Mental Causation and Exclusion: Why the Difference-making Account of Causation is No Help*

José Luis Bermúdez†
jbermudez@tamu.edu

Arnon Cahen‡
acahen@go.wustl.edu

ABSTRACT

Peter Menzies has developed a novel version of the exclusion principle that he claims to be compatible with the possibility of mental causation. Menzies proposes to frame the exclusion principle in terms of a difference-making account of causation, understood in counterfactual terms. His new exclusion principle appears in two formulations: upwards exclusion – which is the familiar case in which a realizing event causally excludes the event that it realizes – and, more interestingly, downward exclusion, in which an event causally excludes its realizer. This paper shows that one consequence of Menzies's proposed solution to the problem of mental causation is a ubiquitous violation of the principle of closure – a fact that forces him into a trilemma to which we see no satisfactory response.

Keywords: Mental causation, Exclusion, Difference-making accounts of causation, Causal Closure, Non-reductive Physicalism

1. Menzies's Exclusion Principle and the Problem of Closure

In a series of papers, most recently Menzies (2013), Peter Menzies has devel-

* While working on this paper we were saddened to learn of the death of Peter Menzies. We hope that this paper, which addresses just one of his many important contributions, will be taken in the spirit of admiration with which it was written. We are grateful to the editors and to Christian List, Antonella Corradini and Michael Schimtz for comments on a previous draft.

† Texas A&M University.
‡ The Open University of Israel.

oped a novel version of the exclusion principle that he argues avoids certain difficulties with, and counterexamples to, Kim's (2005) original formulation, and that is compatible with the possibility of mental causation. Unlike earlier versions of the principle, which do not incorporate a robust account of causation, Menzies proposes to frame the exclusion principle in terms of a difference-making account of causation, understood in counterfactual terms. His new exclusion principle appears in two formulations: upwards exclusion – which is the familiar case in which a certain event causally excludes the event that it realizes – but, more interestingly, downward exclusion, in which an event causally excludes its realizer. According to Menzies, the availability of this latter formulation shows that, when properly understood, causal exclusion is not a threat to mental causation. It is possible for a mental event to causally exclude its physical realizer (rather than, as is traditionally understood, the other way around). In this paper, we explore Menzies's new exclusion principle and argue that one consequence of his proposed solution to the problem of mental causation is a ubiquitous violation of the principle of closure – a fact that forces him into a trilemma to which we see no satisfactory response.

1.1 Menzies on the Exclusion Principle

The exclusion principle was originally introduced as a component in an argument for the causal inertness, or the epiphenomenalism, of the mental within a non-reductive physicalist metaphysics. In brief, the argument for epiphenomenalism of the mental proceeds from the conjunction of the following independently plausible claims:

> *Non-reductive physicalism* – Mental events are realized by *distinct* physical events
> *Closure* – Every physical effect has a sufficient physical cause (Papineau, 2009, p. 53)
> *Exclusion* – If an event *e* has a sufficient cause *c* at *t*, no event at *t* distinct from *c* can be a cause of *e* (Kim, 2005, p. 17).

Given closure, any purported instance of mental causation – say, where some mental event, M, appears to cause some physical behavior, B – has a sufficient physical cause, N. Yet, if N is a sufficient cause of B (at *t*), then, by exclusion, no event distinct from N can be a cause of B (at *t*). Therefore, the mental event, M, which according to non-reductive physicalism is distinct from its physical realizer, is excluded as a cause of B, even if N is the realizer of M. Given that this is a

fully general argument, it follows that the mental in general is epiphenomenal.

As mentioned, Menzies confronts this argument by denying Kim's original formulation of the exclusion principle. In particular, he notes that Kim's exclusion principle fails to make a distinction between causal sufficiency and causation (p. 71). For Menzies this distinction is pivotal, because he believes that mental events can be causally efficacious in generating behavior, despite being realized by physical events that are causally sufficient for the physical behavior in question.

The significance of the distinction between causal sufficiency and causation proper can best be appreciated in the following quotation from Yablo (1992):

> Notice some important differences between causal relevance and sufficiency, on the one hand, and causation, on the other: x can be causally sufficient for y even though it incorporates enormous amounts of causally extraneous detail, and it can be causally relevant to y even though it omits factors critical to y's occurrence. What distinguishes causation from these other relations is that causes are expected to be commensurate with their effects: roughly, they should incorporate a good deal of causally important material but not too much that is causally unimportant. (pp. 273–274)

To illustrate this distinction consider one of Yablo's examples from the same paper (also discussed by Menzies). Yablo asks us to consider a pigeon trained to peck at all and only red objects. When presented with a crimson object, the pigeon pecks at it, and the question is whether it was the redness of the object or its crimsonness that caused the pigeon's pecking. Here is how Menzies describes the example:

> The exclusion principle would say that since being red is realized by being crimson and being crimson is causally sufficient for the pigeon's pecking, the redness of the target is not the cause. But this seems wrong, as Yablo points out: the target's being red is of the right degree of specificity to count as a cause of the pigeon's action. In contrast, the target's being crimson is too specific to count as the cause: citing it as the cause of the pecking might give the erroneous impression that the pigeon would not peck at anything non-crimson. (Menzies, p. 72)

In sum, the object's being crimson is causally sufficient for the pecking, yet it seems to be the wrong event to mention as a cause. After all, the pigeon would have pecked had the object been differently colored, so long as it were

some determinate shade of red.

There is room to take issue with Menzies's use of Yablo's example, since it is not clear that we should think about the relation between red and crimson in terms of realization. To many (including Yablo) it seems more accurate to think of crimson as a determinate of red, rather than as a realizer of red. Nonetheless, the distinction between causation and causal sufficiency comes across very clearly, as does the idea that an event could have causally sufficient antecedents that might not properly be described as the causes of the event.

A further problem with the principle of exclusion, as originally formulated, arises from what Menzies refers to as the transmission of causal sufficiency across realization. The principle states that if some event M is causally sufficient for some behavior B, then its realizer, N, is also causally sufficient for B. If an object's being red is causally sufficient for the pigeon's pecking, then being crimson (or any other realizer of red) is likewise causally sufficient for the pigeon's pecking. Given that being red and being crimson are distinct (compresent) events, we have a clear conflict with the principle of exclusion. We have two distinct events, being red and being crimson, each of which is causally sufficient for the effect, and therefore each of which causally excludes the other. The problem is further compounded by the fact that the realization relation is omnipresent in the physical world – it goes all the way down, so to speak. Exclusion thus implies a (perhaps unending) chain of mutually excluding events.

Additionally, if we wish to avoid this mutual causal exclusion by granting causal priority to the realizer over the realized, as many physicalists do and as Menzies makes explicit in his first revision of the exclusion principle,[1] then we reach the conclusion that there is no causation anywhere but at the most fundamental physical level. And if, as some physicists believe to be a possibility, there is no fundamental physical level, then there can be no causation at all! This is the problem that Ned Block has aptly named the problem of causal drainage (Block, 2003).

To avoid these difficulties, Menzies argues we must reject the above formulation of the principle of exclusion, which infers the causal exclusion of some event from the causal sufficiency of a distinct event. Causal sufficiency is too weak to generate exclusion. Instead we need to formulate exclusion using a

[1] Menzies's first revision of the exclusion principle states that «if a mental state M is realized by a distinct physical state P that is causally sufficient for B, then M does not cause B» (p. 64). This he takes to hold true even if both M and P are causally sufficient for B.

stronger notion of causation. In order to capture just the right amount of detail required for genuine causation, Menzies appeals to a difference-making criterion for causal relevance.[2] According to Menzies,

> Truth conditions for causal relevance (or making a difference): The state S1 makes a difference to the state S2 in the actual world just in case (i) if in any relevantly similar possible situation S1 holds, S2 also holds; and (ii) if in any relevantly similar situation S1 does not hold, S2 does not hold. (p. 73)

Or, in terms of counterfactuals:

> The S1 makes a difference to S2 in the actual world if and only if it is true in the actual world that (i) S1 holds $\Box\rightarrow$S2 holds; and (ii) S1 doesn't hold $\Box\rightarrow$S2 doesn't hold. (p. 74)

The first condition aims to rule out events that are not specific enough (for example, the thought that the object's being colored is the cause of the pecking), whereas the second aims to rule out events that are too specific, which though perhaps causally sufficient for the effect, are too fine grained to be considered causes of the effect (such as the thought that the object's being crimson is the cause of the pecking).[3]

At this point, Menzies is ready to introduce his new, and improved, version of exclusion, which focuses on causation rather than causal sufficiency.

> Revised exclusion principle: For all distinct states S and S* such that S* is realized by S, S and S* do not both cause state T. (p. 77)

This allows for two formulations:

[2] Menzies is inspired by, and elaborating upon, interventionists accounts of causation, most significantly as they are developed in Woodward (2003, 2006). See also List and Menzies (2009) in which the ties to interventionism are even more explicit.

[3] It is worth mentioning a potential terminological confusion that may arise here. Though Menzies presents the difference-making criteria as criteria for *causal relevance*, the counterfactuals involved aim to capture both of the conditions for causation identified by Yablo above, causal sufficiency as well as causal relevance. The first counterfactual relates to the sufficiency condition on being a cause, and the second relates to the relevance condition. The reason Menzies's discussion is focused on the notion of causal relevance should be understood in the context of his dissatisfaction with the original exclusion principle, which merely mentions causal sufficiency. To supplement the exclusion principle with a genuine notion of causation we must provide an account of causation that does not only guarantee causal sufficiency, but also secures causal relevance. Thus, Menzies's difference-making criteria are criteria for being a cause, not merely for causal relevance, though it is the latter notion that is central in the context of his project.

> Revised exclusion principle (upwards formulation): If a state S causes a state T, then no distinct state S* that supervenes on S causes T.
>
> Revised exclusion principle (downwards formulation): If a state S causes a state T, then no distinct state S* that realizes S causes T. (ibid.)

The difference-making criterion, above, allows us to specify the cause of some event with the correct degree of detail and the revised exclusion principle makes room for the possibility that the realizer, rather than the realized event, is excluded as a cause.

Let us look at one such example. Suppose a certain intention to move one's hand, M, is a sufficient cause of one's hand moving, B. Given non-reductive physicalism, the intention M is realized by some distinct neural event N_{47}, which is also causally sufficient for event B (by the principle of transmission of causal sufficiency across realization). However, to establish which of the two causally sufficient events is the cause of B, we must evaluate their respective difference-making counterfactuals.

The first pair of counterfactuals relates to the mental event:

(1) $M \mathrel{\square\!\!\rightarrow} B$
(2) $\sim M \mathrel{\square\!\!\rightarrow} \sim B$

The second pair of counterfactuals relates to the realizing neural event:

(3) $N_{47} \mathrel{\square\!\!\rightarrow} B$
(4) $\sim N_{47} \mathrel{\square\!\!\rightarrow} \sim B$

According to Menzies, the first two conditionals regarding the intention, M, are true. In all the closest possible worlds where M holds, so does B, and in all those closest possible worlds in which M does not hold, where there is no intention to move one's hand, B also does not hold, i.e., the hand is not moved. This is not the case with the pair of counterfactuals relating to the neural realizer. The first counterfactual (3) is true – in all those closest possible worlds in which N_{47} obtains, so does the effect B. Yet, the second counterfactual (4) is false – in those closest worlds where N_{47} does not obtain it does not follow that the effect B does not obtain. On the contrary, in those worlds in which $\sim N_{47}$, the intention M is realized by a different neural realizer, say N_{48}, and the hand still moves. Given that the difference-making counterfactuals relating to the intention are satisfied, we can conclude that the cause of the moving of the hand is the intention to move the hand. In contrast, the difference-making counterfac-

tuals relating to the neural realizer of the intention are not both satisfied. The second counterfactual fails, thus indicating that the neural event is too specific to serve as the cause of the moving of the hand. Therefore, by the revised exclusion principle, the neural realizer is causally excluded from being the cause of the moving of the hand.

According to Menzies, this serves as a solution to the problem of mental causation, avoiding the original pitfalls of exclusion by reversing the causal exclusion relation.

1.2. Downward Exclusion and the Violation of Closure

The previous section presented Menzies's strategy of supplementing the principle of exclusion with a difference-making account of causation in order to block the common epiphenomenalist objection to non-reductive physicalism. However, we believe that the price is too high. The manoeuvre secures the causal efficacy of the mental by excluding the causal efficacy of its physical realizer. The inevitable consequence, therefore, is a ubiquitous violation of the principle of closure.

As we saw in the example above, a consequence of downward exclusion is that the physical realizer of the intention, while causally sufficient for the movement of the hand, is not the cause of the movement. Yet, according to closure, every physical event must have a sufficient physical cause. Thus, the result of downward exclusion is a violation of closure. We will have such a violation of closure in every case of mental causation that involves downward exclusion. (Indeed, in any case where the difference-making counterfactuals relating to the physical realizer are not satisfied.)

Admittedly, Menzies makes room also for cases of mental causation that do not involve downward exclusion. These are cases in which the difference-making counterfactuals relating to *both* the realized and the realizer are satisfied. Such cases correspond to Menzies's 'Compatibility Result': «If M causes B, then N causes B if and only if the causal relation between M and B is realization-sensitive» (p. 78), where the relation between M and B is realization-sensitive if behavior «...B fails to hold in all those M-worlds that are closest ~N-worlds (i.e. where M has a different realizer from the actual one)» (ibid.). In other words, we have mental causation without downwards exclusion precisely when the occurrence of B depends upon M being realized by N.

However, instances of the Compatibility Result are supposed to be rare. Furthermore, to the extent that such cases exist they are uninteresting from the

point of view of the problem of mental causation as it arises for non-reductive physicalism. As Menzies says:

> When might we expect the conditions for realization-sensitivity to obtain? If the mental property M were identical to the neural property N, then we would certainly expect instances of M to stand in realization-sensitive causal relations with respect to instances of N. The fact that M-instances had certain effects when and only when N-instances are present would simply reflect the identity of the properties. (p. 79)

Certainly, if M is identical to its realizer N, then the casual efficacy of the one just is the causal efficacy of the other. Yet non-reductive physicalism is the claim that M and N are distinct events. It is on the assumption that they are distinct that the causal efficacy of the mental is brought into question in the light of the exclusion principle.

How then might Menzies respond to the evident conflict between the principle of closure and his difference-making formulation of exclusion? We see three possible responses.

> (1) Bite the bullet:
> Menzies can accept the conflict between the principle of closure and the downwards formulation of the exclusion principle, and maintain that in the majority of cases involving genuine mental causation (excluding those rare cases corresponding to the Compatibility Result) closure is violated.
>
> (2) Revise the principle of closure:
> Unlike exclusion, which Menzies argued requires a robust notion of causation, perhaps the principle of closure only requires the weaker notion of causal sufficiency. This weaker principle would require that for every physical event there is a causally sufficient physical event.
>
> (3) Deny the rarity of the Compatibility Result:
> Menzies could also retain the original formulation of closure in terms of the availability of a sufficient cause, and avoid its violation by acknowledging that in *all* cases of genuine mental causation the relation between the mental event and its effect is realization-sensitive.

In the remainder of the paper we will explore each of these strategies and show why none will preserve Menzies's strategy as a viable solution to the problem of mental causation. We begin in the next section with an evaluation of the first strategy.

2. Violating Closure

In this section we put aside those supposedly rare cases in which the relation between the mental event and the behavior it causes is realization-sensitive, and focus only on those cases in which the difference-making counterfactuals are both true for the realized mental event, but are not both true for the realizing physical event. These are cases of genuine downward exclusion according to Menzies's new exclusion principle.

In such cases, the evaluation of Menzies's difference-making counterfactuals indicates that the behavior, B, would have occurred whether or not N_{47}, the physical realizer of M, the intention to so behave, had occurred. This holds because, in those cases in which the physical realizer of the intention failed to occur, N_{48}, which is a different realizer of M, would have occurred, giving rise to the very same behavior. But then, at least with respect to the occurrence of the behavior in question, there is nothing that the physical realizer causally or explanatorily contributes. As Kim says, when presenting the original (upward) exclusion argument, the fact that the behavior would have occurred regardless of «...whether or not the rationalizing belief and desire occurred surely demonstrates the causal and explanatory irrelevance of the belief and desire» (Kim, 1989a, p. 82). By parity of reasoning, the fact that the behavior would have occurred regardless of whether or not N_{47}, the physical realizer of the intention, occurred «surely demonstrates the causal and explanatory irrelevance» of the physical realizer (with respect to the behavior in question).[4] Or, to paraphrase Menzies's own words «...if the [intention] did all the causal work, it would appear that [its realizer], which we are assuming is numerically distinct from its [less] specific [realized mental state], has no causal role and so is epiphenomenal» (p. 62).

Yet the causal closure principle requires that the behavior in question – as a physical effect – does have a sufficient physical cause. By Menzies's exclusion argument, it appears we must reject the most plausible candidate physical cause

[4] The parenthetical qualification is just to note that the conclusions with respect to the physical realizer in the case of downward exclusion are weaker than those drawn from the original exclusion principle with respect to the mental. The conclusion of the original exclusion argument is epiphenomenalism of the mental. In the case of downward exclusion the conclusion is doubly limited. It does not apply to physical events in general, but only to those physical events that are realizers of mental events whose relation to the behavior they cause is realization-insensitive, and it is limited to their causal and explanatory role with respect to those behaviors, leaving open the question of their causal and explanatory roles more generally. That is, N_{47} might be 'causally and explanatory irrelevant' with respect to the movement of the hand, but not so with respect to various other neural occurrences.

of the behavior, namely the physical realizer of the intention. The behavior is therefore a counter example to closure.

Interestingly, in his contribution to this issue, Menzies appears to endorse this outcome of his position. He accepts that if we understand the principle of closure as requiring that for each physical effect there is a difference-making physical cause, then the principle of closure is violated in any case of realization-insensitive mental causation.[5] As he says:

> I take [Downward Exclusion] to provide a basis for saying that the reformulated causal closure principle (5) [in terms of difference-making physical events] is not true; or alternatively, that physical effects need not always have physical difference-making causes. (p. 27 of proofs)

And

> If this reasoning is correct, then it follows that wherever a mental event is a realization-insensitive cause of a physical effect, we will have a violation of the modified causal closure principle. (p. 28 of proofs)

Yet the fact that closure is violated in any case of downward exclusion seems rather a steep price to pay. After all, as Kim (1989b) emphasizes, to deny the assumption of 'the causal closure of the physical domain' «...is to accept the Cartesian idea that some physical events need nonphysical causes, and if this is true there can in principle be no complete and self sufficient physical theory of the physical domain. If the causal closure failed, our physics would need to refer in an essential way to nonphysical causal agents, perhaps Cartesian souls and their psychic properties, if it is to give a complete account of the physical world. I think most physicalists would find that picture unacceptable» (pp. 43–44). We agree with Kim that no physicalist, including non-reductive physicalists, can accept such a picture of physical reality.

This is especially significant, for, as we have seen, Menzies attempts to secure the causal efficacy of the mental within a non-reductive physicalist metaphysics. Not only is the principle of closure a central tenet of physicalism, it is also the central motivation for the 'physicalism' in non-reductive physicalism.

[5] List and Spiekermann (2013) similarly deny the principle of closure within a more general social science context. They argue that «once causation is understood as difference making, the causal-closure and exclusion principles are by no means conceptual truths about causation, but rather contingent principles that may apply to some causal systems but not to others» (p. 637).

Non-reductive physicalism is the preferred metaphysical position for those who attempt to locate the mental within the causal nexus while acknowledging that the mental is distinct from the physical and that the causal nexus is solely composed of physical events – i.e., acknowledging closure. If the causal efficacy of the mental conflicts with the causal closure of the physical world, the very motivation for physicalism, reductive or otherwise, is loosened.

In Menzies own words:

> Non-reductive physicalists express the hope that they can explain how mental causation is possible within the austere metaphysical framework of physicalism while avoiding the reductionism of the identity theory. Indeed, they hope that it is possible to vindicate not only the reality of mental causation, but also its independence and autonomy from physical causation. (p. 60)

We submit that the independence of mental causation from physical causation that results from Menzies's exclusion principle tears apart 'the austere metaphysical framework of physicalism'.

We have argued that cases of downward exclusion provide counterexamples to the principle of causal closure, which is a central tenet of physicalism. However, this was on the assumption that closure is to be understood as requiring that every physical effect have a sufficient physical cause. Downward exclusion violates closure by denying that the realizing physical event is a cause of the behavioral effect, yet it nonetheless allows that the realizing physical event is *causally sufficient* for the behavioral effect. Causal sufficiency, unlike causation proper, is not excluded; on the contrary it is transmitted across realization. Given that causation implies causal sufficiency (Menzies, p. 63), the intention that is the cause of the behavior is also causally sufficient for the behavior, and given transmission of causal sufficiency across realization, the physical realizer of the intention is also causally sufficient for the behavior. A rebuttal of the difficulties raised above suggests itself: rather than take closure to require the existence of some physical cause for every physical effect, it should only require the existence of some causally sufficient physical event for every physical effect. In the next section we turn to evaluate this weaker formulation of closure.

3. Revising the Closure Principle

Section 2 has shown that the price of biting the bullet and abandoning the causal closure of the physical is too high for Menzies. In this section we turn to the second of the three strategies identified in Section 1. As an alternative to biting

the bullet and abandoning the principle of Causal Closure, Menzies can exploit the distinction between causation and causal sufficiency to give a revised version of the principle.[6]

The basic tenet of Menzies's position is that a physical event can have causally sufficient physical antecedents that do not, properly speaking, cause that event. Perhaps it is enough for the causal closure of the physical that all physical events have causally sufficient physical antecedents? That suggests the following principle.

> *Revised Closure* – For every physical effect there is a causally sufficient physical event

Revised Closure says nothing about causation and plainly allows for the possibility that the causally sufficient physical event is not the genuine cause of the physical effect – exactly as Menzies maintains to be the case in all instances of mental causation where the realizing physical event is not identical to the realized mental event.

How should we evaluate Revised Closure? The only way we can think of is to look behind the original Causal Closure principle to what motivates it and then consider whether Revised Closure does justice to those motivations.

The original Causal Closure principle is often described as affirming the Completeness of Physics. The Completeness of Physics, in turn, is really a thesis about explanation. It says, in effect, that there is no need to go outside physics in order to explain a physical effect. For that reason the Completeness of Physics is, one might say, a regulative ideal of philosophical naturalism. And the most natural justification for the Completeness of Physics come from the combination of two claims. The first claim is that all explanation in physics is causal explanation. The second is the Causal Closure principle.

By the same token, many philosophers have argued that the Causal Closure principle is, in the words of David Papineau, «a highly empirical claim, whose acceptance derives from detailed empirical evidence about the causes of physical effects» (Papineau, 2009, p. 60). Another way of putting this would be to say that the best evidence for Causal Closure is the Completeness of Physics – the fact that, up to now, there has not been any need to go outside physics in order to explain any physical effect.

So, although one is a principle about metaphysics and the other a principle about explanation, Causal Closure and the Completeness of Physics are very

[6] This possibility is identified but not explored in Weslake, forthcoming.

closely related. Plausibly they stand or fall together. This is certainly how many philosophers have viewed the matter. Here are two representative passages from Jaegwon Kim.

> Physics is causally and explanatorily self-sufficient; there is no need to go outside the physical domain to find a cause, or a causal explanation, of a physical event. (Kim, 2005, p. 16)

> The only thing that physical causal closure protects is the causal and explanatory self-sufficiency of the physical domain. (Kim, 2009, p. 39)

Our proposal, therefore, is that we evaluate Revised Closure by looking at its implications for causal explanation. Does Revised Closure stand to the Completeness of Physics in anything like the relation that Causal Closure does?

With this question in mind, let us consider the cases that Menzies claims to be possible. These are cases where a physical effect has causally sufficient physical antecedents, but its genuine cause is distinct from those physical antecedents. The question we would like to pose is: Which way does the explanation go when cause and causally sufficient antecedents diverge in the way that Menzies suggests occurs in the vast majority of cases of mental causation?

There are two options. The first is that causal explanation tracks the genuine cause, not the causally sufficient physical antecedents. So, in order to explain the movement of the hand (to go back to the original example), we need to appeal to the intention, which comes out as the genuine cause on Menzies's view, and not to the causally sufficient neural realizer of the intention. The second option is that causal explanation tracks the causally sufficient physical antecedents and not the genuine cause. On this option the explanatory event would be the causally sufficient neural realizer.

In the mental causation case, the first option seems plainly to lead to a breach in the Completeness of Physics. The whole point of Menzies's discussion of mental causation is that the genuine causes of physical actions are intentions (or some other mental event). As such these genuine causes are not identical to their neural realizers – they cannot be identical to their neural realizers, because of their different modal characteristics (as revealed by the fact that different counterfactuals are true of the mental events and of their neural realizers). If the genuine cause of a physical action is neurally realized, but not identical to its neural realizer, then it follows that the genuine cause of a physical action cannot be identical to any physical antecedent of the physical action. Hence (as-

suming that the causal explanation tracks the genuine cause, not the causally sufficient physical antecedents) we will need to go outside the physical antecedents of the physical action in order to explain it causally – which contravenes the Completeness of Physics, as standardly defined.

Suppose then that causal explanation tracks the causally sufficient physical antecedents, not the genuine cause. If this is the case then, in the mental causation cases that we are considering, it is the neural realizers of the relevant mental events that do the explanatory work. This would certainly preserve the Completeness of Physics. But there is a heavy price to pay. The price is that we lose the connection between causal explanation and explanatory counterfactuals. Let us explain. It is widely, perhaps universally, held to be a necessary condition upon event P causally explaining event E that the following counterfactual hold: ~P $\Box\rightarrow$ ~E. We can term this the explanatory counterfactual. If, as we are assuming, causal explanation tracks causally sufficient physical antecedents, then in the mental causation cases we are considering the explanatory counterfactual will in effect say that, in the absence of the neural realizer, the relevant action would not have taken place. The whole point of Menzies's argument, however, is that this counterfactual does not hold. According to Menzies, the nearest worlds where the neural realizer does not occur are worlds where the same mental event is realized by some different neural realizer and hence causes the same physical action. So we have ~P and E.

We conclude, therefore, that the second of the three strategies open to Menzies is no more successful than the first, given the close connection between Causal Closure and the Completeness of Physics. Reformulating causal closure in terms of causally sufficient physical antecedents (as opposed to genuine causes) offers a choice between two highly unpalatable options. One option abandons the Completeness of Physics and accepts that an important class of physical events cannot be explained in terms of their physical antecedents. The second option preserves the Completeness of Physics, but at the price of abandoning the connection between causal explanation and explanatory counterfactuals.

4. Extending the Compatibility Result

The final option available to Menzies, if he wishes to retain closure while affirming the reality of mental causation, is to deny the rarity of the Compatibility Result. As a reminder, the Compatibility Result is the claim that if M causes B, then M's physical realizer N causes B if and only if the causal relation

between M and B is realization-sensitive. And the relation between M and B is realization-sensitive if the occurrence of the latter depends on M's having a specific realizer.

In these cases, the difference-making counterfactuals relating to both the realized mental event, M, and the realizing physical event, N, are satisfied. The satisfaction of the former is an affirmation of the causal efficacy of the mental, and the satisfaction of the latter secures conformity with the principle of closure. No causal exclusion is involved in either direction. In all other cases (those that do not correspond to the Compatibility Result), either the mental event is causally excluded by its physical realizer – in which case the mental is epiphenomenal – or the physical realizer is excluded by the mental event it realizes – in which case the principle of closure is violated. If we are to avoid both these options, we must insist that every purported case of mental causation is realization-sensitive.

However, this suggestion does not sit well with Menzies's original project – defending non-reductive physicalism against the epiphenomenalist consequences of the original exclusion principle. As was mentioned above, and as Menzies recognizes, the conditions for realization-sensitivity will obtain most readily when the mental event and its physical realizer are identical. That is, we will expect realization-sensitivity, and hence the simultaneous negation of epiphenomenalism and satisfaction of closure, when non-reductive physicalism is false. It is for this reason that Menzies insists that realization-sensitive causal relations should be the exception rather than the rule. Yet, given the Compatibility Result, realization-sensitivity is a condition on avoiding the violation of closure in all cases of genuine mental causation.

However, there is a further question that must be addressed. Although identity is a natural explanation of the realization-sensitivity of the causal relation between M and B, perhaps there is an alternative way of thinking about realization-sensitivity that is compatible with non-reductive physicalism.[7] If so then this leaves open the possibility of affirming mental causation without violating closure while still retaining the original motivation of non-reductive physicalism.

There certainly seems to be room in logical space for the non-identity of a mental event and its physical realizer even when the relation between that mental event and its physical effect is realization-sensitive. Suppose thatM is *dis-*

[7] Menzies too, does «not rule out the possibility of other explanations of the existence of realization-sensitive causal relations» (p. 80). However, he does not indicate what such an alternative might look like.

tinct from its physical realizer N. Then, obviously, there exists a possible world, say W*, in which M occurs and N does not occur. In possible world W* mental event M has a different neural realizer, say N*. Assuming that W* is a world where the functional role of M closely tracks its functional role in this world, then we can say that N* causes B in W*. All this is consistent, on Menzies's theory, with M being realization-sensitive, because realization-sensitivity simply requires that B fails to hold in all those M-worlds that are closest ~N-worlds. Consistency is secured precisely when W* is not one of the M-worlds that is among the closest ~N-worlds.

When we step back from the formalism, however, and consider what this actually amounts to, it is very unclear that it will provide much comfort to the non-reductive physicalist. Consider the modal profile of the scenario just sketched out. In this world and all nearby worlds M is realized by N and B holds. When we reach the limit of the N-worlds we are still in the sphere of the M-worlds, but B no longer holds. In other words, when M is realized differently from how it is realized in this world, then B no longer holds. Yet, as we continue to travel through the space of possible M-worlds that are also ~N-worlds, moving ever further away from the actual world, we eventually arrive at worlds where B starts to hold again. This all sounds bizarre.

Connected to this is a concern to do with the individuation of mental states. Suppose that mental states are individuated by causal role, as many philosophers believe. Then one would expect, at a minimum, that the causal profile of the mental state will be in some sense included in the causal profile of the realizing neural state. One way to put this might be: The set of causal powers of the mental state will be a subset of the set of causal powers of the realizing neural state.[8]And so, in any case of multiple realizability, there will be multiple realizing neural states with causal profiles including the causal profile of the realized mental state. One would expect this to hold both within and across worlds. But this cannot occur in the realization-sensitivity cases discussed earlier. The nearest ~N-worlds that are M-worlds are all worlds where B does not take place. So the M-relevant parts of the causal profile of M's neural realizer N* in the ~N worlds are different from M-relevant parts of the causal profile of N. In what sense, then, can N and N* both be realizers of M? In fact, one could press this further and ask the question: In what sense is M the same mental state when its causal profile differs with respect to B?

[8] Compare Shoemaker (2007).

In any event, the envisaged strategy does not seem to help at all with the intuitions and arguments driving non-reductive physicalism. Let us go back to the original formulation.

> *Non-reductive physicalism* – Mental events are realized by *distinct* physical events

What gives the thesis that mental events are realized by distinct physical events its force is the idea of multiple realizability. The distinctness of mental events and their physical realizers is underwritten by the fact that a single mental event could be realized by different physical events. Historically it was arguments from multiple realizability that first put non-reductive physicalism on the map as a potential alternative to versions of the type-identity theory (see, e.g., Putnam, 1967).

Different types of multiple realizability have been formulated and extensively discussed. For present purposes the key point is that multiple realizability has typically been proposed as an intraworld phenomenon. The example that Hilary Putnam originally gave in 'Psychological predicates' (Putnam, 1967) was pain. Pain, he claimed, is differently realized across the animal kingdom. Others have moved beyond the idea of species-relative multiple realizability and explored the idea that mental states may be differently realized in different individuals within a given species – or even differently realized at different times within the same individual. Consider the following passage from Terence Horgan.

> Multiple realizability might well begin at home. For all we now know (and I emphasize that we really do *not* now know), the intentional mental states we attribute to one another might turn out to be radically multiply realizable at the neurobiological level of description, *even in humans*; indeed, even in *individual* humans; indeed, even in an individual human *given the structure of his central nervous system at a single moment of his life*. (Horgan, 1993, p. 308; author's emphases)

Plainly, however, all of these types of intraworld multiple realizability are incompatible with the scenario that we have been discussing.

The dialectical situation with regard to the third horn of the trilemma, then, is the following. The causal efficacy of the mental can indeed be reconciled with the causal closure of the physical if all instances of mental causation have

the general pattern highlighted by the Compatibility Result. The Compatibility Result would require causally efficacious mental states to be realization-sensitive. The problem is that, even though the Compatibility Result does not entail the identity of mental event and realizer, it completely rules out any form of multiple realizability in this world – or even across nearby possible worlds. For that reason this strategy is, we submit, unavailable to anyone who wants to adopt a meaningful version of non-reductive physicalism.

5. Conclusion

We argued that a significant consequence of adopting Menzies's suggested modification of the exclusion principle is that genuine cases of mental causation (with, supposedly rare, exceptions) causally exclude their physical realizers. Such 'downward exclusion', we argued,poses a threat to a core principle of physicalism, the principle of the causal closure of the physical world.

To confront this threat, we proposed and evaluated three possible replies. The first reply, discussed in Section 2, is to bite the bullet by accepting the ubiquity of violations of physical closure. Closure is violated in all those cases in which the causal relation between some mental event and its physical effect is not realization-sensitive and thus involves downward exclusion. We argued that the ubiquitous violation of closure is too big of a bullet to bite. Given that Menzies's original suggestion aims to block the epiphenomenalist consequences of exclusion within a non-reductive physicalist metaphysics, the denial of such a central tenet of physicalism is especially problematic.

The second reply, presented in Section 3, starts off by noting that downward exclusion is only a threat to closure if we understand the latter as requiring the same kind of robust notion of causation as is excluded in cases of mental causation. A weaker formulation of the principle of physical closure – one that only requires that there exist some *causally sufficient* physical antecedent to any physical effect – is not similarly threatened. Indeed, because causation implies causal sufficiency, and causal sufficiency is transmitted across realization, genuine mental causation implies the causal sufficiency of its physical realizer.

Yet, such a weak formulation of the principle of closure, we argued, is unacceptable. The principle of closure is meant to support a related, explanatory, principle – the Completeness of Physics – which claims that the causal explanation of any physical event need only appeal to antecedent physical events. Yet the weak notion of causal sufficiency is inadequate for the purpose of causal

explanation, because, at the very least, causal explanation must preserve exactly those counterfactuals that in cases of downward exclusion are satisfied with respect to the mental cause but fail with respect to its (merely) causally sufficient physical realizer. Thus, we argued, such weakening of the principle of closure entails either a rejection of the Completeness of Physics or a severing of the connection between causal explanation and explanatory counterfactuals. Neither of these option is desirable.

The final suggestion, evaluated in Section 4, was to extend Menzies's, supposedly rare, Compatibility Result to all cases of mental causation, so that the causal relation between the mental event and its physical effect is at all times realization-sensitive. This suggestion would effectively eliminate all cases of downward exclusion and thereby the threat to causal closure. However, the demand that all instances of mental causation are realization-sensitive leads us to an unacceptable account of the relation between a mental event and its physical realizer. As Menzies admits, the most natural explanation of realization-sensitivity is *identity* between the mental event and its physical realizer. Yet if all mental causation rests on identity, we have abandoned non-reductive physicalism rather than resolved its potential epiphenomenalist consequences. An alternative explanation of realization-sensitivity that avoids identity nonetheless invokes an unacceptable notion of multiple realizability – one that does not allow for intra-world, or even close-world, multiple realizability.

We conclude that Menzies's modification of exclusion does not supply an adequate response to the epiphenomenalist threat to non-reductive physicalism. To the extent that it allows for genuine mental causation it does so at the cost either of denying a core principle of physicalism – closure – or of abandoning any acceptable notion of non-reductive physicalism.

REFERENCES

Block, N. (2003). Do Causal Powers Drain Away? *Philosophy and Phenomenological Research*, 67, 133–150.

Horgan, T. (1993). Nonreductive Materialism and the Explanatory Autonomy of Psychology. In S. Wagner, & R. Warner (Eds.), *Naturalism: A Critical Appraisal*. Notre Dame, IN: University of Notre Dame Press, 295–320.

Kim, J. (1989a). Mechanism, Purpose, and Explanatory Exclusion. *Philosophical Perspectives*, 3, 77–108.

Kim, J. (1989b). The Myth of Nonreductive Materialism. *Proceedings and Addresses of the American Philosophical Association*, 63, 31–47.

Kim, J. (2005). *Physicalism, or Something Near Enough*. Princeton: Princeton University Press.

Kim, J. (2009). Mental Causation. In B. McLaughlin, A. Beckermann, & S. Walter (Eds.), *The Oxford Handbook of Philosophy of Mind*. Oxford: Oxford University Press, 29–52.

List, C. and Menzies, P. (2009). Nonreductive Physicalism and the Limits of the Exclusion Principle. *Journal of Philosophy*, 1006, 475–502.

List, C., & Spiekermann, K. (2013). Methodological Individualism and Holism in Political Science: A Reconciliation. *American Political Science Review*, 107(4), 629–643.

Menzies, P. (2013). Mental Causation in the Physical World. In S.C. Gibb, & R. Ingthorsson (Eds.), *Mental Causation and Ontology*. Oxford: Oxford University Press.

Papineau, D. (2009). The Causal Closure of the Physical and Naturalism. In B. McLaughlin, A. Beckermann, & S. Walter (Eds.), *The Oxford Handbook of Philosophy of Mind*. Oxford: Oxford University Press, 53–65.

Putnam, H. (1967). Psychological Predicates. In W.H. Capitan, & D.D. Merrill (Eds.), *Art, Mind, and Religion*. Pittsburgh: University of Pittsburgh Press, 37–48.

Shoemaker, S. (2007). *Physical Realization*. Oxford: Oxford University Press.

Weslake, B. (forthcoming). Difference Making, Closure and Exclusion. In. H. Beebee, C. Hitchcock, & H. Price. (Eds.), *Making a Difference: Essays in Honour of Peter Menzies*. Oxford: Oxford University Press.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Woodward, J. (2006). Sensitive and Insensitive Causation. *Philosophical Review*, 115, 1–50.

Yablo, S. (1992). Mental Causation. *Philosophical Review*, 101, 245–280.