

(5) (From (3), (4) and the transitivity of causality), if *M'* counterfactually depends upon *M*, then *M* would causes *P**;

(6) (From (1), (2) and (5)), if *M* causes *M**, then *M* would cause *P**;

(7) But (as I discussed earlier), *M* cannot cause *P**;

(8) Therefore, *M* cannot cause *M**.

4. Concluding Remarks

In the preceding sections, I have discussed three versions of the exclusion argument. As I argued thus far, although a counterfactual theory of causation can help to block the first two arguments, it fails to reject the third version of the exclusion argument. For the counterfactual theory of causation actually implies the Downward Causation Principle, a premise of the third argument (given that other premises of the argument are unproblematic).

At the end of the paper, I want to remark on one possible response from the counterfactualist approach. The counterfactualist might reply that even though the current version of the counterfactual theory has to license the Downward Causation Principle, a future upgraded version, which will solve the putative problems with the counterfactual account, could coherently reject downward causation. I disagree. To see this, we should first consider those well-known problems and the possible solutions to them. Of course, I cannot discuss all of them here; for the purpose of illustration, I will just focus on a typical problem, the problem of effects.

Consider an example. Marie's having a disease causes her to have a fever. Suppose that the occurrence of the fever counterfactually depends upon the occurrence of the disease. Then the converse counterfactual *seems* also true to many people:

(i) If Marie had not gotten the fever, then she would not have gotten the disease.

And it seems to follow that the fever is the cause of the disease (according to the counterfactual theory of causation). But that is definitely odd. The so-called problem of effects (and other similar problems[12]) actually results from the troubling conditional as below:

(*a*) If *e* counterfactually depends upon *c*, then *c* would also counterfactually depend upon *e*.

Now let's turn back to our discussion of the exclusion problem. As I discussed earlier, in showing that the counterfactualist approach implies the Downward Causation Principle, my argument relies on a crucial thesis as below:

(*b*) If M^* counterfactually depends upon M, then the realizer of M^*, P^*, also counterfactually depends upon M.

---

[12] The problem of effects will generate a similar problem, 'the problem of epiphenomena'. Suppose further that Marie's having the disease is the common cause of her fever and rash (but the rash is not caused by the fever). If (i) were true, then the following counterfactual seems also true:

(ii) If Marie had not gotten the fever, then she would not have gotten the rash.

The reason is this. If you hadn't had the fever, you wouldn't have had the disease. But if you didn't have the disease, you wouldn't have gotten the rash. On the counterfactual theory of causation, it seems to follow that the fever causes the rash. But that is also odd.

Should, and could, the upgraded version of the counterfactual theory that will rule out (*a*) also deny (*b*)? I don't think so. I wish to indicate that there are two significant differences between (*a*) and (*b*). *First*, (*a*) and (*b*) have very different implications (given the counterfactual theory of causation). Whereas (*a*) leads to the troublesome thesis that if *c* cause *e*, then *e* causes *c*—an absurd consequence any account of causation must reject, the consequence of (*b*), "if *M* causes *M**, then M causes the realizer of *M**, *P**", is pretty intuitive to many philosophers. While the upgraded version would rule out (*a*) in order to avoid the problem of effects, it's unclear why it should also reject (*b*).

Second, and what's more important, whereas (*a*) is logically invalid, (*b*) seems to be a conceptual truth. As I showed in Section 3, if *M** counterfactually depends upon *M* and if *M** is realized by *P**, then it would logically follow that *P** also counterfactually depends upon *M*. How can any version of the counterfactual theory (including the upgraded version) reasonably reject (*b*)? By contrast, it doesn't logically follow from the fact that *e* counterfactually depends upon *c* that *c* also counterfactually depends upon *e*. Precisely because (*a*) is logically invalid, the counterfactualist can have the conceptual room to deny that *c* counterfactually depends upon its effect *e*, without giving up the basic framework of the counterfactual account of causation. For instance, according to Lewis's 'non-backtracking' reading of counterfactuals, it is less of a departure from actuality to get rid of *e* by holding *c* fixed and giving up some or other of the laws and circumstances in virtue of which *c* could not have failed to cause *e,* rather than to hold those laws and circumstances fixed and get rid of *e* by going back and abolishing its cause *c* (Lewis 1973). Thus, on this approach, if *e* had been absent, *c* would have still occurred just as it did but would have failed to cause *e*. Such solutions to rule out troubling sentences like (*a*) would leave (*b*) totally intact.[13]

Perhaps the counterfactualist might further reply that even though *P** counterfactually depends upon *M*, *P** is somehow not *caused* by *M*. But we must ask: Why? What distinguishes this case from other normal cases in which counterfactual dependence entails causation? It seems the only special feature of this case is just that *P** is the realizer of *M** which is alleged to be caused by *M*. But why does this matter to non-causality? If the counterfactualist intends to make a stipulation that whenever *A* causes *B*, *A* doesn't cause the realizer of *B*, that would be groundless and *ad hoc*.

So, we have good reason to believe that the counterfactualist solution even armed with the upgraded version still fails to block the third

[13] For other solutions, see Collins *et al* 2004b.

exclusion argument. Therefore, even if other theories of causation might help the non-reductive physicalist to solve the exclusion problem, the counterfactual theory of causation cannot. In this paper, I have left open the question of whether non-reductive physicalism can save mental causation—personally I'm pretty sympathetic with the position of non-reductive physicalism. All I have done is just to argue that the popular counterfactualist approach cannot solve the exclusion problem, and therefore the non-reductive physicalist should find other ways to save mental causation.[14]

References

Antony, L. and Levine, J. (1997) "Reduction with Autonomy," *Philosophical Perspectives* 11, pp. 83–105.

Baker, L. R. (1993) "Metaphysics and Mental Causation," in Heil and Mele 1993, pp. 75–95.

Block, N. (1990) "Can the Mind Change the World?" in G. Boolos (ed.), *Meaning and Method: Essays in Honor of Hilary Putnam*, Cambridge: Cambridge University Press, pp. 137–70.

Burge, T. (1993) "Mind-Body Causation and Explanatory Practice," in Heil and Mele 1993, pp. 97–120.

Collins, J., Hall, N. and Paul, L. A. (eds.) (2004a) *Causation and Counterfactuals*, Cambridge, Mass.: MIT Press.

Collins, J., Hall, N. and Paul, L. A. (2004b) "Counterfactuals and Causation: History, Problems, and Prospects," in Collins, Hall and Paul 2004a, pp. 1–58.

Crisp, T. M. and Warfield, T. A. (2001) "Kim's Master Argument," *Noûs* 35, pp. 304–16.

Gibbons, J. (2006) "Mental Causation without Downward Causation," *Philosophical Review* 115, pp. 79–103.

Heil, J. and Mele, A. (eds.) (1993) *Mental Causation*, Oxford: Clarendon Press.

Horgan, T. (1989) "Mental Quausation," *Philosophical Perspectives* 3, pp. 47–76.

—— (1997) "Kim on Mental Causation and Causal Exclusion," *Philosophical Perspectives* 11, pp. 165–84.

Kim, J. (1973) "Causes and Counterfactuals," *Journal of Philosophy* 70, pp. 570–2.

[14] In the future, I plan to explore two other prominent theories of causation: the regularity account and the probabilistic account, and discuss whether either of the two theories of causation can have the resources to reject the exclusion argument(s)—this could be a thesis on another occasion.

—— (1976) ''Events as Property Exemplifications,'' in M. Brand and D. Walton (eds.), *Action Theory*, Dordrecht: Reidel, pp. 159–77.
—— (1984) ''Concepts of Supervenience,'' *Philosophy and Phenomenological Research* 45, pp. 153–76.
—— (1993) ''The Non-Reductivist's Troubles with Mental Causation,'' in Heil and Mele 1993, pp. 189–210.
—— (1998) *Mind in a Physical World*, Cambridge: MIT Press.
—— (2005) *Physicalism, or Something Near Enough*, Princeton: Princeton University Press.
LePore, E. and Loewer, B. (1987) ''Mind Matters,'' *Journal of Philosophy* 84, pp. 630–42.
Lewis, D. (1973) ''Causation,'' *Journal of Philosophy* 70, pp. 556–67.
—— (2000) ''Causation as Influence,'' *Journal of Philosophy* 97, pp. 182–97.
Loewer, B. (2002) ''Comments on Jaegwon Kim's *Mind in a Physical World*,'' *Philosophy and Phenomenological Research* 65, pp. 555–662.
—— (2007) ''Mental Causation, or Something Near Enough,'' in McLaughlin and Cohen 2007, pp. 243–64.
Marras, A. (1998) ''Kim's Principle of Explanatory Exclusion,'' *Australasian Journal of Philosophy* 76, pp. 439–51.
McLaughlin, B. and Cohen, J. (eds.) (2007) *Contemporary Debates in the Philosophy of Mind*, Malden, MA: Blackwell Publishing.
Thomasson, A. (1998) ''A Nonreductivist Solution to Mental Causation,'' *Philosophical tudies* 89, pp. 181–95.
Yablo, S. (1992) ''Mental Causation,'' *Philosophical Review* 101, pp. 245–80.