CHAPTER 2

Mental Causation by Counterfactual Dependence

2.1 Introduction

Mental events cause physical events because they make a difference to whether or not these physical events occur. This is the idea that is elaborated in this chapter. We saw in the previous chapter that an event causes a later event if it makes a difference to the occurrence of that event. The main task of this chapter is to show that mental events do in fact make a difference to physical events (technically speaking, to show that physical events counterfactually depend on mental events). If non-reductive physicalism is true, showing this is straightforward. If dualism is true, it is less straightforward but still manageable. Dualists will have to assume not just a naturalistic version of their view, but also a special status of the laws that connect mental and physical properties. The strategy of the argument for the counterfactual dependence of physical events on mental events is similar in the non-reductive physicalist case and the dualist case. In both cases, the argument proceeds as follows: the instantiation of a mental property is equivalent, in a sense to be spelled out in more detail, to the instantiation of some physical realizer or base of that mental property. Whether or not a realizer or base is instantiated makes a difference to whether or not future physical events occur. It follows that the instantiation of the mental property makes a difference to whether or not those physical events occur.

Section 2.2 presents the argument for the non-reductive physicalist case. We shall see that the argument generalizes to virtually all properties that supervene on physical properties: virtually all of these properties can also be shown to make a difference to the physical future and hence to have physical effects. For some supervenient properties this result is an interesting corollary. For others it looks more problematic. Section 2.3 discusses one of the more problematic cases and suggests several responses. A recent

https://doi.org/10.1017/9781108762717.004 Published online by Cambridge University Press

argument by Lei Zhong (2011, 2012) also attempts to show that supervenient mental property-instances have physical effects by drawing on certain counterfactuals. Section 2.4 argues that the argument presented here is superior to Zhong's in several respects. The dualist case is discussed in Section 2.5, which argues that dualists can show the efficacy of the mental and thus solve the interaction problem if they adopt what I shall call super-nomological dualism, that is, a version of dualism that assigns a special modal status to the psychophysical laws. Section 2.6 discusses an objection according to which the account of mental causation presented here falls short of explaining genuine agency.

2.2 Non-Reductive Physicalism

If non-reductive physicalism is true, then many physical events counterfactually depend on mental events and, therefore, are caused by these mental events. This section presents a simple argument for that conclusion. Some authors have invoked counterfactuals in order to show that non-reductive physicalism allows the mind to have physical effects,[1] but in general they have not attempted to show why the relevant counterfactuals are true (see Kim 1998: 71). The argument presented here gives a rigorous derivation of those counterfactuals.

The argument employs some of the assumptions that were defended in the previous chapter. It uses the strong Kimian account of events, according to which events are constituted by an object, a property, and a time and according to which actual and possible events are identical just in case they are constituted by the same object, property, and time. (For simplicity I will sometimes suppress reference to the object and the time in question and simply refer to events by talking about the instance of the property.) The argument uses Lewis's truth-conditions for counterfactuals and the logic that results from them. Recall that, according to the truth-conditions, a counterfactual is non-vacuously true just in case there is a world where both the antecedent and the consequent are true that is closer (that is, more similar overall) to the actual world than any worlds where the antecedent is true while the consequent is false; if there is no world where the antecedent is true, the counterfactual is vacuously true. As an account of the relation of overall similarity, the simple asymmetry-by-fiat approach will suffice. According to this approach, the closest antecedent-worlds of a given

[1] For instance, Baker (1993), Lepore and Loewer (1987), Keil (2001), Loewer (2007), and List and Menzies (2009).

counterfactual whose antecedent is actually false are exactly like the actual world until just before the time that the antecedent talks about; then the truth of the antecedent is brought about with minimal difference to the actual world; then things evolve lawfully again. (The asymmetry-by-fiat approach will suffice at least while we are dealing with non-reductive physicalism. For the dualist case that will be discussed in Section 2.5, the more elaborate miracles approach will prove more useful.) The argument uses our principle about causation according to which an event *c* causes a later event *e* if *e* counterfactually depends on *c*, that is, if *e* would not have occurred had *c* not occurred.

The argument draws on a consequence of the definition of strong supervenience, namely that the instantiation of a supervening property is strictly equivalent to the instantiation of some or other subvening property. Recall the definition from Section 1.2: a set of properties **A** *strongly supervenes* on a set of properties **B** if and only if, necessarily, if anything instantiates some property *F* in **A** at a given time, then there is a property *G* in **B** such that that thing instantiates *G* at that time, and, necessarily, everything that instantiates *G* at a given time also instantiates *F* at that time. Put less formally, **A**-properties strongly supervene on **B**-properties just in case any instantiation of an **A**-property has to be accompanied by an instantiation of some **B**-property, which in turn necessitates that the **A**-property is instantiated whenever it is itself instantiated. We have already used the example of dot-matrix pictures and their symmetry properties. Those symmetry properties strongly supervene on the arrangement of the dots in the picture's matrix. That is, any symmetry property of a dot-matrix picture has to be accompanied by the picture's instantiating some arrangement of dots or other, and any such arrangement that can underlie the symmetry of a picture necessitates its symmetry.

Now if a set of properties **A** strongly supervenes on a set of properties **B**, then the following is true: for each **A**-property *F* there is a subset of the **B**-properties – call this subset the *realizers* of *F* – such that, first, the instantiation of *F* necessitates the instantiation of a realizer of *F* and, second, the instantiation of a realizer of *F* necessitates the instantiation of *F* (at the same time and by the same object).[2] Take, for instance, the property of being point-symmetric for 3×3 dot-matrix pictures. By the supervenience of symmetry properties on dot arrangements, any point-symmetric

[2] For a similar result, see Kim 1984. Sometimes realization is taken to be a notion that is different from the one defined here. Advocates of such a notion of realization can simply substitute another term for what I have called a realizer. For a recent discussion of various notions of realization in the context of mental causation, see Walter 2010.

picture (actual or merely possible) has some dot arrangement that underlies this symmetry. Take all the possible dot arrangements that can underlie point-symmetry: ∴, ⋱, ⋮ ⋮, etc. We have already established that, by the supervenience of symmetry properties on dot arrangements, any point-symmetric picture has to have one of the arrangements ∴, ⋱, ⋮ ⋮, etc. It also follows from the supervenience – more precisely, from the second 'necessarily' in the definition – that any picture (actual or merely possible) that has one of the arrangements ∴, ⋱, ⋮ ⋮, etc. is point-symmetric. Thus, the set {∴, ⋱, ⋮ ⋮, etc.} is the set of realizers of point-symmetry for 3×3 dot-matrix pictures.

Let us return to the general case and expand the notation. If we are dealing with supervenient **A**-properties, let $\mathbf{P}_F$ be the set of realizers for each **A**-property *F* ('**P**' for 'physical', as we shall be dealing exclusively with physical realizers). For a set of properties **S**, let ∪**S** be the proposition that some member of **S** is instantiated. Let a roman capital letter stand for the proposition that the property referred to by the corresponding italicized capital letter is instantiated. Then we can formulate the consequence of **A**'s strong supervenience on **B** as follows: for each property *F* in **A**, there is a set of realizers $\mathbf{P}_F$ (where $\mathbf{P}_F$ is a subset of **B**) such that

(i) necessarily, if *F* is instantiated, then a realizer of *F* is instantiated ($\Box[F \supset \cup\mathbf{P}_F]$); and

(ii) necessarily, if a realizer of *F* is instantiated, then *F* is instantiated ($\Box[\cup\mathbf{P}_F \supset F]$).

We can express the consequence of **A**'s supervenience on **B** more concisely by turning (i) and (ii) into a strict biconditional: for each property *F* in **A**, there is a set of realizers $\mathbf{P}_F$ (where $\mathbf{P}_F$ is a subset of **B**) such that

(iii) necessarily, *F* is instantiated if and only if a realizer of *F* is instantiated. ($\Box[F \equiv \cup\mathbf{P}_F]$)

Applied to the strong supervenience of mental properties on physical properties, (iii) says that, necessarily, a given mental property is instantiated if and only if one of its realizers is instantiated. For instance, that someone is in pain is strictly equivalent to her instantiating a realizer of pain. Thus, that someone is in pain is strictly equivalent to her having firing c-fibres or having firing x-fibres or having an active semiconductor network of a certain kind in her head, etc. The strict equivalence of the instantiation of a mental property with the instantiation of one of its realizers is an important ingredient of the argument for mental causation under non-reductive physicalism, which we can now state.

According to non-reductive physicalism, mental properties strongly supervene on physical properties. We just saw that the instantiation of a property that strongly supervenes is strictly equivalent to the instantiation of one of its realizers. Let *M* be a specific mental property. Given that mental properties strongly supervene on physical properties, we have:

(1) Necessarily, *M* is instantiated if and only if a realizer of *M* is instantiated. ($\Box[\mathrm{M} \equiv \cup\mathbf{P}_M]$)

Unless *M* is instantiated at the last moment of history, some physical properties are instantiated later than *M*. Plausibly, some of them would not have been instantiated if *M*'s actual realizer had not been instantiated. Even more plausibly, some of them would not have been instantiated if none of *M*'s realizers had been instantiated. The asymmetry-by-fiat approach says so too. In the closest worlds where no realizer of *M* is instantiated, things are exactly as they actually are until just before the time at which *M*'s actual realizer is actually instantiated; then the non-occurrence of any realizer of *M* is brought about with minimal difference to the actual world; then things evolve lawfully again. It is hard to see how the absence of any realizer of *M* could leave no physical trace whatsoever. Indeed, we should expect many later physical events that actually occur not to occur in the closest worlds where no realizer of *M* is instantiated. Let *P** be a corresponding physical property that is instantiated later than *M* and that would not have been instantiated if none of *M*'s realizers had been instantiated:

(2) If none of *M*'s realizers had been instantiated, then *P** would not have been instantiated. ($\sim\cup\mathbf{P}_M \mathrel{\Box\!\!\rightarrow} \sim\mathrm{P}^*$)

We saw in Section 1.4 that Lewis's truth-conditions for counterfactuals allow us to replace the antecedent of a counterfactual with a strictly equivalent proposition. Thus, from (1) and (2) it follows logically that the *P**-instance counterfactually depends on the *M*-instance:

(3) If *M* had not been instantiated, then *P** would not have been instantiated. ($\sim\mathrm{M} \mathrel{\Box\!\!\rightarrow} \sim\mathrm{P}^*$)

We saw that counterfactual dependence is sufficient for causation that is forward in time. Applied to our case, this yields:

(4) If *P** is instantiated later than *M*, and *P** would not have been instantiated if *M* had not been instantiated, then the instance of *M* causes the instance of *P**.

https://doi.org/10.1017/9781108762717.004 Published online by Cambridge University Press

We have assumed that

(5) P^* is instantiated later than M.

From (3), (4), and (5) it follows logically that

(6) The instance of M causes the instance of P^*.

It follows, in other words, that there is causation of physical events by mental events.[3]

As it stands, the argument merely makes an existence claim, namely that there is some physical effect or other of a given mental property. We can also run the argument with reference to a specific physical event. I have a headache and reach for an aspirin. Having a headache is strictly equivalent to instantiating one of the realizers of having a headache. If I had instantiated none of these realizers, my hand would not have moved towards the aspirin. It follows that my hand's moving towards the aspirin counterfactually depends on my headache. Given our sufficient condition for causation,[4] it follows that my headache causes my hand's moving towards the aspirin.[5]

This is not to say, of course, that the argument can show an arbitrary physical event to be caused by a given mental event. And some physical events that have a good claim to be caused by a given mental event may not counterfactually depend on that mental event. (Thus, counterfactual dependence fails to be necessary for mental causation, just as it fails to be necessary for causation in general.) Perhaps a hospital patient has a headache and takes an aspirin, but if she had not had the headache, an overzealous nurse would have moved her hand towards the aspirin anyway. Cases like that of the hospital patient are the exception rather than the rule,

[3] If causation itself is non-hyperintensional (that is, if causal claims allow the substitution *salva veritate* of events whose occurrence is strictly equivalent), one could formulate an even easier argument for the causal efficacy of the M-instance. Assuming that the instance of the disjunctive property that some member of $\mathbf{P}_M$ is instantiated causes the instance of P^*, it would follow by the strict equivalence of M and $\cup\mathbf{P}_M$ that the instance of M causes the instance of P^*. The assumption that the instance of the disjunctive property is a cause is not without problems, however; see Sections 2.4 and 4.4 for further discussion.

[4] If I continue to have a headache after I have started reaching, let 'my headache' refer to the earlier temporal part of the continuing headache.

[5] We can also run the argument for (occurrent) propositional attitudes. If externalism about mental content is true, the realizers of those attitudes are at least partly extrinsic, but this does not threaten the efficacy of the attitudes, for the argument does not require the realizers themselves to be causes (see Sections 2.4 and 4.4). It seems to me that the account of mental causation presented here by itself neither solves nor exacerbates the problem of the efficacy of content. For discussion of that problem in the context of counterfactual accounts of causation, see Yablo 1997.

however, and in a wide range of cases the argument can show specific mental events to have specific physical effects.

The argument, both in its general and in its specific variety, assumes non-reductive physicalism about mental properties, but uses only the strong supervenience of mental properties on physical properties that non-reductive physicalism claims and no other specific assumptions about mental properties. Thus, the argument easily generalizes. Indeed, it can be used to show that virtually any instance of a property that strongly supervenes on physical properties has physical effects. For any such property *F*, it seems, we can find a physical property *P** that is instantiated later than *F* and that would not have been instantiated if none of *F*'s realizers had been instantiated. It follows from the argument that the instance of *F* causes the instance of *P**. Thus, it follows that there is downward causation of physical property-instances by virtually any supervenient property-instance.

Before assessing this result, we need a clarification. It does *not* follow from the argument that the instances of any property that is necessitated by a property with certain physical effects inherit all those physical effects. Suppose that an instance of a physical property *P** counterfactually depends on, and hence is caused by, an earlier instance of property *F*, which in turn necessitates the instantiation of property *H*. These suppositions do not entail that the instance of *P** counterfactually depends on, and hence is caused by, the instance of *H*, for the inference from $\sim F \mathrel{\Box\!\!\rightarrow} \sim P^*$ and $\Box[F \supset H]$ (contrapositively, $\Box[\sim H \supset \sim F]$) to $\sim H \mathrel{\Box\!\!\rightarrow} \sim P^*$ is invalid (see Section 1.4 and Lewis 1973b: 32). Thus, we do not get the result that every higher-level property-instance takes on all the effects of any lower-level property-instance that necessitates it. But of course it is consistent with the argument that sometimes higher-level property-instances do take on such effects.

Higher-level causes are not in general objectionable. Assume, as many do, that moral and aesthetic properties strongly supervene on physical properties.[6] Then our argument yields that they have some physical effects, for the absence of all physical realizers of a moral or aesthetic property would have made a difference to the physical future. Sometimes the argument can even be employed to show that they have certain specific effects. By the supervenience of aesthetic properties on physical properties,

[6] Even moral particularists like Dancy (1993) can accept the strong supervenience of moral properties on physical properties and the corresponding corollaries of forms (i)–(iii), for the realizers of moral properties are likely to be so complex that the supervenience claim does not yield any action-guiding principles.

beauty has certain physical realizers. If Helen of Troy had not instantiated any of those realizers while at Sparta, the arrowhead would not have moved towards Achilles' heel some nine years later. Hence the arrowhead's movement counterfactually depends on, and is caused by, the instance of beauty.

Cases like this are interesting corollaries of the argument for downward causation rather than problems for it. At least, I think so. The more cautious may simply restrict our principle about causation so that instances of moral and aesthetic properties are no longer allowed. As we saw in Section 1.5, some restrictions to rule out properties whose instances are generally ill-suited to enter into causal relations – restrictions to properties that are sufficiently intrinsic and temporally intrinsic, for instance – need to be imposed anyway, so this manoeuvre would not be *ad hoc* (or at least no more *ad hoc* than the original restrictions). Besides, other accounts of causation have to do the same, so our argument faces no special difficulty.[7] Restricting the sufficient condition for causation is unlikely to pose a threat to the efficacy of mental property-instances, for it is a desideratum of common sense that they can be causes. That they cannot, after all, be causes should be the conclusion of an argument, not a premise.

The argument I have presented in this section shows that, given non-reductive physicalism, particular mental events have physical effects. One might have lingering doubts about the efficacy not of particular mental events, but of mental events *qua* mental. Such doubts can easily be dispelled, however. For we have assumed the strong Kimian account of events, according to which events are not merely constituted by a property, an object, and a time, but have these constituents essentially. Given the combination of the strong Kimian account of events and non-reductive physicalism about the mind, mental events (that is, events that are constituted, *inter alia*, by mental properties) are not identical to physical events (that is, events that are constituted, *inter alia*, by physical properties) because of the distinctness of mental and physical properties that non-reductive physicalism claims. Thus, mental events do not have physical effects *qua* physical. One might still be worried that they have physical effects *qua* nothing, but this possibility can be ruled out, too. For clearly it is the mental properties that constitute, *inter alia*, mental events that are relevant for their causal efficacy. Unlike Quinean events or tropes, these mental properties are general features of the mental events, not particulars.

[7] For instance, a view on which event *c* causes event *e* if the occurrence of *c* and the actual laws of nature entail the occurrence of *e* also needs to be restricted to properties that are sufficiently temporally intrinsic. Otherwise, properties such as the property of shattering-in-a-minute yield counterexamples. The relation between causation and nomological sufficiency will be discussed in Section 4.5.

The mental properties are causally relevant because they do the work in the counterfactual dependence that implies the causal relation: if a given mental property had not been instantiated, then the later physical property would not have been instantiated.

2.3 The Problem of Overlapping Realizers

The argument for mental causation under non-reductive physicalism from the previous section shows that instances of supervenient mental as well as non-mental properties have physical effects. We saw in the Helen of Troy example that it can also be used to show that a supervenient non-mental property-instance has a specific physical effect. Sometimes, however, the argument seems to ascribe the wrong effects to supervenient property-instances. In particular, we seem to get the result that sometimes a supervenient property-instance has an effect that really is due to the instance of a different supervenient property that shares realizers with the first supervenient property. This section discusses that problem and explores several responses to it.

The problem arises as follows. I hold an aluminium ladder against a power line and subsequently get electrocuted.[8] Being made of aluminium, the ladder is an electrical conductor. Conductivity supervenes on physical properties and can be realized in different ways. If the ladder had not instantiated any realizer of conductivity, I would not have been electrocuted. It follows from the argument for downward causation that the instance of conductivity causes my electrocution. So far, so good. But being made of aluminium, the ladder is also opaque. Opacity too supervenes on physical properties and can be realized in different ways. The realizers of opacity are closely related to the realizers of conductivity. Almost all conductors are opaque. Some conductors are transparent (see Ginley *et al.* 2010), but they are not used to make ladders. Thus, it seems that if the ladder had not instantiated any realizer of opacity, I would not have been electrocuted either. It follows from the argument for downward causation that the instance of opacity causes my electrocution. That, however, does not seem very plausible, at least at first sight.[9]

[8] I borrow this example from Menzies (1988), with slight modifications. Jackson and Pettit (1990) also use the example, albeit in a different context.

[9] If artefacts such as ladders have their origin essentially, as Kripke (1980) holds, the ladder could not have been made of a different material from the one it is actually made of. If that is the case, the problem can be reformulated by taking the relevant events to be constituted by (i) the spatial region that is occupied by the ladder, (ii) the property of containing a ladder that is made of such-and-such a material, and (iii) the time in question.

Let us formulate the argument for the implausible conclusion along the lines of the argument from the previous section by using the following abbreviations:

C: being an electrical conductor
O: being opaque
E: being electrocuted

(In the example the object that instantiates property *E* (that is, myself) is distinct from the object that instantiates properties *C* and *O* (the ladder). In the original argument, property *P** might or might not be instantiated by the same object as *M*.) By the supervenience of opacity, we have:

(1-O) Necessarily, opacity is instantiated if and only if a realizer of opacity is instantiated. ($\Box[O \equiv \cup\mathbf{P}_O]$)

The close relation between the opacity-realizers and the conductivity-realizers seems to give us:

(2-O) If no opacity-realizer had been instantiated, then I would not have been electrocuted. ($\sim\cup\mathbf{P}_O \mathrel{\Box\!\!\rightarrow} \sim E$)

From (1-O) and (2-O) it follows logically that

(3-O) If opacity had not been instantiated, then I would not have been electrocuted. ($\sim O \mathrel{\Box\!\!\rightarrow} \sim E$)

By the sufficiency of counterfactual dependence for (forward-in-time) causation, from (3-O) we get the implausible conclusion:

(4-O) The opacity-instance causes my electrocution.

In the following I shall discuss several responses to the argument for this conclusion. We shall see that it is possible to deny the conclusion, but that this denial comes at a price. Ultimately, the best response will turn out to be the acceptance of the conclusion, coupled with an explanation of why it seems implausible.

The first response follows a strategy analogous to the strategy for dealing with backtracking counterfactuals that was discussed in Section 1.5 and denies the counterfactual that expresses the counterfactual dependence of my electrocution on the opacity-instance, (3-O).[10] Since (3-O) follows logically from (1-O) and (2-O), denying (3-O) requires denying either (1-O) or (2-O). Statement (1-O) seems unassailable, so one has to deny

[10] Menzies (1988: 573) denies this counterfactual dependence, but does not give an argument against it.

https://doi.org/10.1017/9781108762717.004 Published online by Cambridge University Press

(2-O). One has to deny, that is, that I would not have been electrocuted if no opacity-realizer had been instantiated. To see what denying (2-O) amounts to, consider the following argument *for* (2-O):

(5-O) If no opacity-realizer had been instantiated, then no conductivity-realizer would have been instantiated. ($\sim\cup\mathbf{P}_O \mathrel{\square\!\!\rightarrow} \sim\cup\mathbf{P}_C$)

(6-O) If no opacity-realizer had been instantiated and no conductivity-realizer had been instantiated, then I would not have been electrocuted. ($\sim\cup\mathbf{P}_O$ & $\sim\cup\mathbf{P}_C \mathrel{\square\!\!\rightarrow} \sim$E)

(2-O) If no opacity-realizer had been instantiated, then I would not have been electrocuted. ($\sim\cup\mathbf{P}_O \mathrel{\square\!\!\rightarrow} \sim$E)

The argument from (5-O) and (6-O) to (2-O) has the form of the restricted transitivity inference, which is valid (see Section 1.4 and see Lewis 1973b: 35). Given the validity of the argument, denying (2-O) requires denying either (5-O) or (6-O). Statement (6-O) looks very plausible. If all conductivity-realizers had been absent, I certainly would not have been electrocuted. It would be strange if the additional absence of all opacity-realizers were to bring back my electrocution.[11]

So denying (3-O), that is, denying the counterfactual dependence of the electrocution on the opacity-instance, ultimately requires denying (5-O). Denying (5-O) comes at a price, however. It is natural to think that if the ladder had not instantiated any opacity-realizer, then it would have been made of some middle-of-the-road transparent material (glass or transparent plastic, say), which would not have been conductive. This natural thought must be given up if (5-O) is denied. Instead, worlds where the ladder is made of some exotic transparent conductive material[12] have to be taken to be just as close to the actual world as worlds where the ladder is made of some middle-of-the-road transparent non-conductive material.

The second response is to drop the strong Kimian account of events in favour of a conception that allows for more flexibility in the modal relation between the event and the property that is instantiated, like the weak Kimian account or the Lewisian account. It does not matter for our

[11] Which is not to say that (6-O) follows logically from $\sim\cup\mathbf{P}_C \mathrel{\square\!\!\rightarrow} \sim$E, for it does not, owing to the invalidity of antecedent-strengthening for counterfactuals (see Section 1.4 and Lewis 1973b: 31).

[12] These days, transparent conductors are not exotic *per se*. They are used in smartphone screens and solar panels, for example (see Ginley *et al.* 2010). But they are certainly exotic in the context of ladders.

purposes which of these two alternatives we accept. (As we saw in Section 1.3, they are very similar in any case.) In order to spell out the response, we merely need an account of events that allows for a more relaxed connection between events and properties than the strong Kimian account does. Given such an account, it seems promising, at least *prima facie*, to proceed as follows: let *o* be the event of the ladder's being opaque. Event *o* should essentially involve the instantiation of opacity by the ladder. Otherwise we would have to say that the ladder's being opaque could have occurred while the ladder was not opaque, which seems strange.[13] Thus, we have:

(9) Necessarily, if *o* occurs, then opacity is instantiated. ($\Box[\mathrm{Oc}(o) \supset \mathrm{O}]$)

(In this addition to the notation, 'Oc(*x*)' stands for the proposition that event *x* occurs.) By itself, (9) does not relax the connection between events and their constituent properties, because (9) is also true on the strong Kimian account, according to which events have their constituent properties essentially, too. The converse of (9) is the claim that, necessarily, if opacity is instantiated, then *o* occurs; equivalently, that, necessarily, if opacity is not instantiated, then *o* does not occur. The converse of (9) is true on the strong Kimian account as well (assuming, as we tacitly do, that we are holding the constituent object and time fixed). We can relax the connection between events and their constituent properties by assuming not the converse of (9), which is a strict conditional, but the following counterfactual, which is logically weaker:

(10) If *o* had not occurred, then opacity would not have been instantiated. ($\sim\mathrm{Oc}(o) \mathrel{\Box\!\!\rightarrow} \sim\mathrm{O}$)

Lastly, we should demand that it is not the case that if *o* had not occurred, then the ladder would not have been conductive:

(11) It is not the case that if *o* had not occurred, then conductivity would not have been instantiated. ($\sim[\sim\mathrm{Oc}(o) \mathrel{\Box\!\!\rightarrow} \sim\mathrm{C}]$)

Claim (11) allows us to deny that event *o* causes my electrocution: if *o* had not occurred, I might still have been electrocuted because the ladder might still have been conductive.

The trouble with this response is that it is at least as problematic as the previous response, which sought to deny the claim that my electrocution

[13] At least it sounds strange in our case. In general, properties that feature in the description of a weak Kimian or Lewisian event do not have to be essential to that event. See Lewis 1986b: 247–254 for discussion.

https://doi.org/10.1017/9781108762717.004 Published online by Cambridge University Press

counterfactually depends on the opacity-instance. By contraposition, claim (9) is equivalent to the claim that

(12) Necessarily, if opacity is not instantiated, then *o* does not occur. ($\Box[\sim O \supset \sim Oc(o)]$)

Since strict conditionals logically imply the corresponding counterfactual conditionals, from (12) we get:

(13) If opacity had not been instantiated, then *o* would not have occurred. ($\sim O \mathrel{\Box\!\!\rightarrow} \sim Oc(o)$)

Claims (10), (11), and (13) logically imply:[14]

(14) It is not the case that if opacity had not been instantiated, then conductivity would not have been instantiated. ($\sim[\sim O \mathrel{\Box\!\!\rightarrow} \sim C]$)

By our earlier assumption (1-O), the instantiation of opacity is strictly equivalent to the instantiation of a realizer of opacity. Similarly, the instantiation of conductivity is strictly equivalent to the instantiation of a realizer of conductivity:

(1-C) Necessarily, conductivity is instantiated if and only if a realizer of conductivity is instantiated. ($\Box[C \equiv \cup\mathbf{P}_C]$)

Given (1-O) and (1-C), (14) is equivalent to:

(15) It is not the case that if no opacity-realizer had been instantiated, then no conductivity-realizer would have been instantiated. ($\sim[\sim\cup\mathbf{P}_O \mathrel{\Box\!\!\rightarrow} \sim\cup\mathbf{P}_C]$)[15]

Claim (15) is the negation of claim (5-O). We saw earlier that denying (5-O) is problematic because it requires giving up the natural thought that the ladder would have been made of some middle-of-the-road transparent material if it had not instantiated any opacity-realizer. Thus, the response that adopts a more coarse-grained conception of events instead of the strong Kimian account is at least as costly as the first response.

[14] The inference has the form of an inference from $\phi \mathrel{\Box\!\!\rightarrow} \chi$, $\chi \mathrel{\Box\!\!\rightarrow} \phi$, and $\sim[\chi \mathrel{\Box\!\!\rightarrow} \psi]$ to $\sim[\phi \mathrel{\Box\!\!\rightarrow} \psi]$, which is valid if and only if the inference from $\phi \mathrel{\Box\!\!\rightarrow} \chi$, $\chi \mathrel{\Box\!\!\rightarrow} \phi$, and $\phi \mathrel{\Box\!\!\rightarrow} \psi$ to $\chi \mathrel{\Box\!\!\rightarrow} \psi$ is, which we saw to be valid in Section 1.4; see also Lewis 1973b: 33.

[15] We saw in Section 1.4 that we may substitute necessarily equivalent antecedents in counterfactuals. The substitution of necessarily equivalent consequents is likewise allowed – if two propositions are true at exactly the same worlds, then they are either both true or both false at the closest worlds where a given antecedent is true.

The third response is to refine the sufficient condition for causation by taking into account counterfactuals with more complex antecedents. If opacity had not been instantiated, then I would not have been electrocuted. But if opacity had not been instantiated *while conductivity had still been instantiated*, then I would still have been electrocuted. On the other hand, if conductivity had not been instantiated *while opacity had still been instantiated*, then I would not have been electrocuted. More generally, the idea is that one event causes another if the first event makes a difference to the occurrence of the second event if we hold the occurrence of certain other events fixed.

What other events should we hold fixed? This is not an easy question to answer. For virtually any pair of events that are related by counterfactual dependence, we can find other events that actually occur and for which holding them fixed makes no difference to the occurrence or non-occurrence of the dependent event. Take the example of my throwing a dart at a balloon. I throw the dart; the balloon bursts. If I had not thrown the dart, then the balloon would not have burst. If I had not thrown the dart and there had been just as many grains of sand on Mars as there actually are, then the balloon would not have burst either. On the other hand, for virtually any pair of events that are related by counterfactual dependence, we can also find other events that actually occur and for which holding them fixed does make a difference to the occurrence or non-occurrence of the dependent event. For instance, if I had not thrown the dart and the dart had been on its actual trajectory a second later (somehow materializing there despite not having been thrown), then the balloon would still have burst. Why should we hold fixed the ladder's being conductive when assessing whether the ladder's being opaque causes the electrocution, but not hold fixed the dart's being on its later trajectory when assessing whether my throw causes the balloon's bursting? Intuitively, the difference is that the dart's being on its later trajectory is on the causal path from my throw to the bursting, while the ladder's being conductive is not on a causal path – if such there be – from the ladder's being opaque to the electrocution. (Nor, for that matter, is the sand event on Mars on a causal path from my throw to the balloon's bursting.) Only off-path events, it seems, should be held fixed.

As it stands, this suggestion is rather vague. It also smacks of circularity. How can we identify causal paths without making prior assumptions about what causes what? So-called causal modelling theories of causation can be used to make the suggestion more precise and to avoid the apparent circularity. I will elaborate in Section 3.5, but one difficulty with the solution can be outlined here without going into details. Causal modelling theories of causation use causal

models (unsurprisingly), which consist, *inter alia*, of a set of variables that represent the occurrence of events. In order to spell out the idea that counterfactual dependence is sufficient for causation if the dependence persists when all off-path events are held fixed, it is not enough to demand that there be *some* causal model where the dependence thus persists. If this were enough, we could take a simple model that merely contained variables for the putative cause and the putative effect and for no other events. In that simple model, it would be trivially true that the counterfactual dependence between the putative cause and the putative effect persists if all off-path events are held fixed, for there are no off-path events in the model. Instead of merely stating an existential condition for models of a certain kind, it seems that we should demand that the counterfactual dependence persists in an appropriate model. This requires spelling out what an appropriate model is, however. As we shall see, that is no easy task.

If the responses discussed so far all seem unsatisfactory, we have two more options, which are more radical. The fourth response is to deny that counterfactual dependence is sufficient for causation without attempting to replace it with a modified sufficient condition (such as the sufficient condition in terms of holding off-path events fixed). The fifth response is to maintain the original sufficient condition and accept that the opacity-instance causes the electrocution. Denying that counterfactual dependence is sufficient for causation is simple. But so is the idea that what makes a difference is a cause. It seems premature to give that idea up unless all alternatives turn out to be untenable. The other radical option, namely accepting that the opacity-instance causes the electrocution, might initially seem like excessive bullet-biting. But a closer look reveals it to be not so unattractive. If we choose that option, we can hold on to our original simple and elegant sufficient condition for causation. We shall have to accept the result that the opacity-instance causes the electrocution, but we can try to explain away the implausibility of this result. It is because of the intimate relation between the realizers of conductivity and the realizers of opacity that the electrocution counterfactually depends on the opacity-instance. This intimate relation may well eventuate in the opacity-instance's causing the electrocution. We might still hesitate to call the opacity-instance a cause of the electrocution, but we hesitate because the opacity-instance is a cause that has little explanatory relevance in our context, not because it is not a cause at all.[16] (Recall the example from Section 1.5 of my bumping into Albert as an explanatorily irrelevant cause of Berta's death.)

[16] It might seem promising to apply Swanson's (2010) account of the context-sensitivity of causal talk to our case. Unfortunately, there are some *prima facie* difficulties with this application. Swanson

https://doi.org/10.1017/9781108762717.004 Published online by Cambridge University Press

This defence of the fifth response, which holds on to our principle about causation but denies that the opacity-instance is a cause that is explanatorily relevant in our context, does not threaten the status of mental causes. For mental events do typically count as explanatorily relevant. And the argument from the previous section showed that they can cause physical events. Thus, it is clearly appropriate to say that they cause those physical events. For instance, it is clearly appropriate to name my headache as a cause of my hand's moving towards the aspirin, because the headache is not merely a cause of my hand's moving (because my hand's moving counterfactually depends on it), but a cause that we would cite in an explanation of my hand's moving. That mental events are explanatorily relevant seems obvious (see Burge 1993). It can also be established through argument. As we shall see in Section 3.5, causal modelling theories allow us to formulate a criterion for explanatory relevance within the counterfactual approach to causation.

2.4 Comparison with Zhong's Argument

Lei Zhong has suggested an argument that is similar to the argument for mental causation under non-reductive physicalism that I have presented. He argues as follows.[17] Assume non-reductive physicalism. Assume further that an instance of a mental property M causes an instance of a mental property M^* that is realized by a physical property P^*. By the realization of M^* by P^*, that P^* is instantiated entails that M^* is instantiated ($\Box[P^* \supset M^*]$). Contrapositively, that M^* is not instantiated entails that P^* is not instantiated ($\Box[\sim M^* \supset \sim P^*]$). Thus, the P^*-instance counterfactually depends on whatever the M^*-instance counterfactually depends on, since

appeals to the principle that when ascribing causal responsibility for a given effect to a causal path, one should use good representatives of that path (2010: 225). One cannot use this principle to show that the conductivity-instance is a better representative of a path that contains both the conductivity-instance and the opacity-instance than the opacity-instance is, for both by Swanson's definition and by the causal modelling definition (which will be presented in more detail in Section 3.4) the two property-instances are on different paths. Perhaps it could be shown that the opacity-instance is a poor representative of a path that contains it but does not contain the conductivity-instance. But showing this would not be straightforward either, since one of Swanson's principal criteria for an event's being a *good* representative, the effect's counterfactually depending on the representative, does apply to the opacity-instance and the electrocution.

[17] See Zhong 2011: 141–143; 2012: 80–81. I follow the 2012 version of the argument here, which Zhong prefers (2012: 81 n. 7). Zhong uses his argument in the context of a strategy that is different from the one I pursue. Instead of arguing that non-reductive physicalism can accommodate mental causation, he uses his argument to strengthen the exclusion problem for non-reductive physicalism. Zhong's argument is also discussed in Pernu 2016.

$\sim X \mathrel{\Box\!\!\rightarrow} \sim P^*$ follows logically from $\sim X \mathrel{\Box\!\!\rightarrow} \sim M^*$ and $\Box[\sim M^* \supset \sim P^*]$.[18] Now if the *M*-instance causes the *M**-instance, then either

(i) the *M**-instance counterfactually depends on the *M*-instance; or
(ii) there is an intermediary, namely an instance of a mental property *M′* which is caused by the *M*-instance and on which the *M**-instance counterfactually depends.

In case (i), it follows that the *P**-instance counterfactually depends on the *M*-instance; hence the *M*-instance causes the *P**-instance. In case (ii), it follows that the *P**-instance counterfactually depends on the *M′*-instance; hence the *M′*-instance causes the *P**-instance; hence, by the transitivity of causation, the *M*-instance causes the *P**-instance. In sum, if some instances of mental properties cause instances of other mental properties, then they also cause instances of the realizers of these mental properties.

Zhong's conclusion is weaker than mine. He concludes that a mental property-instance causes the instance of the realizer of another mental property *if* the first mental property-instance causes the second mental property-instance. I conclude that some mental property-instances cause physical property-instances *tout court*. That some mental property-instances cause other mental property-instances is not very controversial, however,[19] so the fact that Zhong's conclusion is a conditional one while mine is not does not make for a substantial difference between our arguments.

Zhong's argument is more specific than mine. My argument can easily be generalized to supervenient properties besides mental properties as conceived of by non-reductive physicalism. All that is required for this generalization is that the absence of all realizers of the supervenient property in question would have made a difference to the physical future. In order to generalize Zhong's argument to other supervenient properties, we would first have to identify a future instance of another supervenient property that is caused by the instance of the original supervenient property. It might not always be straightforward to find such a future instance. We saw in the previous section that the ease with which my argument can be generalized is a mixed blessing. It might therefore be taken to be an advantage of Zhong's argument that it does not generalize so easily. His argument has a number of disadvantages, however.

[18] This valid inference should not be confused with the similar but invalid inference from $\sim\psi \mathrel{\Box\!\!\rightarrow} \sim\chi$ and $\Box[\psi \supset \phi]$ (contrapositively, $\Box[\sim\phi \supset \sim\psi]$) to $\sim\phi \mathrel{\Box\!\!\rightarrow} \sim\chi$; see Section 1.4.

[19] Proponents of the so-called autonomy approach such as Gibbons (2006) accept mental-to-mental causation while denying mental-to-physical causation.

Zhong's assumptions about causation are stronger than mine. He assumes that counterfactual dependence is sufficient for causation and that counterfactual dependence – or counterfactual dependence via a caused intermediary – is necessary for causation. I merely assume that counterfactual dependence is sufficient for causation. (Strictly speaking, I assume even less: that counterfactual dependence between property-instances is sufficient for causation that is forward in time. But Zhong could do so as well without jeopardizing the validity of his argument, so we are on a par here.) In spite of the worries we have discussed in previous sections, the sufficiency of counterfactual dependence for causation is very plausible. The necessity of counterfactual dependence – or counterfactual dependence via a caused intermediary – is not very plausible. Consider a case of late pre-emption, such as Billy and Suzy throwing rocks at a bottle. Billy's rock arrives at the bottle first and causes it to shatter. We saw in Section 1.4 that this is a case of causation without counterfactual dependence: it is not the case that the bottle would not have shattered had Billy not thrown, because in this case Suzy's rock would have shattered it. It is also a case of causation without counterfactual dependence on a caused intermediary. Whatever intermediate event we choose that is caused by Billy's throw, the shattering does not counterfactually depend on it. For instance, the event of Billy's rock being on its actual trajectory a split-second after the throw is caused by Billy's throw, but the shattering would still have occurred if that event had not occurred, because in that case, too, Suzy's rock would have shattered the bottle. Cases of overdetermination, such as deaths by firing squad, also yield counterexamples not only to the necessity of counterfactual dependence for causation, but also to the necessity of counterfactual dependence via a caused intermediary for causation.[20]

Another controversial assumption about causation that Zhong makes is that causation is transitive. He assumes, that is, that if a first event causes a second event and the second event causes a third event, then the first event causes the third event. The transitivity of causation is subject to various counterexamples. Here is one of them.[21] I throw a railway switch, diverting a train to a side track. The side track later rejoins the main track. Further down on the main track someone left a cart, which is run over by the train. My throwing the switch causes the train to be on the side track a moment later. The train's being on the side track at that moment causes it to run over the cart later on. But my throwing the switch does not seem to cause the train

[20] On a similar issue, see Lewis 1986d: 193–212.

[21] So-called switching cases like this one are due to McDermott (1995: 532). For further discussion of the transitivity of causation, see Paul 2000 and Paul and Hall 2013: 215–244.

to run over the cart. (The case is *not* also a counterexample to the sufficiency of counterfactual dependence for causation, for the running over of the cart does not counterfactually depend on my throwing the switch to start with: if I had not thrown the switch, the train would simply have stayed on the main track and would still have run over the cart later on.)

Perhaps Zhong's argument would still be valid if the controversial assumptions about causation were appropriately weakened. He could assume, for instance, that counterfactual dependence – or counterfactual dependence via a caused intermediary – is necessary for causation *in the absence of redundancy* and that causation is transitive *in standard cases*. The weakened assumptions would be less controversial. But controversy might arise over whether they can be applied in particular cases. If we can give them up completely, so much the better.

Zhong's argument is open to an objection to which my argument is immune. Jonas Christensen and Jesper Kallestrup (2012) object to Zhong's argument as follows. The necessitation of M^* by its realizer P^*, which is required to establish that the P^*-instance counterfactually depends on the M-instance or the M'-instance, holds only if P^* is a 'total realizer' of M^* (2012: 515). That is, P^* has to be a conjunctive property that includes various 'background properties' such as 'properties pertaining to pertinent causal laws of nature' besides its 'core realizer' properties, which are more narrowly circumscribed (2012: 514). The background properties, however, are not themselves 'causal properties' that could feature as causes or effects (2012: 515). Given that P^* includes those background properties, the claim that the instance of P^* is an effect becomes problematic. Moreover, the background properties are shared between P^* and the actual realizer of M. Thus, Christensen and Kallestrup claim, M and P^* are no longer sufficiently distinct to be causally related (2012: 516).

Whatever the success of Christensen and Kallestrup's objection to Zhong's argument, their objection does not touch mine.[22] Granted, I would have to restrict the sufficient condition for causation to causal properties if the objection were sound, for otherwise background properties would yield counterexamples. (As we have seen, we need to impose some restrictions along these lines in any case.) Granted, the realizers of M that featured in claims (1) and (2) from Section 2.2 would have to be read as total realizers if the objection were sound, for otherwise the instantiation of M would no longer be strictly equivalent to the instantiation of one of its realizers, as (1) claims.

[22] Zhong (2015) addresses the objection by Christensen and Kallestrup. We shall return to the issue of whether the instances of realizers can be causes in Section 4.4.

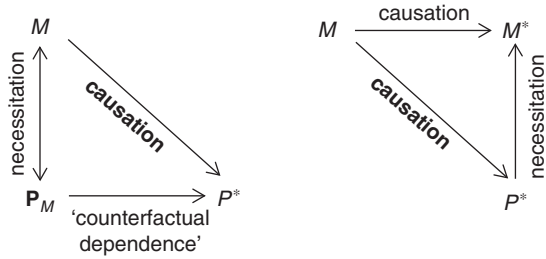

Figure 2.1. Zhong's argument (right) vs mine (left)

But these concessions would not threaten the causal relation between our *M*-instance and our *P**-instance. Zhong's argument is open to the Christensen–Kallestrup objection because there is a realizer (namely the realizer of *M**) whose instance is claimed to be an effect. In my argument no realizers need to have instances that are causes or effects. The set of *M*'s realizers (as represented by proposition $\cup\mathbf{P}_M$ in (1) and (2)) is merely a logical intermediary, not a causal one. And our property *P** can be as causal as one likes, since it need not realize anything.

What is the role of the actual realizer of *M* (call it *P*)? One might object that I cannot avoid treating at least the instance of *P* as a causal intermediary, since it follows from *M*'s being necessitated by *P* that *P* would not have been instantiated if *M* had not been instantiated, wherefore the *M*-instance causes the *P*-instance. I have to concede only the first half of this reasoning, however. It does follow that the *P*-instance counterfactually depends on the *M*-instance. But the sufficient condition for causation I have used remains silent on whether or not the *M*-instance causes the *P*-instance, because the two instances are simultaneous. Further, given that we restrict the sufficient condition for causation to causal properties, it would remain silent on whether or not the *P*-instance causes the *P**-instance should it turn out that our *P**-instance counterfactually depends on the *P*-instance while *P* is not a causal property. So it neither follows that the *P*-instance is an effect of the *M*-instance nor that the *P*-instance is a cause of the *P**-instance. Nonetheless it still follows that the *M*-instance is a cause of the *P**-instance.

Figure 2.1 summarizes the structure of Zhong's argument and mine. Zhong's argument proceeds from causation on the mental level and concludes that the realizer of the mental effect (more precisely: the instance of the realizer of the property of which the mental effect is an instance) has the same mental cause as the mental effect. My argument does not proceed from causation on the mental level. Nor does it proceed from causation on

https://doi.org/10.1017/9781108762717.004 Published online by Cambridge University Press

the physical level. Nor does it proceed from counterfactual dependence on the physical level, strictly speaking. In a loose sense, the P^*-instance counterfactually depends on the set of M's realizers, $\mathbf{P}_M$ (namely in the sense that P^* would not have been instantiated had no member of $\mathbf{P}_M$ been instantiated). Together with the strict equivalence of the instantiation of M with the instantiation of a member of $\mathbf{P}_M$, it follows that the P^*-instance counterfactually depends on the M-instance (in the strict sense) and hence is caused by it. On the level of realizers, Zhong's argument merely takes into account the actual realizer P^* of the mental effect M^*; hence the relation of necessitation is merely one-way, unlike the relation between M and the set of realizers $\mathbf{P}_M$.

2.5 Dualism

Of all positions about the nature of mind, dualism has been considered the one for which mental causation spells most trouble. The interaction problem is particularly severe for dualism. If the mental is neither identical to nor necessitated by the physical, how can it interact with the physical at all? The exclusion problem, too, is particularly severe for dualism. Even if the mental can interact with the physical in principle, how can it do so without making it the case that physical effects are caused twice over, like in a firing squad? This section deals with the interaction problem for dualism. (The exclusion problem will be discussed in Chapter 4.) It will turn out that dualists can solve the interaction problem provided they make certain assumptions about the status of the psychophysical relation.

I will not discuss varieties of dualism that do not even assume that the relation between mental and physical properties is a matter of natural law. Those varieties cannot avail themselves of the solution I am going to suggest, and I doubt that there is an alternative solution for them. Let us assume, then, that what we called 'naturalistic dualism' in Section 1.2 is true:

Naturalistic dualism: Each mental property is distinct from all physical properties. No subset of mental properties strongly supervenes on physical properties, but mental properties nomologically supervene on physical properties.

Recall that the notion of nomological supervenience was in turn defined as follows:

Nomological supervenience: A set of properties **A** *nomologically supervenes* on a set of properties **B** if and only if it is nomologically necessary that if

https://doi.org/10.1017/9781108762717.004 Published online by Cambridge University Press

anything instantiates some property *F* in **A** at a given time, then there is a property *G* in **B** such that that thing instantiates *G* at that time, and it is nomologically necessary that everything that instantiates *G* at a given time also instantiates *F* at that time.

Applied to the case of mental and physical properties, we get the following claim of nomological supervenience:

Nomological psychophysical supervenience: It is nomologically necessary that if anything instantiates some mental property at a given time, then there is a physical property such that that thing instantiates the physical property at that time, and it is nomologically necessary that everything that instantiates the physical property at a given time also instantiates the mental property at that time.

According to nomological psychophysical supervenience, it is a matter of nomological necessity that a mental property is accompanied by some physical property whenever it is instantiated, and it is also a matter of nomological necessity that the mental property is instantiated whenever one of the physical properties that can underlie its instantiation is instantiated.

As was the case with strong supervenience, nomological supervenience allows us to correlate each supervenient property with a disjunction of subvening properties, although, unlike in the case of strong supervenience, this correlation holds only with nomological necessity, not with metaphysical necessity. In the case of nomologically supervenient mental properties, for each mental property *M*, there is a set $\mathbf{P}_M$ of physical properties – call them the *bases* of *M* – such that

(i) it is nomologically necessary that if *M* is instantiated, then a base of *M* is instantiated; and
(ii) it is nomologically necessary that if a base of *M* is instantiated, then *M* is instantiated.[23]

(For simplicity, I am again leaving reference to times and to the things that instantiate the properties in question implicit.) We can turn (i) and (ii) into a biconditional that holds with nomological necessity:

[23] Alternatively, one could call the members of $\mathbf{P}_M$ *realizers*, as in the non-reductive physicalist case. But, since talk of realization invokes a relation of metaphysical necessitation, it seems preferable to use a different term.

(iii) It is nomologically necessary that *M* is instantiated if and only if a base of *M* is instantiated.

For example, according to nomological psychophysical supervenience, it is nomologically necessary that someone is in pain if and only if they have firing c-fibres or they have firing x-fibres or an active semiconductor network of a certain kind in their head etc. Thus, the properties of having firing c-fibres, of having firing x-fibres, of having an active semiconductor network of a certain kind in one's head, etc. are the bases of pain.

Let us assume that the nomological necessity of the fact that a mental property is instantiated just in case one of its bases is instantiated is due to a psychophysical law that has the status of a fundamental law of nature.[24] Thus, the psychophysical laws are nomologically necessary, which they should be, for nomological necessity is truth in all worlds where all the actual laws of nature hold (see Section 1.2), and in any worlds where all the actual laws of nature hold, *a fortiori* the actual psychophysical laws hold. The converse, that all the actual laws hold in any worlds where the actual psychophysical laws hold, does not follow. There might be worlds where the actual psychophysical laws hold, but some of the other actual laws of nature do not hold. If this is the case, then the actual psychophysical laws are in a sense 'more necessary' than the remaining actual laws of nature, while still being nomologically necessary. As we shall see, it is worth taking this possibility seriously.

The psychophysical laws cannot be metaphysically necessary as well as nomologically necessary, at least by dualists' lights, for if they were metaphysically necessary, mental properties would strongly supervene on physical properties, and dualism does not allow this.[25] Given the failure of mental properties strongly to supervene, we can no longer use claim (1) from Section 2.2, according to which the instantiation of a mental property is strictly equivalent to the instantiation of a realizer of that property, as a starting-point for an argument for mental causation. We can, however, use the weaker claim that the instantiation of a mental property is, as it were, counterfactually equivalent to the instantiation of a physical base of that property: if the mental property had not been instantiated, then none of its bases would have been instantiated, and if none of its bases had been

[24] On such laws, see Chalmers 1996: 127.

[25] If one included irreducible psychophysical laws in the subvening properties and also stipulated that the subvening properties include physical properties and perhaps physical laws, the result would be that mental properties strongly supervene. This result would not vindicate physicalism, however, for physicalism requires that the mental supervene on the physical alone; thus, physicalism cannot allow irreducible *psycho*physical laws in the subvening properties.

instantiated, then the mental property would not have been instantiated. As before, we can combine this with the claim that the absence of all physical bases would have made a difference to the physical future. Thus, we get the following argument (where M is a specific mental property and P^* is a physical property that is instantiated later than M and that would not have been instantiated if none of M's bases had been instantiated):

(16) If none of M's bases had been instantiated, then M would not have been instantiated. ($\sim\cup\mathbf{P}_M \mathrel{\Box\!\!\rightarrow} \sim$M)

(17) If M had not been instantiated, then none of M's bases would have been instantiated. ($\sim$M $\mathrel{\Box\!\!\rightarrow} \sim\cup\mathbf{P}_M$)

(18) If none of M's bases had been instantiated, then P^* would not have been instantiated. ($\sim\cup\mathbf{P}_M \mathrel{\Box\!\!\rightarrow} \sim$P*)

---

(19) If M had not been instantiated, then P^* would not have been instantiated. ($\sim$M $\mathrel{\Box\!\!\rightarrow} \sim$P*)

Claim (19) says that the P^*-instance counterfactually depends on the M-instance. Once this is established, we can continue as we did in the case of non-reductive physicalism. Applied to the case of M and P^*, our sufficient condition for causation yields:

(4) If P^* is instantiated later than M, and P^* would not have been instantiated if M had not been instantiated, then the instance of M causes the instance of P^*.

We have assumed that

(5) P^* is instantiated later than M.

From (4), (5), and (19) it follows logically that

(6) The instance of M causes the instance of P^*.

The argument from (16)–(18) to (19) is the most controversial part of this reasoning for the causal claim, (6). The validity of the argument, however, is beyond reproach. We saw in Section 1.4 that inferences of the form of the present inference are valid, in spite of the general failure of transitivity for counterfactuals. Given this failure, (19) does not follow from (17) and (18) alone, but (19) does follow if we add premise (16). Together with premise (16), (17) guarantees that the closest worlds where M is not instantiated coincide with the closest worlds where none of M's physical bases is

instantiated. Since by (18) the latter are worlds where P^* is not instantiated, so are the former.[26]

What about the premises, claims (16)–(18)? Let us consider them in reverse order. I take it that premise (18) – more precisely, the existence of a physical property P^* that is instantiated later than M and that makes (18) true – is as plausible as it was in the case of non-reductive physicalism. Taking away the actual physical base of a mental event and not replacing it by an alternative base certainly makes a difference to the physical future.

Premise (17), by contrast, looks problematic. Let us try to apply the asymmetry-by-fiat approach. According to this approach, the closest worlds where M is not instantiated are exactly like the actual world until just before the time at which M is actually instantiated. Then the instantiation of M is prevented with minimal difference to the actual world; then things evolve lawfully again. What does it mean to prevent the instantiation of M with minimal difference to the actual world? We could prevent the instantiation of M by not having any of M's physical bases instantiated. If this makes for a minimal difference to the actual world, (17) comes out true. Or we could prevent the instantiation of M by eliminating its instantiation while leaving everything as it is in the physical world; given dualism, this is a metaphysical possibility since the instantiation of a base of M does not metaphysically necessitate the instantiation of M. If the second option makes for a minimal difference to the actual world, (17) comes out false. It is not entirely clear which way of preventing M's instantiation is the way of minimal difference to the actual world, but one might suspect that it is the way that leaves the physical world as it is, for this way differs from actuality merely with respect to one event (namely the M-instance) and not with respect to two events (the M-instance and the instance of its actual physical base).

The problem looks even worse if we consider the verdict of the miracles approach to closeness or overall similarity. Recall Lewis's criteria for overall similarity to the actual world:

(1) It is of the first importance to avoid big, widespread, diverse violations of law.

(2) It is of the second importance to maximize the spatiotemporal region throughout which perfect match of particular fact prevails.

[26] The inference would still be valid if (17) were replaced with the corresponding 'might' conditional (see Lewis 1973c: 433).

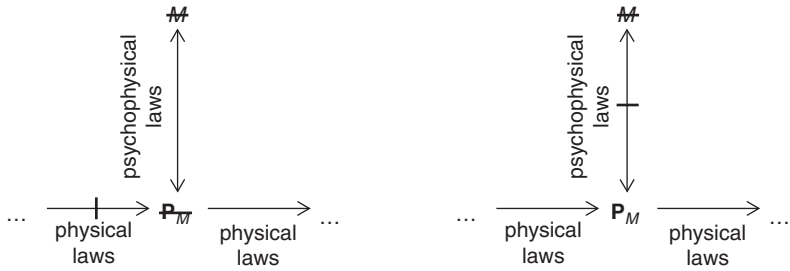

Figure 2.2. Type-1 worlds (left) vs type-2 worlds (right) as candidate antecedent-worlds for premise (17)

(3) It is of the third importance to avoid even small, localized, simple violations of law.

(4) It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly. (Lewis 1979: 472)

Compare the following two types of antecedent-worlds of (17) for closeness to the actual world (see Figure 2.2): worlds of type 1 match the actual world perfectly in particular fact until just before the time at which *M* is instantiated; then the physical laws are violated while the psychophysical laws are not, such that *M*'s failure to be instantiated implies the failure of any of its physical bases to be instantiated. Worlds of type 2 match the actual world perfectly in particular fact until just before the time at which *M* is instantiated too; then the psychophysical laws are violated while the physical laws are not, such that *M*'s actual physical base is still instantiated, but *M* is not. By Lewis's criteria, type-2 worlds are closer to the actual world than type-1 worlds. While worlds of the respective types are on a par as far as violations of law are concerned, there is vastly more match of particular fact to the actual world in the type-2 worlds, for type-2 worlds match the actual world perfectly at all times after *M*'s actual instantiation. Type-1 worlds, by contrast, cannot equal this match; owing to the failure of any of *M*'s physical bases to be instantiated, they lawfully evolve into a different future.[27] In type-2 worlds, the antecedent of (17) is true while its consequent is false; in type-1 worlds, both are true. Thus, if Lewis's similarity criteria apply, (17) is false.

Naturalistic dualists can avoid this result, however. They hold that the relation between mental events and physical events is contingent owing to

[27] See Loewer 2001a: 51–52 for an argument along these lines.

https://doi.org/10.1017/9781108762717.004 Published online by Cambridge University Press

the failure of mental properties strongly to supervene on physical properties. Specifically, they hold that it is contingent that the instantiation of a physical base of *M* implies the instantiation of *M*. The psychophysical laws that entail such contingent implications must be contingent as well. But nothing forces dualists to accept that psychophysical laws are modally on a par with ordinary laws of nature, such as the laws of physics. They are within their rights to claim that psychophysical laws could not have failed so easily as the other laws. They can claim, in other words, that worlds where the psychophysical laws are violated are further from actuality than any worlds where only the ordinary laws are violated. (One might object that this claim is *ad hoc*. This objection is addressed at the end of this section.) Lewis's account of the similarity relation does not make provisions for a special status of the psychophysical laws. This is not surprising, since Lewis himself was a materialist (see Lewis 1994b). His account can easily be modified to accommodate the distinction, however.[28] Then the new principal criterion for overall similarity to the actual world is that none of the actual psychophysical laws be broken. (Call a violation of the actual psychophysical laws a *psychophysical miracle*.) The new principal criterion can be grafted onto Lewis's original criteria:

(1*) It is of the first importance to avoid violations of psychophysical laws.
(2*) It is of the second importance to avoid big, widespread, diverse violations of ordinary laws of nature.
(3*) It is of the third importance to maximize the spatiotemporal region throughout which perfect match of particular fact prevails.
(4*) It is of the fourth importance to avoid even small, localized, simple violations of ordinary laws of nature.
(5*) It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly.

According to the new set of criteria, a world where a violation of the actual psychophysical laws occurs is always less similar overall to our world than a world without such violations; among worlds that are on a par with respect to violations of the actual psychophysical laws, a world where a large-scale violation of the ordinary actual laws of nature occurs is always less similar overall to our world than a world without such large-scale violations; and so on. According to the new set of criteria, type-1 worlds

[28] A number of authors have suggested different modifications of Lewis's account of similarity recently, including Woodward (2003: 133–145) Kment (2006a), Williams (2008), and Dunn (2011).

come out more similar overall to the actual world than type-2 worlds since they involve no violation of the psychophysical laws. Hence, on the modified account, (17) is true.

Given the new similarity criteria, premise (16) comes out true too. Worlds where the antecedent of (16) holds in the absence of a psychophysical miracle are closer to the actual world than any worlds where such a miracle takes place. But if the actual psychophysical laws are intact in a world where none of *M*'s physical bases is instantiated, *M* is not instantiated there either. So in the closest worlds where none of *M*'s physical bases is instantiated, *M* is not instantiated either. Hence, (16) is true.[29]

The new similarity criteria (1*)–(5*) do not commit us to backtracking evaluations of counterfactuals, at least not any more than the original criteria (1)–(4) do. Our psychophysical laws are synchronic, so holding them fixed by itself never requires changing the past. Even if the new similarity criteria sometimes yield backtracking evaluations because the old criteria sometimes do, the argument for dualist mental causation emerges unscathed. None of its counterfactual premises involves such a backtracking reading. The instantiation of a base of *M* can be prevented by a small miracle just before the time at which *M*'s actual base is actually instantiated. Thus, the closest worlds where no base of *M* is instantiated match the actual world perfectly in particular fact until just before the time at which *M* is instantiated in the actual world. By (16) and (17), the closest worlds where no base of *M* is instantiated coincide with the closest worlds where *M* is not instantiated. Thus, the closest worlds where *M* is not instantiated match the actual world perfectly in particular fact until just before the time at which *M* is instantiated in the actual world too. Hence, no backtracking ensues under the counterfactual supposition that *M* is not instantiated.[30] In particular, the truth of counterfactual (19), which expresses the counterfactual dependence of the *P**-instance on the *M*-instance, is not due to a backtracking evaluation. We saw in Section 1.5 that we can restrict our principle about causation to cases of counterfactual dependence that are not due to backtracking evaluations of the relevant counterfactuals. This restricted principle can be applied here, so the conclusion that the *M*-instance causes the *P**-instance still follows.[31]

[29] For a discussion of (16), albeit in the context of non-reductive physicalism, see Kallestrup 2006: 473.

[30] Except perhaps into the very near past, as was the case for the original criteria (1)–(4).

[31] Instead of modifying the similarity criteria of the miracles approach in order to make room for dualist mental causation, one could try to modify the asymmetry-by-fiat approach. Instead of requiring that the antecedent be made true with minimal difference to the actual world, one

So far the argument merely establishes that mental events cause some physical events or other, because the absence of their physical bases would have made some difference to the physical future. But, as in the case of non-reductive physicalism, we can easily apply the argument to specific pairs of mental and physical events, such as my headache and my hand's moving towards the aspirin. For a specific physical effect, the relevant premise, (18), is just as plausible as the corresponding premise (2) was in the non-reductive physicalism case.

We have seen that, assuming naturalistic dualism, the critical condition for establishing that behavioural events counterfactually depend on, and hence are caused by, mental events is that worlds where the actual psychophysical laws are violated are always less similar overall to our world than worlds without such violations, irrespective of violations of ordinary laws of nature. Call the conjunction of this condition and the position of naturalistic dualism as it was defined earlier *super-nomological dualism*. Put less technically, the position of the super-nomological dualist is that the relation between the mental and the physical is a matter of law, but that the relevant laws are 'more necessary' than ordinary laws of nature.[32] The upshot so far is this: while other varieties of dualism may struggle at the task, super-nomological dualism can show that mental events have physical effects. Thus, super-nomological dualism can solve the interaction problem.

The argument for mental causation under super-nomological dualism can be illustrated geometrically. Imagine the modal universe spread out on a plane, with the actual world (@) at the centre. In this framework, Figure 2.3 represents a typical way for a counterfactual $\phi \mathrel{\Box\!\!\rightarrow} \psi$ to be non-vacuously true (see Lewis 1973b: 17). Applied to our case, we may represent the truth of premise (18) similarly (see Figure 2.4). What premises (16) and (17) add to this picture is a sphere of worlds **S** in which *M* is instantiated just in case a base of *M* is instantiated and which contains a world *w* where neither *M* nor a base of *M* is instantiated (see Figure 2.5). (Intuitively, we

could require that it be made true with minimal difference to the actual world *provided that this does not involve a psychophysical miracle*. As it stands, this suggestion is incomplete, however. It does not say, for instance, how the truth of the antecedent is to be brought about if the antecedent can only be made true at the cost of a psychophysical miracle. The modified miracles approach provides a neater solution.

[32] Super-nomological dualism is incompatible with Armstrong's (1983) theory of laws of nature, for Armstrong's theory involves a single universal of necessitation that is responsible for lawhood. By contrast, super-nomological dualism is compatible in principle with Lewis's (1973b, 1983, 1994a) 'best system' theory of laws of nature, for Lewis's theory is merely about what the laws of a given world are, not about how easily they could have failed.

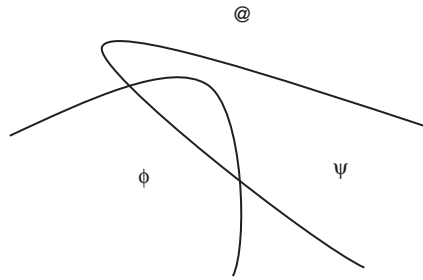

Figure 2.3. $\phi \mathrel{\Box\!\!\rightarrow} \psi$ true

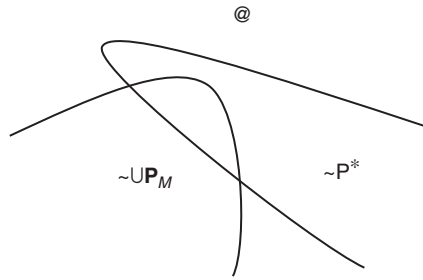

Figure 2.4. (18) (~∪$\mathbf{P}_M \mathrel{\Box\!\!\rightarrow}$ ~P*) true

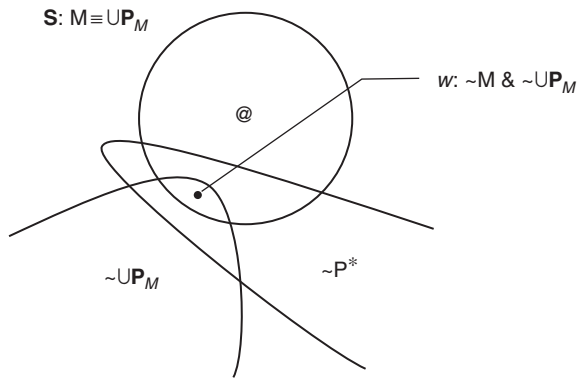

Figure 2.5. (16) (~∪$\mathbf{P}_M \mathrel{\Box\!\!\rightarrow}$ ~M), (17) (~M $\mathrel{\Box\!\!\rightarrow}$ ~∪$\mathbf{P}_M$), (18) (~∪$\mathbf{P}_M \mathrel{\Box\!\!\rightarrow}$ ~P*) true

can think of a sphere of worlds as a set of worlds that resemble the actual world at least to a certain degree. More formally, we can define a sphere as a set of worlds that are closer to the actual world than all the worlds that

https://doi.org/10.1017/9781108762717.004 Published online by Cambridge University Press

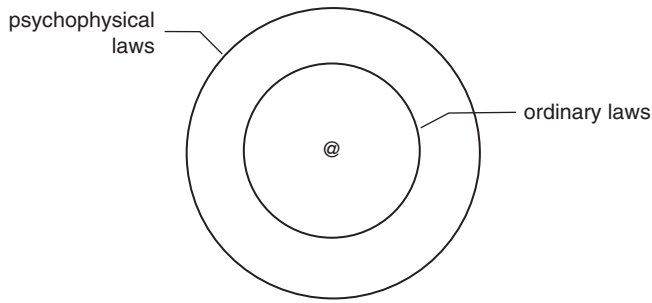

Figure 2.6. A misleading picture of super-nomological necessity

are not in the set (see Lewis 1973b: 4–19). When talking about a proposition being *true in a sphere*, I mean that the proposition is true in all worlds in that sphere.) In the situation represented in Figure 2.5, claim (19) is true, that is, it is true that P^* would not have been instantiated if M had not been instantiated. (The details are explained in Appendix 1.)

Thus, from (16), (17), and (18) we get the existence of a sphere **S** in which M is instantiated just in case a base of M is instantiated and which contains a world w where neither M nor a base of M (nor P^*) is instantiated. In **S**, the psychophysical laws that govern the relation between M and its bases hold. This is not the case for the physical laws. The physical laws are violated at w, where a small miracle prevents the instantiation of a base of M. We need to depart further from actuality to find worlds where the psychophysical laws are broken than we do to find worlds where the ordinary laws of nature are broken. This is of course just what super-nomological dualism says.

It is tempting to generalize Figure 2.5 to a spherical model of nomological and, as it were, super-nomological necessity. According to this model, all worlds where the ordinary laws of nature hold are contained in a sphere, and all worlds where the psychophysical laws hold are contained in a larger sphere (see Figure 2.6).

While there is a certain elegance to this model, super-nomological dualists should not endorse it. If the model is correct, there are no worlds where the psychophysical laws are violated while the ordinary laws of nature are not, because worlds outside of the sphere where the psychophysical laws hold are *ipso facto* outside of the sphere where the ordinary laws of nature hold. Like other dualists, however, super-nomological dualists are likely to hold that there are zombie worlds that are physically like our world but where the psychophysical laws are violated. In zombie worlds, not just the particular physical facts, but also the ordinary laws of nature, are supposed to be like they are in our

world. If there are zombie worlds, we cannot have the picture of super-nomological necessity that is depicted in Figure 2.6. For super-nomological dualists, modal space is less orderly than the spherical model has it.[33]

There is an obvious objection to the account of dualist mental causation presented in this section. The account assumed on behalf of the dualist that the psychophysical laws have a privileged status in the similarity criteria for worlds. Correspondingly, it assumed that these laws could not have failed so easily as the ordinary laws of nature. Assuming such a special modal status for psychophysical laws, however, seems distinctly *ad hoc*.

I offer two replies. First, assuming a distinct modal status for psychophysical laws might be more congenial to dualism than it initially seems. Dualists hold that the mind is special, so they may well hold that the mind is modally special. More specifically, they may hold that a special modal status of the psychophysical laws has independent epistemological virtues. Perhaps it is easier to imagine electricity without magnetism than it is to imagine my body without my mind. If so, this could be straightforwardly explained if the physical laws that link magnetism to electricity could have failed more easily than the psychophysical laws that link my mind to my body.

Second, even if the assumption that the psychophysical laws have a special modal status is made without independent motivation, it may be worthwhile in order to save mental causation, at least for those independently convinced of the truth of dualism. If astrophysicists are allowed to posit dark matter to save their convictions about gravity, why should dualists not be allowed to posit a special modal status for the psychophysical laws to save their conviction that there is mental causation? Jaegwon Kim has argued for reductive physicalism from the existence of mental causation (see, e.g., Kim 1998, 2005). Proponents and detractors of the trope identity theory agree that it is legitimate to make substantial metaphysical assumptions in order to fit the mind into the causal order of the physical world (see Robb 1997, Nordhoof 1998, and Robb 2001). If this general kind of argument is acceptable, it should likewise be acceptable for dualists to fine-tune their metaphysics of mind and adopt super-nomological dualism in order to accommodate mental causation.

[33] Kment endorses a spherical model of different kinds of necessity, but does not discuss the possibility of psychophysical laws that have a special modal status. The context of his discussion also differs from ours in that he allows laws to have exceptions. See Kment 2006a, 2006b, 2014.

2.6 Agency, Transference, and Physical Causes

The arguments for mental causation under non-reductive physicalism and dualism that I have presented in this chapter have drawn on the sufficiency of counterfactual dependence for causation. It might be objected that, even if these arguments solve the interaction problem in the sense that they show that physical events have mental causes, they do not, in the end, give us all we expect of mental causation. In this vein, Kim claims that agency requires more than counterfactual dependence. He holds that agency requires causal processes between mental causes and bodily movements. 'These causal processes', Kim holds,

> all involve *real connectedness* between cause and effect, and the connection is constituted by phenomena such as energy flow and momentum transfer, an actual movement of some (conserved) physical quantity.[34]

We saw in Section 1.6 that it is doubtful that causation generally requires the transfer of some physical quantity because of cases of double prevention. One might still claim that at least mental causation requires such a transfer if it is to yield genuine agency. That claim, however, runs into difficulties that arise from empirical facts about human physiology. The causal processes from mental events to bodily movements involve muscle contractions. As Jonathan Schaffer points out, muscle contractions work by double prevention and thus do not involve the transfer of a physical quantity.[35] In the muscle, myosin proteins are tense. They would bind to actin filaments, move them forward and thus make the muscle contract if it weren't for the obstruction of the binding sites by tropomyosin molecules. If the muscle receives a nerve signal, calcium is released at the neuromuscular junction, which causes the tropomyosin to move away from the binding sites. Muscle contraction works like the examples of the spring and the pillar that were discussed in Section 1.6: an event (here the calcium release) prevents something from happening (the obstruction of the binding sites) which, unless prevented, prevents the another event (the muscle contraction) from happening. Figures 2.7 and 2.8 are neuron diagram representations of the case. Figure 2.7 shows the actual situation; Figure 2.8 shows the situation where no nerve signal is received.[36]

Whatever the exact nature of the causation of a bodily movement by a mental cause, it seems that the mental cause has to operate via the calcium

[34] Kim 2007: 236. Esfeld (2007) advances a similar objection.
[35] See Schaffer 2000a and 2004a. For the physiological details, see Guyton and Hall 2006: 72–84. For another example from biology that involves double prevention, see Woodward 2002.
[36] I borrow these diagrams from Schaffer 2000a: 288 and 2004a: 200 with slight variations.

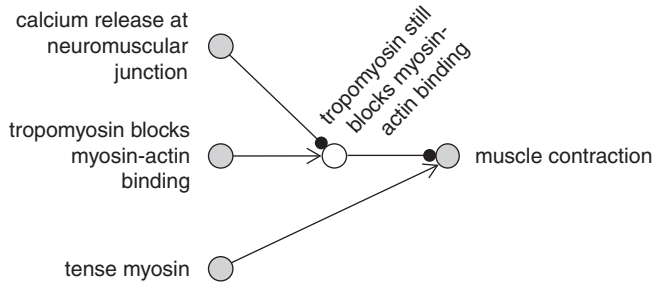

Figure 2.7. Double prevention in muscle contraction

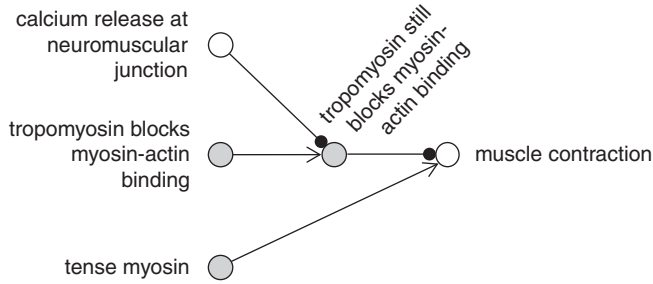

Figure 2.8. If no calcium had been released . . .

release at the neuromuscular junction. (Indeed, it seems that the mental cause already has to operate via intermediate events that are further upstream in the nervous system, for the calcium release takes place quite some time after the mental event does – at least by physiological standards.) If the mental cause transfers something to the bodily movement, it seems that this transfer, too, has to go through the calcium release. But given the facts about human physiology, there is no such transfer because of the double-prevention structure of the case.[37] Thus, if human agency requires there to be a transfer of a conserved physical quantity from mental causes of bodily movements to those bodily movements, then there is no human agency. If we believe that there is human agency, we should conclude that, contra Kim, agency does not require the transfer of a physical quantity from the

[37] This is not to say that no energy is ever transferred on the muscle, of course, but the energy does not come from the nervous system. We shall discuss energy-transferring causes of muscle contractions in Section 4.5.

https://doi.org/10.1017/9781108762717.004 Published online by Cambridge University Press

mental causes to the bodily movements after all.[38] We can still have counterfactual dependence of bodily movements on mental events, however. While this might not give us 'real connectedness' in Kim's sense, it still allows our minds to make a difference to what we do (see Loewer 2007: 255).

(Do the physiological facts about muscle contraction not also have ramifications about the *physical* causes of bodily movements if causation is understood in terms of transfer? They do. We will take up this issue in the context of the exclusion problem in Section 4.5.)

It might seem that the argument against Kim's claim about agency, causal processes, and transfer is somehow parasitic on the assumption that double prevention is causation, which some might not find convincing, despite the strong case that we saw can be made for it. But in fact the issue of whether double prevention is causation is a red herring here. Anyone who agrees with Kim's claim and the assumption that the mental cause operates via the calcium release at the neuromuscular junction will also agree with the following modified claim: in human agency, a physical quantity is transferred to bodily movements from earlier events via the calcium release at the neuromuscular junction. The modified claim does not talk about causation; it merely talks about transfer. The modified claim is still false because of the facts about human physiology. Thus, the Kimian approach to agency is flawed for reasons that are independent of whether double prevention is causation.

So far, I have followed Kim in talking about the transfer (or lack thereof) of a physical quantity in muscle contraction. The arguments generalize to the transfer (or lack thereof) of powers. Appealing to powers has become popular not just in the philosophy of causation in general, but also in attempts to solve the problems of mental causation.[39] We saw in Section 1.6 that powers theories that take the shape of powers transference views have the same trouble with double prevention as standard transference views, because no power is passed from the double preventer to the event that would have been prevented but for the occurrence of the double preventer. Given the mechanism of muscle contraction, no power is transferred from the calcium release at the neuromuscular junction to the movement of the muscle, just as no physical quantity

[38] The applicability of the double-prevention structure of muscle contraction to Kim's claim about agency was discovered independently by Russo (2016). Schaffer (2000a, 2004a, 2012) makes similar points about the kind of causation involved in human agency.

[39] Recent discussions of mental causation that explicitly appeal to powers theories of causation include Heil 2012: 133–134, Gibb 2013 and 2015a, Lowe 2013, Hornsby 2015, Robb 2015, and Mayr 2017. Gibb advocates a powers-based solution to the exclusion problem according to which certain mental events are double preventers that do *not* cause the physical events that would have been prevented but for the occurrence of the mental events. Her suggestion is not motivated by purely physical cases of double prevention like the muscle contraction case, however.

is transferred. Thus, proponents of powers transference views cannot endorse an analogue of Kim's claim that talks about a transfer of powers, any more than proponents of standard transference views can endorse Kim's original claim. Likewise for a modified claim that does not talk about causation, but merely demands a transfer of powers to bodily movements via the calcium release at the neuromuscular junction: the modified claim is empirically false too.

In sum, Kim's 'real connectedness' is not to be had, either as a transfer of a physical quantity or as a transfer of powers. It is more sensible if we do not endorse it in the first place and satisfy ourselves with the result that, as agents, we can make a difference in the physical world because our bodily movements counterfactually depend on what is going on in our minds.

Let me briefly address a question of intra-physical causation before concluding this chapter. So far, I have mostly talked about the causal relation between the instance of a mental property *M* and the later instance of a physical property *P**. What I have said about the actual base or realizer of *M, P*, has been negative. In the context of non-reductive physicalism, I said in Section 2.4 that the argument for the causation of the *P**-instance by the *M*-instance commits us neither to claiming that the *M*-instance causes the *P*-instance nor to claiming that the *P*-instance causes the *P**-instance. The argument for mental causation under dualism yields no such commitments either. There, too, the set of *M*'s physical bases functions merely as a logical intermediary between the *M*-instance and the *P**-instance, not as a causal one, and nothing follows about the role of the actual physical base of *M*.

While no commitment to the *P*-instance's being a cause or effect follows from the arguments I have presented, our sufficient condition for causation can be used independently to make a case for the claim that the *P*-instance causes the *P**-instance. Assume that the *P*-instance is a c-fibre firing and the *P**-instance is my hand's moving towards the aspirin. We may assume that there are no redundant additional physical causes of my hand's moving. We may also assume that there are no pre-empted alternative physical causes of my hand's moving, such as the intervention of the overzealous nurse who would move my hand towards the aspirin if I were not to do it myself. Given these assumptions, it seems that my hand would not have moved if my c-fibres had not fired.[40] The movement occurs later than the c-fibre firing.[41] Therefore, by our sufficient condition for causation, the c-fibre firing causes my hand to move.

[40] For further discussion of this counterfactual, see Lowe 2008: 103–107 and Paprzycka 2014.

[41] Once more, we might have to take a suitable temporal part of the c-fibre firing in order to avoid temporal overlap between the putative cause and the putative effect.

This sounds like a commonsensical result, but besides putting the exclusion problem on the agenda, it conjures up the issues from Section 2.4 about whether realizers can be causes or effects, for we are now committed to the claim that the instances of certain realizers or bases of mental properties are causes. We will return to this issue in Section 4.4.

2.7 Conclusion

This chapter has presented arguments for the existence of mental causation under non-reductive physicalism and dualism. Both views allow us to establish that physical events counterfactually depend on, and hence are caused by, mental events. For non-reductive physicalists, showing that physical events counterfactually depend on mental events is straightforward. For dualists, showing this is less straightforward, but it can still be done if one endorses the super-nomological variety of dualism that assigns a special modal status to the psychophysical laws. Like counterfactual dependence in general, mental causation by counterfactual dependence falls short of showing that mental causes transfer a physical quantity or a power to their physical effects. But that there be such a transfer should not be a requirement for agency, for it is an empirical fact that in humans bodily movements are caused by double prevention and hence do not involve a transfer of a physical quantity from cause to effect.

On the face of it, having accommodated mental causation looks like good news for non-reductive physicalists and super-nomological dualists. This result can be employed in different ways, however, depending on how serious one takes the exclusion problem to be. One could read the result as a *reductio* of non-reductive physicalism and super-nomological dualism: the physical effect has a physical cause that is simultaneous with its mental cause (namely the instance of the realizer or base of the relevant mental property). Thus, the physical effect is overdetermined, like a death by firing squad. But the physical effects of mental causes are not thus overdetermined. Contradiction! Non-reductive physicalism and super-nomological dualism have to go. Alternatively, my argument can be read in favour of non-reductive physicalism and super-nomological dualism: it brings the good news that these positions allow the mental to have physical effects. Granted, a physical effect of a mental cause has a physical cause simultaneous with its mental cause. Depending on what we mean by overdetermination, we might or might not have to call the physical effect overdetermined. But even if we call it overdetermined, there is nothing objectionable or particularly firing-squad-like about the

situation. Far from being a coincidence, the fact that the physical effect has a mental cause in addition to its physical cause is explained by the relation of strong supervenience and nomological supervenience that non-reductive physicalism and super-nomological dualism posit. I prefer the second use of our result, but discussion will have to wait until Chapter 4.