CHAPTER 1

Theories of the Mind and Theories of Causation

1.1 Introduction

Explaining how various theories about the nature of mind can accommodate mental causation requires some groundwork. It requires formulating these theories about the nature of mind more precisely. It also requires getting clearer about the nature of causation, which in turn has two aspects: the nature of the relata of causation and the nature of the relation itself. These are the tasks of this chapter.

The aims of the chapter are modest in several ways. I will not attempt to give a complete taxonomy of views about the nature of mind. Instead, I will confine myself to views that are both common and that stand a *prima facie* chance of accommodating mental causation. I will not say much about the comparative advantages and disadvantages of these positions, not least because I do not wish to commit myself to any of them in this book. I will not try to defend a full-blown theory of causation that states necessary as well as sufficient conditions for causation. For one thing, giving such a full-blown theory of causation may well be impossible. For another, it is not necessary for our purposes. Giving a sufficient condition for causation is enough, at least if this condition can be applied in sufficiently many cases of (putative) mental causation. Fortunately, sufficient conditions for causation are easier to find than necessary ones.

The plan for the chapter is as follows. Section 1.2 formulates different theories about the nature of mind. In particular, it defines reductive physicalism, non-reductive physicalism, and dualism. It also defines a version of dualism, naturalistic dualism. Sections 1.3–1.5 discuss the nature of the causal relata and the causal relation. In the context of mental causation, the causal relata are best conceived of as particular or token events, and particular events are best conceived of as being constituted by, and having their identity determined by, triples of an object, a property, and a time (see Section 1.3).

As for the causal relation, we can give a sufficient condition for causation in terms of difference-making or counterfactual dependence: one event causes another if (roughly) the second event would not have occurred had the first event not occurred. The counterfactual conditionals that are used to formulate claims about counterfactual dependence exhibit some logical peculiarities (see Section 1.4). The sufficient condition for causation in terms of counterfactual dependence is subject to a few *prima facie* problems, but they can be overcome by making certain assumptions about how the relevant counterfactual conditionals should be evaluated and by restricting the sufficient condition to suitable kinds of causal relata (see Section 1.5). The sufficient condition for causation in terms of counterfactuals conflicts with the view that causation requires the transfer of a physical quantity from cause to effect. Section 1.6 argues that this conflict should be resolved in favour of the counterfactual condition. It also discusses the requirement that causation should involve an intrinsic connection between cause and effect.

1.2 Varieties of Physicalism and Dualism

Generally, physicalists with respect to X hold that X is physical.[1] Cartesians deny that mental substances are physical and thus fail to be physicalists about mental substances. This view has few adherents today;[2] I shall follow the mainstream view and assume that all substances are physical. Instead of focusing on substances, I shall focus on mental and physical properties. Physicalists about mental properties hold that mental properties are physical. This claim can be formulated in different ways. The most straightforward reading is that mental properties are identical to physical properties. This is the position called reductive physicalism, defined as follows:

Reductive physicalism: Each mental property is identical to a physical property.

According to reductive physicalism, the property of having a headache is identical to, say, the property of having firing c-fibres:[3] to have a headache is one and the same thing as to have firing c-fibres.[4]

[1] Characterizing the physical raises some problems of its own, which I will ignore here. See Crane and Mellor 1990 and Crook and Gillet 2001 for discussion.

[2] Hart (1988) is an exception. Lowe (1996) endorses a more attenuated substance dualism.

[3] Associating pains with c-fibre-firings is empirically questionable, but has a longstanding philosophical tradition, which I follow here. Indeed, I take the liberty of associating c-fibre firings not with pains *per se*, but with headaches.

[4] As was mentioned in Section 0.2, reductive physicalism was originally called the identity theory and was pioneered by Place (1956) and Feigl (1958). For a recent defence, see Polger 2004.

Since identity entails mutual necessitation, reductive physicalism makes it impossible to have a headache without having firing c-fibres and impossible to have firing c-fibres without having a headache.[5] Some physicalists wish to accept the latter impossibility claim while rejecting the former. They hold that mental properties are physical in the sense that mental properties are necessitated by physical properties and necessitate the instantiation of physical properties, but they also hold that the relation between physical and mental properties is many–one, not one–one, as reductive physicalism has it.[6] Thus, there can be different physical properties besides having firing c-fibres that necessitate pain. This yields a form of physicalism that is weaker than reductive physicalism and hence non-reductive.

We can make non-reductive physicalism more precise by using the notion of strong supervenience, which is defined as follows:

Strong supervenience: 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.[7]

Here and throughout, 'necessarily', unless qualified further, expresses metaphysical necessity, that is, truth in all possible worlds or, somewhat more informatively, truth come what may: if pigs were to fly, donkeys were to talk, and particles were to travel faster than light, what is metaphysically necessary would still have been the case. The definition of strong supervenience is a bit cumbersome, but the underlying idea is simple. As an illustration of the definition, consider dot-matrix pictures and their symmetry properties.[8] The symmetry properties of a dot-matrix picture strongly supervene on the arrangement of dots in the picture's matrix. According to the definition, this means that, necessarily, if a picture instantiates a symmetry property at a given time, the dots in the picture's matrix are arranged in a certain way and that this arrangement necessitates the symmetry property whenever

[5] See Kripke 1980. I assume, with Kripke, that the names for mental and physical properties are rigid designators, that is, that they name the same property at every possible world. Lewis, by contrast, takes the names of mental properties to be non-rigid designators, which opens up the possibility of contingent psychophysical identities (see Lewis 1980).

[6] Alternatively, we could characterize the position as saying that mental properties are identical to *higher-order* physical properties, which are distinct from, but stand in the necessitation relation to, sufficiently *fundamental* physical properties. For further discussion of this strategy, see Pauen 2002.

[7] My formulation of the definition follows what McLaughlin calls 'Modal-Operator Strong Supervenience' (1995: 95). For further discussion, see Kim 1984 and McLaughlin 1995.

[8] This kind of example is due to Lewis 1986c: 14.

a picture has it. For instance, a 3×3 dot-matrix picture that is point-symmetric has to have the dot arrangement ∴ or ∵ or ⁝ ⁝, etc., and any picture that has the arrangement ∴ has to be point-symmetric, any picture that has the arrangement ∵ has to be point-symmetric, any picture that has the arrangement ⁝ ⁝ has to be point-symmetric, etc. More generally, strong supervenience says that a supervenient property has to be accompanied by some subvening property (that is, by a **B**-property), which in turn necessitates the supervenient property whenever it is instantiated.

If we combine the claim that mental properties strongly supervene on physical properties with the claim that mental properties are distinct from physical properties, we get a version of non-reductive physicalism. Generally, non-reductive physicalists claim that mental properties are distinct from physical properties but maintain that mental properties stand in a relation of metaphysical dependence to physical properties (see Baker 2009). The canonical way of spelling out this notion of metaphysical dependence is to read it as strong supervenience (see Kim 1993). For the purposes of this book, I shall identify non-reductive physicalism with the combination of the distinctness claim and the strong supervenience claim. (I do this mainly because it yields a clear-cut terminology, but nothing hinges on it; alternatively, one could use a different label, say 'strong supervenience physicalism', for the position thus characterized.) Thus, we get the following definition:

Non-reductive physicalism: Each mental property is distinct from all physical properties, but mental properties strongly supervene on physical properties.

According to non-reductive physicalism, the property of having a headache is distinct from the property of having firing c-fibres and from all other physical properties. But, owing to the strong supervenience of mental properties on physical properties, it is impossible for someone to have a headache without instantiating some physical property that in turn necessitates having a headache.[9] In my case, that physical property is the property of having firing c-fibres, but in other cases (actual or merely possible), it might be the property of having firing x-fibres (which, let us assume, are actually present not in humans but merely in octopuses), the property of having an active semiconductor network of a certain kind in

[9] It seems that this necessitation relation holds only if the relevant physical properties include properties to the effect that certain background conditions obtain and that certain laws of nature hold. This issue will be taken up again in Section 4.4.

one's head, etc.[10] Thus, non-reductive physicalism allows for the multiple realizability of mental properties by physical properties.

Some theorists reject not only that mental properties are identical to physical properties, but also that they strongly supervene on physical properties. The view they advocate is dualism, defined as follows:

Dualism: Each mental property is distinct from all physical properties, and no subset of mental properties strongly supervenes on physical properties.

According to dualism, the property of having a headache is not merely distinct from all physical properties, but can be instantiated without a physical property that would in turn necessitate the property of having a headache. In my case, having a headache is accompanied by having firing c-fibres, but, according to dualism, it is possible for there to be someone with firing c-fibres who does not have a headache. Likewise, according to dualism, it is possible in principle for someone to have a headache without being in any physical state whatsoever.

Dualism can take more or less extreme forms. An extreme form might have it that in some remote corner of the universe there are disembodied creatures with headaches, and that, next year, humans with firing c-fibres will no longer have headaches while erupting geysers will have headaches. Few scientifically minded people accept such an extreme form of dualism. According to a more moderate form of dualism, mental properties are tied to physical properties by laws of nature in a way that is structurally similar to, but modally weaker than, strong supervenience (see Chalmers 1996: 123–171). We can express this moderate form of dualism more precisely by introducing the notion of nomological supervenience:

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 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.

The definition of nomological supervenience is just like the definition of strong supervenience, except that the two occurrences of 'necessarily' (which, remember, we stipulated to mean metaphysical necessity) are

[10] The doctrine of non-reductive physicalism is often associated with the view that particular, token mental events are identical to particular physical events. We shall see in the following section, however, that not all accounts of token events allow this. For further discussion of the relation between non-reductive physicalism and token identity, see Schneider 2012.

replaced by 'it is nomologically necessary that'.[11] Nomological necessity is necessity in view of the laws of nature: something is nomologically necessary just in case it is strictly implied by the actual laws of nature. (Equivalently, something is nomologically necessary just in case it is true in all possible worlds in which all actual laws of nature hold.) We can use the notion of nomological supervenience to formulate the moderate form of dualism, which, following Chalmers (1996), I will call naturalistic dualism:

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.

Naturalistic dualism is obviously a version of dualism as we have defined it. Both reductive physicalism and non-reductive physicalism are versions of physicalism about mental properties, but do not exhaust it. One could hold that some but not all mental properties are identical to physical properties, while the remaining mental properties strongly supervene on physical properties. Then one would neither be a reductive physicalist nor a non-reductive physicalist. The best way to describe such a view is to divide the subject-matter and say that its adherents are reductive physicalists about those mental properties that they take to be identical to physical properties, but non-reductive physicalists about those mental properties that they take to merely strongly supervene on physical properties. Similarly, dualism and physicalism do not exhaust logical space. One could hold that some subset of mental properties strongly supervenes on physical properties (perhaps even that the members of this subset are identical to physical properties), while some other subset of mental properties does not strongly supervene on physical properties. Then one would be neither a physicalist nor a dualist. This kind of view seems odder than the version of physicalism that is neither reductive nor non-reductive, but one could again say that its adherents are physicalists about the first subset of mental properties and dualists about the second.

[11] Indeed, if it weren't for the stipulation that 'necessarily' mean metaphysical necessity, nomological supervenience would be a kind of strong supervenience, for without the stipulation, strong supervenience could accommodate different kinds of necessity, including nomological necessity. If one preferred this more flexible notion of strong supervenience, one could define two species of this genus, say 'metaphysical strong supervenience' and 'nomological strong supervenience', which would correspond to our notions of strong supervenience and nomological supervenience, respectively. Our terminology has the advantage of brevity, however.

One could avoid the possibility of hybrid views by defining the positions differently. In particular, one could define non-reductive physicalism as the strong supervenience thesis conjoined with the claim that *some* mental properties are distinct from physical properties. Similarly, one could define dualism as the claim that *some* mental properties are distinct from physical properties conjoined with the claim that *some* subset of mental properties does not even strongly supervene on physical properties. The resulting positions would exhaust logical space, and they would be entailed by, but not entail, the corresponding positions according to the definitions I have given.[12] It seems to me that the stronger, albeit non-exhaustive, definitions capture the general usage better, but one could make a different terminological choice.

Setting aside the issue of whether or not the definitions should exhaust logical space, there is also some controversy about whether the labels 'reductive physicalism', 'non-reductive physicalism' and 'dualism' are apt for the positions to which I have attached them. Some would object to the definition of reductive physicalism because they think that reduction requires more than identity.[13] Some would object to the definition of non-reductive physicalism because they think that physicalism (non-reductive or otherwise) requires more than strong supervenience.[14] Some would object to the definition of naturalistic dualism because they think that nomological supervenience is sufficient for physicalism and hence for the falsity of dualism (see Kim 2005: 49). (Consequently, they would also object to the definition of dualism *simpliciter*, since, according to the definition, naturalistic dualism is a species of dualism.) I will not address these objections here. It suffices for our purposes that the positions that I have labelled 'reductive physicalism', 'non-reductive physicalism' and 'dualism' have been sufficiently prominent in the philosophy of mind. It does not matter for our purposes whether they really deserve these labels.

There are certain standard objections to each of the positions about the nature of mind. There is one class of objections that have nothing to do with mental causation. Non-reductive physicalists argue that reductive

[12] Strictly speaking, the entailments hold only on the assumption that there are mental properties, which I take for granted here.

[13] See van Riel 2013. Kim (2005: 34) defines non-reductive physicalism (*inter alia*) as the conjunction of the claim that mental properties are not reducible to physical properties and the claim that they are not identical to physical properties. He explicitly acknowledges that identity is necessary for reduction, however (2005: 34), and his arguments for reductive physicalism (2005: 32–69) suggest that he takes it to be sufficient for reduction too. For discussion of the notion of reduction in the context of emergentism, see Stephan 2002.

[14] See Wilson 2005; see also Jackson 1998: 22–23 and Melnyk 2003: 49–70.

physicalists cannot explain the multiple realizability of mental properties, such as the fact that headaches can be accompanied by c-fibre firings as well as x-fibre firings. Dualists argue that reductive physicalists and non-reductive physicalists cannot explain the possibility of zombies, that is, beings that are physically exactly like us but without any conscious thoughts. I am not going to discuss this class of objections. The comparative merits and difficulties of the different views about the nature of mind will occupy us only where mental causation is concerned.

1.3 The Relata of Causation

Statements of causation come in many stripes. We say that my throwing the stone caused the shattering of the bottle (more idiomatically: that it caused the bottle to shatter); that the reef caused the leakage; that smoking causes cancer. When we say that the throwing caused the shattering, we are talking about token events, that is, particular events (the throwing and the shattering). When we say that the reef caused the leakage, we are talking about a particular thing (the reef) and a particular event (the leakage). When we say that smoking causes cancer, we are talking about general phenomena (smoking, cancer); these might in turn be event types or properties. Perhaps these different statements of causation talk about different kinds of causation that are mutually irreducible; perhaps they do not.[15] In any event, the problems of mental causation are primarily problems about causation between particular events, so my focus will be on this. The interaction problem is about how, in principle, particular mental events such as my headache can cause particular physical events such as my hand's moving towards the aspirin. The exclusion problem is about how a particular physical event such as my hand's moving towards the aspirin can have both a particular mental event and a simultaneous particular physical event as its cause without being like a case of overdetermination.[16] This is not to say that properties play no role in these causal relations – indeed, we shall see that they play a crucial role – but the causal relata are best taken to be particular events. Henceforth, when talking about events without further qualification, I shall mean particular events.

[15] One promising approach analyses claims such as 'Smoking causes cancer' as generic statements about token events. See Carroll 1988, 1991 and Swanson 2012b for further discussion.

[16] Lowe (2000, 2008: 41–57) advocates a solution to the exclusion problem according to which mental events cause not physical *events* but *facts* about intra-physical causal relations. If this kind of solution could be made to work, it would still be only a second-best solution. It would be better to have a solution in terms of causation between events.

What are events? In particular, when are we dealing with a single event and when with several events? W.V.O. Quine and others think that events are identical just in case they occur in the same spatiotemporal region.[17] This account individuates events in a rather coarse-grained way. I stroll leisurely. By virtue of strolling leisurely, I stroll. My strolling and my strolling leisurely take place in the same spatiotemporal region. Therefore, according to the Quinean account, my strolling and my strolling leisurely are one and the same event. Someone who is worried about multiplying events beyond necessity will welcome this result. But there are problems. Perhaps my strolling leisurely, but not my strolling *per se*, causes me to feel refreshed afterwards. How can my strolling leisurely cause something that my strolling does not cause, yet be the very same event as my strolling? By Leibniz's law, identical events must have the same effects (and the same causes). Other examples bring out this issue more sharply. A metal sphere rotates and heats up at the same time. (The heating is due to an external source.)[18] The sphere's rotating and the sphere's heating up take place in the same spatiotemporal region. Therefore, according to the Quinean account, the sphere's rotating and the sphere's heating up are one and the same event. This should sound strange even to those who are inclined towards ontological parsimony. And to say that the sphere's rotating and the sphere's heating up have all their causes and effects in common sounds even less plausible than the parallel claim in the case of my stroll. If I place a funny hat on the sphere, it will start rotating too, but the hat's rotation will be caused by the sphere's rotating, not by the sphere's heating up.

These problems force us, I think, to reject the Quinean account. This may seem unfortunate, for the account seems to offer an attractive solution to the problems of mental causation. Suppose that my headache takes place in the same spatiotemporal region as my c-fibre firing does. (The claim that headaches have spatial location is not completely uncontentious, but anyone who is a substance physicalist should find it acceptable.) Then, according to the Quinean account, my headache and my c-fibre firing are one and the same event.[19] Thus, my headache and my c-fibre firing have all their causes and effects in common. The interaction problem

[17] See Lemmon 1967, Smart 1972, Quine 1970: 31–32, Quine 1974: 5, 131–132, and Davidson 1985. Previously, Davidson (1969) had endorsed the individuation of events in terms of sameness of causes and effects, which critics had claimed to be covertly circular.

[18] The example is from Davidson 1969.

[19] This result is a token identity claim, albeit one that, if generalized, remains weaker than Davidson's (1970) famous Anomalous Monism, because it does not claim that the mental is anomalous.

disappears. No one denies that my c-fibre firing can cause my hand to move towards the aspirin; if it does, then *ipso facto* my headache causes my hand to move towards the aspirin. The exclusion problem disappears, too. Since my headache and my c-fibre firing are one and the same cause of my hand's moving towards the aspirin, my hand's moving is not caused twice over. Even non-reductive physicalists and dualists about mental properties could invoke this solution to problems of mental causation, for what matters is merely that particular mental and physical events are identical; the mental and physical properties that are involved can be distinct and need not even stand in a relation of strong supervenience.

Apart from requiring an implausible conception of events, however, this solution to the problems of mental causation loses its appeal on closer scrutiny. When we demand an explanation of how physical effects can have mental causes, we want to know how mental events can cause physical events by virtue of their mental properties. The explanation that has been suggested provides at most an account of how mental events cause physical events by virtue of their physical properties. According to the explanation, my headache causes my hand to move towards the aspirin because it is identical to my c-fibre firing. Presumably, how my c-fibre firing causes my hand to move is in turn explained by the physical properties involved in my c-fibre firing. So we are still lacking an explanation of how mental events can be causes *qua* mental. For all we know, the situation might be like the following: I put an apple on a scale. The apple weighs 100 grams and has a temperature of 20 degrees Celsius. Shortly after putting the apple on the scale, the display flashes '100'.[20] Quineans have to say that the apple's weighing 100 grams (at a certain time shortly before the flashing) is identical to the apple's having a temperature of 20 degrees (at that time), since these events take place in the same spatiotemporal region.[21] We may assume that the apple's weighing 100 grams caused the display to flash '100'. If the apple's weighing 100 grams is identical to the apple's having a temperature of 20 degrees, it follows that the apple's having a temperature of 20 degrees caused the display to flash '100'. This is an implausible result similar to the result that the sphere's heating up caused the hat to turn, but let us accept it for the sake of argument. The temperature causes the flashing, then, but we have not yet explained how it causes the flashing *qua* temperature.

[20] This example is a variation of an example from Honderich 1982.

[21] I am assuming here and throughout that events can be static, that is, that events need not involve change.

One might think that the foregoing considerations do not tell against a coarse-grained individuation of events like the Quinean account as such. Rather, one might think, they show that we should marry a coarse-grained individuation of events to a permissive notion of causation that allows the sphere's warming to cause the hat to rotate and the apple's having a temperature of 20 degrees to cause the display to flash '100'. One might concede that one needs another causal relation, causation-*qua*, that is less permissive and that does not merely relate events, but events together with certain of their aspects. Thus, while the apple's having a temperature of 20 degrees (which is identical to the apple's weighing 100 grams according to the Quinean account of events) does cause the display to flash '100' in the permissive sense, it does not cause the display to flash '100' *qua* temperature, but *qua* mass. Similarly, while the sphere's warming up (which is identical to the sphere's rotating according to the Quinean account of events) does cause the hat to rotate in the permissive sense, it does not cause the hat to rotate *qua* warming up, but *qua* rotating.[22]

Such a view is consistent, but not attractive. First, the problems of mental causation reappear at the level of causation-*qua*. Perhaps my headache causes my hand to move towards the aspirin in the permissive sense, but how can it cause my hand to move *qua* mental? And how can my headache cause my hand to move both *qua* mental and *qua* physical without there being overdetermination at the *qua*-level? Second, the ontological parsimony about events that the Quinean account of events had boasted is outweighed by a proliferation of causal relations. We have few events, but we need an extra kind of causal relation.[23] Proliferating kinds, however, seems worse than proliferating entities *simpliciter* (see Lewis 1973b: 87).

One might try to augment the claim that individual mental events are identical to individual physical events (owing to the Quinean identity condition or for other reasons) in order to save the causal relevance of mental events *qua* mental by adding another ontological layer. Specifically, one could introduce a layer of particularized properties and claim that mental events have physical effects because particularized mental properties are identical to particularized physical properties.

[22] See Horgan 1989 for discussion.

[23] Perhaps one could eschew the permissive notion of causation and hold that causation-*qua* is the only causal relation. But then one would still incur a commitment to extra complexity owing to the higher adicity of causation-*qua*, which seems to outweigh, or at least neutralize, the simplicity of the Quinean account of events.

One way of elaborating this idea is due to David Robb.[24] The details of Robb's suggestion are as follows. Particular mental events are identical to particular physical events. My headache, for instance, is identical to my c-fibres' firing.[25] Whether mental properties are identical to physical properties, Robb holds, depends on what we mean by 'property'. He thinks that 'property' is ambiguous between something particular and something universal (Robb 1997: 186–187). In the particular sense, 'property' means what Robb calls an 'abstract particular' or a 'trope'. Properties in the sense of abstract particulars are supposed to be wholly present in the things that instantiate them, but nowhere else. In the universal sense, 'property' means a universal or a unifying entity; this is (roughly) the sense in which 'property' has been used in this book so far. Robb calls properties in the sense of universals 'types'. To illustrate the difference between types and tropes, consider a case of two wise women, Anna and Hannah.[26] Is Anna's wisdom identical to Hannah's wisdom? If we conceive of Anna's wisdom and Hannah's wisdom as a type, that is, as a property in the sense of a universal, the answer is 'Yes'. If we conceive of Anna's wisdom and of Hannah's wisdom as a trope, that is, as a property in the sense of an abstract particular, the answer is 'No'. Robb holds that mental types are distinct from physical types. In our terminology, Robb denies reductive physicalism. He does, however, hold that each mental trope is identical to a physical trope (1997: 187).

Identifying mental and physical tropes is supposed to ensure not only that mental events cause physical events, but that they do so *qua* mental. They do so, according to Robb, because it is the tropes, not the types, that are responsible for the causal relevance of properties (1997: 187). Thus, for example, the event that is my headache is characterized by a headache-trope, which is identical to a c-fibre-firing-trope. The event has a physical effect, namely my hand's reaching towards the aspirin. The earlier event causes this physical event *qua* mental because it is characterized by a mental trope (which, like all mental tropes, is also a physical trope). It is, as it were, the job of tropes to guarantee causal relevance and, Robb holds, it does not make sense to ask whether tropes in turn get their causal relevance from something else: 'Tropes are not causally relevant *qua* this or that, they are

[24] See Robb 1997. Similar views are defended in MacDonald and MacDonald 1986, Heil 1992: 135–139, Heil and Robb 2003, and Robb 2013. I focus on Robb's position here because it ties in best with the above discussion of event identity.

[25] Robb (1997: 187) endorses this identity claim, but does not derive it from the Quinean identity condition for events.

[26] The example is a variant of an example of Robb's (1997: 186).

causally relevant (or not), period' (1997: 191). The overall picture that Robb advocates is that there are three kinds of entities that are in play in causal relations: events, tropes, and types. Mental events are identical to physical events, and mental tropes are identical to physical tropes, but mental types are distinct from physical types. Tropes are responsible for causal relevance. In particular, they are responsible for mental events being causes *qua* mental.

One might object to Robb's suggestion by saying that, like the earlier suggestion that there is a separate causal relation, causation-*qua*, it needlessly proliferates kinds of entities by postulating the existence of tropes. But this objection could be countered by offering independent arguments for the existence of tropes.[27] It could also be countered by claiming that types reduce to tropes because types are nothing but sets of tropes that resemble one another.[28] A more serious problem for Robb's suggestion is that it is doubtful that it has really solved the problem of how mental events can be efficacious *qua* mental. Any solution to this problem should appeal to something general: when we ask by virtue of what, or *qua* what, a certain event had a certain effect, we expect the answer to tell us something general about that event. Saying that the event caused the effect by virtue of belonging to a certain type would satisfy this expectation, but this answer is not available to Robb, because he holds that mental and physical types are distinct. Saying that the event caused the effect by virtue of being characterized by a certain trope does not satisfy the expectation, for tropes are by definition particular, not general. Locating the causal relevance of events at the level of tropes seems merely to define the *qua* problem away instead of solving it.[29]

Let us return to identity conditions for events. Let us also, from now on, read 'property' in the universal or type sense and not in the trope sense, unless specified otherwise.

Jaegwon Kim thinks that events are constituted by an object, a property that is instantiated by the object, and the time at which the object instantiates the property.[30] Kim endorses the following two conditions,

[27] See Campbell 1990, Heil 2003, and Ehring 2011.

[28] Robb endorses the identity of types with sets of tropes, but holds that his theory of mental causation does not depend on it (1997: 186–188).

[29] A similar worry is expressed in MacDonald and MacDonald 2006; see also Noordhof 1998 and Shoemaker 2001. For further discussion of Robb's view, see Robb 2001, Ehring 2003, Gibb 2004, and Robb 2013.

[30] See Kim 1976. Kim uses 'substance' instead of 'object', but it is clear from his examples that he is not using 'substance' in a metaphysically laden sense. Here and throughout, I will confine myself to monadic events, that is, events that involve a single object (as opposed to multiple objects) and

where '[x, P, t]' stands for the event that is constituted by object x, property P, and time t:

Existence condition: Event [x, P, t] exists [i.e., occurs] just in case object x has property P at time t.
Identity condition: [x, P, t] = [y, Q, t'] just in case $x = y$, $P = Q$, and $t = t'$. (1976: 160–161)

Kim's conditions individuate events more finely than the Quinean account does. The property of rotating is distinct from the property of warming up. Hence, by the identity condition, the sphere's rotating is distinct from the sphere's warming up. Similarly for the apple's weighing 100 grams and the apple's having a temperature of 20 degrees. (The case of my strolling vs my strolling leisurely will be discussed in a moment.) Similarly, too, for my headache and my c-fibre firing if one is a non-reductive physicalist or a dualist: proponents of these positions deem the properties of having a headache and of having firing c-fibres to be distinct; hence, they have to deny the identity of the corresponding events if they accept Kim's identity condition.

Kim's existence and identity conditions allow for a weak and a strong reading. According to the weak reading, the conditions apply only to actual events, objects, properties, and times. According to the strong reading, they apply to all possible events, objects, properties, and times.[31] Let us call the resulting conception of events the *weak Kimian account* and the *strong Kimian account*, respectively.

On the strong reading, Kim's existence condition says that, in any possible world, event [x, P, t] occurs just in case x has P at t in this world. The identity condition says that event [x, P, t] at a possible world w is identical to event [y, Q, t'] at a possible world v (which may or may not be identical to w) just in case x is identical to y, P is identical to Q, and t is identical to t'.[32] On the strong reading, each event has its object, property, and time essentially. That is, no event could have occurred while being constituted by a different object or a different property; nor could any event have occurred at a different time. To see this, take some event [x, P, t] that actually occurs. Assume that object y has property Q at time t' at

a property (as opposed to a relation). I will allow the constitutive time to be an interval that is larger than a point.

[31] Presumably, properties exist at possible worlds at which they are not instantiated. If so, the two readings do not differ with respect to what properties they quantify over.

[32] If one prefers the counterpart relation over trans-world identity, one can substitute 'is a counterpart of' for 'is identical to'.

a different world. Then, by the existence condition, the event [*y*, *Q*, *t′*] occurs at that world. But, by the identity condition, [*y*, *Q*, *t′*] is identical to [*x*, *P*, *t*] only if *y* is identical to *x,* *Q* is identical to *P*, and *t′* is identical to *t*. In other words, [*x*, *P*, *t*] can occur at different possible worlds only if it involves the same object, property, and time there.

Making the object, property, and time essential to an event is problematic, albeit not equally problematic for all the constituents. Suppose that in fact I strolled leisurely between noon and half past noon yesterday. If the object is essential to an event, then my strolling leisurely could not have been someone else's strolling leisurely. This sounds plausible.[33] If the property is essential to an event, then my strolling leisurely could not have happened while I had a property incompatible with strolling leisurely, such as strolling non-leisurely. This may sound less plausible. But we can say that the expression 'my strolling leisurely at noon' can pick out either of two events, one of which involves the property of strolling leisurely and one of which involves the property of strolling *simpliciter*. The former event is essentially a leisurely strolling according to the strong Kimian account; the latter event is merely essentially a strolling and could have happened in a non-leisurely way (see Kim 1976: 163). If the time is essential to an event, then my strolling leisurely could not have occurred at a different time. If I had started to stroll leisurely a second after noon yesterday, I would have strolled a different stroll. This may sound implausible. At least sometimes, it seems, events can be postponed or antedated without being replaced by different events. While the result that events have their time of occurrence essentially is a shortcoming of the strong Kimian account of events, we shall see that this account is still attractive overall.

On the weak reading, Kim's existence condition says that, in the actual world, event [*x*, *P*, *t*] occurs just in case object *x* has property *P* at time *t* in the actual world. The identity condition says that an actual event [*x*, *P*, *t*] is identical to an actual event [*y*, *Q*, *t′*] just in case *x* is identical to *y,* *P* is identical to *Q,* and *t* is identical to *t′*. The weak reading remains silent on whether or not events have their object, property, or time essentially. If an event that differs in one of these constituents from some actual event occurs at a possible world, it may or may not be identical to the actual event. If the possible event is identical to the actual event despite differing

[33] Which is not to say that there are no objections. Kim (1976: 171) discusses the case in which the stroll is a ritual of a secret society that chooses by lottery the person who will stroll. In this case it might seem that, had someone else been chosen, she would have strolled the same stroll.

in some constituent, that constituent is not essential to the event. One can, of course, supplement the weak Kimian account with claims to the effect that certain constituents of events are essential to them.[34] Kim himself sympathizes with the idea that the object is essential to an event, but not the property or the time (1976: 172).

If one endorses the weak Kimian account of events, but, unlike Kim, thinks that properties are essential to events, one can even go further than claiming that it is the constituent property that is essential to an event. One can make the stronger claim that some more specific property is essential to the event. For instance, one could claim that it is not merely essential to the sphere's rotating that the sphere rotates, but that it is essential that the sphere rotates with a certain angular velocity, or with an angular velocity that lies within a certain range. (That range should of course contain the sphere's actual angular velocity.) I shall leave it open for now whether making more specific properties essential to events is a good idea, but at any rate the weak Kimian account of events allows a lot of flexibility with respect to the modal relation between an event and its constituents (and more specific variants of the constituents). Thus, the weak Kimian account of events can be tailored to one's views about the essential and modal properties of events.

David Lewis thinks that events necessarily occur in spatiotemporal regions (1986b). Indeed, he thinks that actual and possible events are identical just in case they occur in the same actual and possible spatiotemporal regions.[35] 'Regions', not 'region': one and the same event may well occur in different spatiotemporal regions in different worlds. According to Lewis, events may have essential features, but we cannot read off these features from the nominalization that we use to pick out events. We use 'my strolling leisurely between noon and half past noon yesterday' to pick out a certain event, but it does not follow that this event necessarily involves me or a leisurely stroll; nor does it follow that this event necessarily occurs between noon and half past noon yesterday.

The Lewisian account and the weak Kimian account are very similar. They are in conflict only if in the actual world (i) a putative event occurs in a spatiotemporal region without being constituted (*inter alia*) by an object and a property or (ii) a non-spatial object has a property at a time.

[34] Strictly speaking, one could endorse a *weak* Kimian account and yet hold that all three constituents are essential to an event. Then some events in different worlds that have all three constituents in common could nonetheless be distinct. It is hard to see the motivation for such a view, however.

[35] In Lewis's own terminology, events correspond to properties, that is (according to Lewis), sets of spatiotemporal regions – those spatiotemporal regions in which the events occur (1986b: 244).

According to the Lewisian account, a genuine event occurs in case (i), but not in case (ii). According to the weak Kimian account (or the strong Kimian account, for that matter), a genuine event occurs in case (ii), but not in case (i). Case (i) is impossible if we can find a suitable property to ascribe to the spatiotemporal region itself.[36] The best candidates for case (ii) seem to be Cartesian souls that exist in time but not in space. Few contemporary philosophers, including contemporary dualists, endorse their existence, however, and it is not even clear that the advocates of Cartesian souls can uphold that they are not spatial (see Lycan 2009, Bailey *et al.* 2011).

Given their similarity, it does not come as a surprise that the weak Kimian account and the Lewisian account of events share strengths and weaknesses. Like the weak Kimian account (and the strong Kimian account), the Lewisian account has the advantage of ruling the sphere's rotating and the sphere's warming up to be distinct events. While they occur in the same spatiotemporal region in the actual world, presumably there are possible regions where either event occurs without the other. Hence, by Lewis's identity condition, they are distinct. Like the weak Kimian account of events, the Lewisian account allows events to have essential features. Like the weak Kimian account, but unlike the strong Kimian account, the Lewisian account does not entail that events have certain specific essential features.

All this seems to tell in favour of the weak Kimian account and the Lewisian account vis-à-vis the strong Kimian account. The strong Kimian account has an advantage over these two views, however: it states necessary and sufficient conditions for the occurrence of a given event that hold in all possible worlds, not merely in the actual world. Moreover, it states these conditions in terms of objects, properties, and times. These two features of the strong Kimian account greatly facilitate the squaring of claims about the occurrence or non-occurrence of events in possible situations with claims about supervenience, which are also formulated in terms of objects, properties, and times. Many of the arguments in the following chapters will derive claims about the occurrence or non-occurrence of events in certain possible situations, and subsequently claims about causation, from claims about supervenience. It will therefore be convenient to assume a strong Kimian account of events. I will not overindulge in convenience, however. Sometimes the weak Kimian account and the Lewisian account will yield different results than the strong Kimian account. In these cases,

[36] For a similar suggestion, see Brand 1977: 335.

I will consider the ramifications that ensue if one of these accounts of events is accepted instead.

1.4 Causation and the Logic of Counterfactuals

What makes a difference is a cause. This is the central principle about causation that I shall use in this book. An event occurs, followed by another event. If the earlier event had not occurred, the later event would not have occurred. Therefore, by the principle, the earlier event caused the later event. For example, I throw a dart at a balloon; an instant later the balloon bursts. If I had not thrown the dart, the balloon would not have burst. Therefore, by the principle, my throw causes the balloon to burst. We can formulate the principle more concisely by using the notion of counterfactual dependence. Say that event *e counterfactually depends* on event *c* just in case *e* would not have occurred if *c* had not occurred.[37] Then the principle says that for any two events *c* and *e* that actually occur, if *e* occurs later than *c*, and *e* counterfactually depends on *c*, then *c* causes *e*.[38]

The principle states a sufficient condition for causation in terms of counterfactual dependence. If there are any plausible claims about causation, the principle is one of them.[39] Its plausibility might, however, be obscured by problems that beset the more ambitious project of giving not merely sufficient conditions for causation in terms of counterfactuals, but necessary conditions as well. There are two kinds of cases that are particularly troublesome for the more ambitious project. First, there are cases of so-called late pre-emption. Billy and Suzy each throw a rock at a bottle at the same time. Billy's rock arrives there first and shatters the bottle. But if Billy had not thrown, Suzy's rock would have shattered the bottle anyway.[40] Second, there are cases of overdetermination. Both members of a firing squad of two simultaneously fire at the victim. The victim dies. Each firing killed the victim, it seems, but if either member had not shot, the victim would still

[37] I follow the common practice of taking counterfactual dependence to capture our informal notion of difference-making. Sartorio (2005, 2016: 94) has a different notion of difference-making that can apply in the absence of counterfactual dependence.

[38] Sometimes the counterfactual dependence of *e* on *c* is taken to require not just that *e* would not have occurred if *c* had not occurred, but also that *e* would have occurred if *c* had occurred. Given Lewis's truth-conditions for counterfactual conditionals, which will be presented shortly, the second counterfactual conditional is redundant given that (i) *c* and *e* actually occur and that (ii) any world (including the actual world) is closer to itself than any other worlds are. The assumption that any world is closer to itself than any other world is known as Strong Centring (see Lewis 1973b: 120). The account of mental causation developed in List and Menzies 2009 rejects Strong Centring.

[39] See Lewis 2004: 78, Schaffer 2004b: 240.

[40] So-called cases of trumping pre-emption give rise to similar problems; see Schaffer 2000b.

have died.[41] These cases show that counterfactual dependence is not necessary for causation. Billy's throw causes the bottle to shatter, but the bottle would have shattered even if Billy had not thrown. The firing of each squad member causes the victim to die, but if either squad member had not fired, the victim would have died anyway from the other member's shot.[42] The cases are not counterexamples to the sufficiency of counterfactual dependence for causation, however, which is all that our principle claims.

In order to apply the principle, we need to know more about how counterfactual conditionals ('counterfactuals' for short), the claims that express counterfactual dependence, work. The counterfactuals we have considered so far have had the specific form 'If this event had not occurred, then that event would not have occurred', but it will be useful to have a general account that can handle any claims of the form 'If ϕ were the case, then ψ would be the case' (in symbols, $\phi \mathrel{\square\!\!\rightarrow} \psi$), irrespective of what the antecedent ϕ and the consequent ψ look like. I shall assume Lewis's (1973b) truth-conditions for counterfactual conditionals. Assume that we can order all possible worlds according to how similar they are, overall, to the actual world. Let us think of the less similar worlds as more distant from the actual world. Lewis's idea is that a counterfactual is true just in case we have to depart further from the actual world to find a world where the antecedent of the conditional is true while its consequent is false than we have to in order to find a world where both the antecedent and the consequent are true. For example, the counterfactual 'If there had been no water on Earth, then no life would have developed there' is true just in case we have to depart further from actuality in order to find a world with a dry Earth where life still developed than we have to in order to find a world with a dry Earth where no life developed.[43] Lewis's idea can be applied only if the antecedent is metaphysically possible; otherwise, Lewis stipulates, the conditional is always vacuously true. More technically, Lewis's truth-conditions are as follows: a counterfactual 'If ϕ were the case, then ψ would be the case' ($\phi \mathrel{\square\!\!\rightarrow} \psi$) is true if and only if either

(i) there is no possible world where ϕ is true; or
(ii) there is a possible world where ϕ and ψ are true that is closer

[41] For an argument for the claim that the individual overdetermining events cause the overdetermined event, see Schaffer 2003. Lewis remains neutral on this claim (1973a: 567 n. 12).

[42] In order to deal with certain other cases of pre-emption, Lewis suggests that not counterfactual dependence by itself, but the existence of a chain of events that are related by stepwise counterfactual dependence, is necessary and sufficient for causation (1973a: 567). This condition is violated in cases of late pre-emption and overdetermination, however. See Lewis 1986d for further discussion.

[43] The imagined journey takes place in the modal universe of possible worlds (sometimes called a pluriverse) and not in the actual universe, so the features of nearby earthlike planets in our actual galaxy have at best an indirect bearing on the truth of the present counterfactual.

(that is, more similar overall) to the actual world than any worlds where ϕ is true while ψ is false.

The truth-conditions also allow us to formulate truth-conditions for another kind of counterfactual conditional, namely conditionals of the form 'If ϕ were the case, then ψ might be the case.' Such 'might' conditionals are not directly relevant to the counterfactual dependence between events, but they will play a role in later arguments. I will follow Lewis in taking 'If ϕ were the case, then ψ might be the case' ($\phi \Diamond\!\!\rightarrow \psi$) to be equivalent to the negation of 'If ϕ were the case, then ψ would *not* be the case.' Thus 'If ϕ were the case, then ψ might be the case' is true just in case there is a world where both ϕ and ψ are true which is at least as close to the actual world as any worlds where ϕ is true while ψ is false. For example, the 'might' conditional 'If the coin had been tossed, it might have fallen heads' is true just in case there is a world where the coin is tossed and it falls heads that is at least as close as any worlds where the coin is tossed but it does not fall heads. Most of the counterfactual conditionals we shall be concerned with are 'would' conditionals rather than 'might' conditionals; I will therefore use 'counterfactual' without qualification to refer to the former kind of conditional.

It is tempting to paraphrase Lewis's truth-conditions for counterfactuals and 'might' conditionals by speaking about the closest worlds where their antecedents are true. According to this paraphrase, a counterfactual is non-vacuously true just in case its consequent is true at the closest worlds where its antecedent is true (for short: at the closest antecedent-worlds). For example, 'If there had been no water on Earth, then no life would have developed there' is non-vacuously true just in case no life developed at the closest worlds where there is no water on Earth. Similarly, a 'might' conditional is true according to the paraphrase just in case its consequent is true in some of the closest antecedent-worlds. 'If the coin had been tossed, it might have fallen heads', for example, is true just in case the coin falls heads in some of the closest worlds at which it is tossed.

The 'closest worlds' paraphrases presuppose that there is a set of closest antecedent-worlds for any 'would' or 'might' conditional. It presupposes, that is, that for any such conditional there is a set of antecedent-worlds such that no worlds are closer to the actual world than the members of this set.[44] It is doubtful whether we can always find such a set, however. Consider the following example. The counterfactual 'If I were nearer to Hammerfest

[44] The claim that there is such a set for each 'would' and 'might' conditional is the so-called Limit Assumption. See Lewis 1973b: 19–21 and Swanson 2012a for further discussion.

now, I would still be alive' is true (I hope). Thus, according to the truth-conditions, there is a world – that is, there is at least one world – in which I am nearer to Hammerfest now and I am alive that is closer to the actual world than any worlds in which I am nearer to Hammerfest now without being alive. According to the 'closest worlds' paraphrase, there is also a set of worlds in which I am nearer to Hammerfest now (and in which I am alive) that are closer to the actual world than any other worlds in which I am nearer to Hammerfest now. There need not be any such set of worlds, however (let alone a unique such world). Presumably, there is a world where I am one metre nearer to Hammerfest now and still alive that is closer to the actual world than any worlds where I am nearer to Hammerfest now without being alive. The existence of such a world suffices to satisfy the truth-conditions for counterfactuals, but it does not suffice for the existence of a set of closest antecedent-worlds for our conditional. For it might well be the case that how close worlds in which I am nearer to Hammerfest now are to the actual world varies with how near I am to my actual position in those worlds. It might well be, that is, that worlds where I am one centimetre nearer to Hammerfest now are closer to the actual world than worlds where I am one metre nearer to Hammerfest now; that worlds where I am one millimetre nearer to Hammerfest now are closer to the actual world than worlds where I am one centimetre nearer to Hammerfest now; etc. Thus, it might well be that for all worlds where I am nearer to Hammerfest now, there are other worlds where I am nearer to Hammerfest now that are closer still to the actual world because in these other worlds I am nearer to where I actually am. If so, there is no set of closest antecedent-worlds, as the 'closest worlds' paraphrase has it. Despite this complication, I will use the paraphrase for convenience in cases where the presupposition is harmless.

From Lewis's truth-conditions for counterfactuals and 'might' conditionals we can assess various inferences that involve counterfactuals as valid or invalid. Having a repertoire of valid inferences that involve counterfactuals to hand will allow us to formulate arguments for claims about counterfactual dependence between mental and physical events and related claims in later chapters.

One inference that we will use repeatedly is the implication of a counterfactual by the corresponding strict conditional (where a strict conditional is a material conditional that is prefixed by 'Necessarily'):

(1) Necessarily, if ϕ is the case, then ψ is the case. ($\Box[\phi \supset \psi]$)

---

(2) If ϕ were the case, then ψ would be the case. ($\phi \mathrel{\Box\!\!\rightarrow} \psi$)

https://doi.org/10.1017/9781108762717.003 Published online by Cambridge University Press

(In the notation, $\supset$ is the material conditional and $\Box$ the metaphysical necessity operator.) The inference from (1) to (2) is valid because if ψ is true in all ϕ-worlds, as (1) says, then *a fortiori* ψ is true in all closest ϕ-worlds, as (2) says.[45]

Certain inferences that are valid for material conditionals and strict conditionals are invalid for counterfactuals. For our purposes, issues of transitivity will be particularly relevant. Material conditionals and strict conditionals are transitive; that is, the following inferences are valid:

(3) If ϕ is the case, then χ is the case. $(\phi \supset \chi)$[46]
(4) If χ is the case, then ψ is the case. $(\chi \supset \psi)$

(5) If ϕ is the case, then ψ is the case. $(\phi \supset \psi)$

(6) Necessarily, if ϕ is the case, then χ is the case. $(\Box[\phi \supset \chi])$
(7) Necessarily, if χ is the case, then ψ is the case. $(\Box[\chi \supset \psi])$

(8) Necessarily, if ϕ is the case, then ψ is the case. $(\Box[\phi \supset \psi])$

Counterfactuals, by contrast, are not transitive; that is, the following inference is invalid (see Lewis 1973b: 32):

(9) If ϕ were the case, then χ would be the case. $(\phi \mathrel{\Box\!\!\rightarrow} \chi)$
(10) If χ were the case, then ψ would be the case. $(\chi \mathrel{\Box\!\!\rightarrow} \psi)$

(11) If ϕ were the case, then ψ would be the case. $(\phi \mathrel{\Box\!\!\rightarrow} \psi)$

That counterfactuals fail to be transitive can be shown abstractly from their truth-conditions, but can also readily be seen from concrete examples, such as the following:

(12) If I were king, I would wear a crown.
(13) If I wore a crown, people would find me ridiculous.

(14) If I were king, people would find me ridiculous.

[45] Friends of false counterpossibles, that is, false counterfactuals with impossible antecedents, will disagree. For a given (allegedly) false counterpossible, they cannot accept that it is logically implied by the corresponding strict conditional, which is trivially true owing to the impossible antecedent. Friends of false counterpossibles can still accept the weaker claim that strict conditionals with possible antecedents logically imply the corresponding counterfactuals. But in any event the existence of false counterpossibles is incompatible with Lewis's truth-conditions, which are assumed here. For further discussion, see Williamson (forthcoming).

[46] For simplicity, I am expressing material conditionals by indicative conditionals in natural language here, but the relation between these two kinds of conditional is notoriously difficult; see Jonathan Bennett 2003 for discussion.

https://doi.org/10.1017/9781108762717.003 Published online by Cambridge University Press

The closest possible worlds where I am king are rather remote from the actual world. In these closest worlds I use the usual insignia of the monarchy, including a crown. Thus, (12) is true. The closest worlds where I wear a crown are not quite so remote. Presumably, the closest worlds where I wear a crown are worlds where I buy one from a fancy dress shop and wear it on my way back home. In those worlds, people find me ridiculous, so (13) is true. But in the more distant worlds where I am king, people do not find me ridiculous, so (14) is false. More generally, counterfactuals can fail to be transitive when the antecedent of the first premise takes us to more distant worlds than the antecedent of the second premise does. In such a case the shared consequent of the second premise and the conclusion can be true in the less distant worlds but false in the more distant worlds.

The failure of transitivity persists even if we strengthen the first premise of the inference by replacing the counterfactual by a strict conditional. This strengthening yields the following inference:

(15) Necessarily, if ϕ is the case, then χ is the case. ($\Box[\phi \supset \chi]$)
(16) If χ were the case, then ψ would be the case. ($\chi \mathrel{\Box\!\!\rightarrow} \psi$)

---

(17) If ϕ were the case, then ψ would be the case. ($\phi \mathrel{\Box\!\!\rightarrow} \psi$)

This inference is invalid, too. Again, this can be shown abstractly or illustrated by a counterexample such as the following:[47]

(18) Necessarily, if I got up at 3 a.m., I got up before 9 a.m.
(19) If I had got up before 9 a.m., I would still have been rested.

---

(20) If I had got up at 3 a.m., I would still have been rested.

Suppose that, actually, I get up at 9 a.m. after a long night's sleep, perfectly rested. Assuming that, if I had got up earlier, I would not have gone to bed earlier (more on this kind of assumption below), (20) is false, because, given this assumption, I am sleep-deprived in the closest worlds at which I get up at 3 a.m. By contrast, (19) is true. For we may assume that, first, the closer in time my getting up is to my actual getting up, the closer the corresponding world is to the actual world and that, second, in such a world I am still rested if I do not get up much earlier than 9 a.m.[48] The strict conditional (18) is obviously true. In sum, the premises of the inference are true, but the conclusion is false, so the inference is invalid.

[47] Lewis (1973b: 32) gives a similar counterexample.
[48] The first assumption makes (19) a temporal analogue of the Hammerfest example discussed above.

That the above inferences involving counterfactuals are invalid should not mislead one into thinking that no interesting transitivity-like reasoning with counterfactuals is possible. For we can find substitutes for those inferences that are valid (see Lewis 1973b: 31–36, 1973c). The idea behind the substitute inferences is to patch up the premises so that the closest antecedent-worlds of the premises no longer come apart in a way that threatens the truth of the conclusion.

This can be done in several ways. The first inference, that from (9) and (10) to (11), can be repaired by adding another premise to the effect that the antecedent of the first premise not only counterfactually implies the antecedent of the old second premise, but that the converse is also true:

(9) If ϕ were the case, then χ would be the case. ($\phi \mathrel{\Box\!\!\rightarrow} \chi$)
(21) If χ were the case, then ϕ would be the case. ($\chi \mathrel{\Box\!\!\rightarrow} \phi$)
(10) If χ were the case, then ψ would be the case. ($\chi \mathrel{\Box\!\!\rightarrow} \psi$)

---

(11) If ϕ were the case, then ψ would be the case. ($\phi \mathrel{\Box\!\!\rightarrow} \psi$)

Together, premises (9) and (21) guarantee that the closest ϕ-worlds coincide with the closest χ-worlds, for (9) says that the closest ϕ-worlds are χ-worlds, and (21) says that the closest χ-worlds are ϕ-worlds. By (10), the closest χ-worlds are ψ-worlds, so together with the coincidence claim it follows that the closest ϕ-worlds are ψ-worlds, as conclusion (11) says.[49]

Another way of repairing the inference from (9) and (10) to (11) is to replace (10) with a premise with a stronger antecedent, viz. claim (22) below; the resulting inference is sometimes called *restricted transitivity*:

(9) If ϕ were the case, then χ would be the case. ($\phi \mathrel{\Box\!\!\rightarrow} \chi$)
(22) If ϕ and χ were the case, then ψ would be the case. ($\phi \mathbin{\&} \chi \mathrel{\Box\!\!\rightarrow} \psi$)

---

(11) If ϕ were the case, then ψ would be the case. ($\phi \mathrel{\Box\!\!\rightarrow} \psi$)

Again, this manoeuvre guarantees that the closest antecedent-worlds of the premises coincide. For if (9) is true, the closest ϕ-worlds are χ-worlds and thus are also the closest ϕ-and-χ-worlds, and the closest antecedent-worlds of (22) are trivially the closest ϕ-and-χ-worlds. The inference is valid because by (22) the closest ϕ-and-χ-worlds are ψ-worlds and by (9) the closest ϕ-worlds are also the closest ϕ-and-χ-worlds and so are ψ-worlds.

[49] For further discussion of the inference from (9), (21) and (10) to (11), see Stalnaker 1968, Lewis 1973b: 33, Tooley 2002 and Cross 2006.

The inference that involved strict as well as counterfactual conditionals, that from (15) and (16) to (17), can be repaired by turning the strict conditional (15) into a strict biconditional, claim (23):

(23) Necessarily, ϕ is the case if and only if χ is the case. ($\Box[\phi \equiv \chi]$)
(16) If χ were the case, then ψ would be the case. ($\chi \mathrel{\Box\!\!\rightarrow} \psi$)

---

(17) If ϕ were the case, then ψ would be the case. ($\phi \mathrel{\Box\!\!\rightarrow} \psi$)

(In the additional symbolism, $\equiv$ is the material biconditional.) The manoeuvre of replacing (15) with (23) makes sure that the closest ϕ-worlds and the closest χ-worlds coincide. Indeed, by (23), *all* ϕ-worlds and χ-worlds coincide.[50] By (16), the closest χ-worlds are ψ-worlds, so with the coincidence result we get that the closest ϕ-worlds are ψ-worlds.

Another way of repairing the inference that involved both counterfactuals and strict conditionals is to switch the role of the counterfactual premise and the strict conditional premise. Thus, we get the following inference:

(24) If ϕ were the case, then χ would be the case. ($\phi \mathrel{\Box\!\!\rightarrow} \chi$)
(25) Necessarily, if χ is the case, then ψ is the case. ($\Box[\chi \supset \psi]$)

---

(17) If ϕ were the case, then ψ would be the case. ($\phi \mathrel{\Box\!\!\rightarrow} \psi$)

While (24) and (25) allow the closest ϕ-worlds to be more distant than the closest χ-worlds, they do not allow ψ to be true only in the relatively close χ-worlds. For by (25) ψ is true in all χ-worlds; *a fortiori* ψ is true in the closest ϕ-worlds, because those, by (24), are also χ-worlds. Thus, (17) follows from (24) and (25).

Besides transitivity, further inferences that are valid for material conditionals and strict conditionals are invalid for counterfactuals. For instance, unlike material conditionals and strict conditionals, counterfactuals do not allow strengthening of the antecedent. That is, the following inference is invalid:

(26) If ϕ were the case, then ψ would be the case. ($\phi \mathrel{\Box\!\!\rightarrow} \psi$)

---

(27) If ϕ and χ were the case, then ψ would be the case. ($\phi \,\&\, \chi \mathrel{\Box\!\!\rightarrow} \psi$)

The inference is invalid for reasons similar to the reasons why counterfactuals fail to be transitive: the closest worlds where both ϕ and χ are true may be more

[50] The inference from (23) and (16) to (17) is valid not only given Lewis's truth-conditions for counterfactuals, but in any logic for counterfactuals that creates non-hyperintensional contexts in Williamson's (2006: 312) sense. That is, the inference is valid in any logic for counterfactuals that allows the substitution of strictly equivalent propositions *salva veritate*.

https://doi.org/10.1017/9781108762717.003 Published online by Cambridge University Press

remote and different in character from the closest worlds where merely ϕ is true.[51] For instance, at the closest worlds where I strike the match, the match is dry and it lights. But at the closest worlds where I strike the match and it is wet, it does not light. Thus, the premise of the following argument is true but its conclusion false:

(28) If I had struck the match, then it would have lit.

---

(29) If I had struck the match and the match had been wet, then it would have lit.

Like the transitivity-inferences, the inference from (26) to (27) can be repaired to restore validity. We can, for example, add the further premise that if ϕ were the case, χ might be the case, which yields the following argument (see Lewis 1973c):

(26) If ϕ were the case, then ψ would be the case. ($\phi \mathrel{\Box\!\!\rightarrow} \psi$)
(30) If ϕ were the case, then χ might be that case. ($\phi \mathrel{\Diamond\!\!\rightarrow} \chi$)

---

(27) If ϕ and χ were the case, then ψ would be the case. ($\phi \,\&\, \chi \mathrel{\Box\!\!\rightarrow} \psi$)

Adding premise (30) makes sure that some of the closest ϕ-worlds, which, by (26), are all ψ-worlds, are χ-worlds. Thus, the closest ϕ-and-χ-worlds as well as the closest ϕ-worlds are ψ-worlds, and (27) is true.

The lesson to be learned from the logical peculiarities of counterfactuals is twofold. First, we may not simply use familiar inferences that are valid for other kinds of conditional, for these inferences may fail for counterfactuals. Second, we can find similar inferences that are valid for counterfactuals and that we can substitute for the invalid ones. In philosophical debates the first point has received more attention than the second. But this should not make us pessimistic about reasoning with counterfactuals, because the second point shows that, if we are sufficiently careful, we can still build powerful arguments with counterfactuals.

The following chapters capitalize on the logic of counterfactuals in order to solve the problems of mental causation. In particular, they derive claims about counterfactual dependence, and thence claims about causation, between mental and physical events from other counterfactuals or from other counterfactuals together with claims about necessity. For this strategy, the substitute inferences for transitivity will be especially relevant, but other inferences will also play a role. New inferences will be explained

[51] Lewis (1973b: 31–36) holds that the transitivity failures of counterfactuals can be regarded as generalizations from the failure of the present inference.

when they first appear in the text. Appendix 2 contains a list of valid and invalid inferences involving counterfactuals for easy reference.

1.5 Counterfactual Dependence and Similarity between Worlds

The central ingredient in the truth-conditions for counterfactuals is the relation of comparative overall similarity or closeness between worlds, that is, the relation expressed by 'world *w* is more similar overall to the actual world than world *v* is' and (equivalently) by 'world *w* is closer to the actual world than world *v* is'. The truth-conditions enable us to classify inferences involving counterfactuals as valid or invalid even if we do not know anything about the relation of comparative overall similarity or closeness apart from its structural features. In order to evaluate individual counterfactuals as true or false, however, we need to specify the details of the relation. This section discusses how we should spell out these details. The results will be important for the arguments in later chapters. In particular, the results will enable us to evaluate counterfactuals about the synchronic relation between the mental and the physical as well as claims about the counterfactual dependence of later physical events on earlier events.

The relation of comparative overall similarity or closeness should not have the result that the past counterfactually depends on the present or on the future. The air pressure drops. A little later, the barometer reading falls. Suppose that our similarity relation had the result that if the barometer reading had not fallen, then the air pressure would not have dropped. We would not like to say that the falling of the barometer reading causes the earlier drop in air pressure, and we do not have to say this, for our principle about causation is restricted to putative effects that occur after their putative causes. So far, so good. But suppose further that after the barometer reading falls, there is a storm. If it is true that the air pressure would not have dropped had the barometer reading not fallen, presumably it is also true that the storm would not have occurred had the barometer reading not fallen. Unlike the drop in air pressure, the storm occurs after the barometer reading falls, so it would follow from our principle that the falling of the barometer reading causes the storm. That claim, however, is almost as implausible as the claim that the falling of the barometer reading causes the earlier drop in air pressure.

Let us define a *backtracking* evaluation of a counterfactual as an evaluation that has the result that the past of the time that the antecedent talks about counterfactually depends on what is going on at that time. We just saw that backtracking can have the result that the future depends on the

present or past in the wrong way. This, in turn, would make trouble for our principle about causation. We should therefore rule out backtracking evaluations at least in contexts where we are interested in causation.

Backtracking is well defined only for counterfactuals whose antecedents talk about a specific time, so let us confine our attention to such counterfactuals for now. Here is a simple suggestion that rules out backtracking. For a given counterfactual, let us stipulate that there is a set of antecedent-worlds that are closer to the actual world than any other antecedent-worlds and that have the following feature: the past of the worlds in the set exactly matches the past of the actual world until just before the time specified in the antecedent; then the antecedent is made true with minimal difference to the actual world, which might involve violations of the actual laws of nature; from the antecedent-time onwards, there are no more violations of the actual laws of nature. Let us call this approach to the similarity relation the *asymmetry-by-fiat approach*.[52]

The asymmetry-by-fiat approach avoids backtracking, because the closest antecedent-worlds match the actual world until just before the antecedent-time. Here is how the approach accounts for ordinary cases of counterfactual dependence, such as the dependence of the balloon's bursting on my throwing the dart: The closest antecedent-worlds for the counterfactual 'If I had not thrown the dart, then the balloon would not have burst' are like the actual world until just before the time at which I throw the dart in the actual world. Then my throwing is prevented in a way that minimized the difference to the actual world. Perhaps I have a change of heart, or perhaps a sudden cramp in my arm. Afterwards, the closest antecedent-worlds evolve in accordance with the actual laws of nature. Thus, the balloon remains intact. (Perhaps the balloon eventually bursts at a much later time, but this would be different bursting.)

The asymmetry-by-fiat approach has two disadvantages. First, it entails, at least for the kind of counterfactuals under consideration, that there is a set of closest antecedent-worlds in the sense that the worlds in that set are closer to the actual world than any other antecedent-worlds. We saw in the previous section that counterfactuals like 'If I were nearer to Hammerfest now, I would still be alive' show that there need not always be such a set and that instead there might be a series of antecedent-worlds that are ever closer

[52] I borrow the term from Lewis (1979), who uses it to refer to a slightly different account of counterfactuals that does not assume the truth-conditions in terms of the similarity of possible worlds. Similar approaches can be found in Maudlin 2007 and Paul and Hall 2013: 47–48.

to the actual world.[53] A second disadvantage of the asymmetry-by-fiat approach is that it can be applied only to counterfactuals whose antecedents talk about what is (or is not) going on at specific times.

Neither disadvantage should worry us too much. The closest-worlds issue can be circumvented. Instead of stipulating that there is a set of such-and-such closest antecedent-worlds, we can rule out backtracking by fiat by formulating similarity criteria in the style of the miracles approach that will be discussed shortly.[54] But presumably the closest-world issue is not much of a problem to start with. In any event, for convenience I shall continue sometimes to talk about 'the closest antecedent-worlds' of a given counterfactual. The restriction to antecedents that talk about specific times is not a problem for us either, because the counterfactuals that we shall be concerned with, namely counterfactuals that express counterfactual dependence, have antecedents that talk about specific times in any case. (Likewise for the counterfactuals with more complex antecedents that will play a role in Chapter 3.)

The asymmetry-by-fiat approach gives a simple and convenient way to evaluate the counterfactuals that are relevant to causation. Sometimes, however, it will prove useful to have a more elaborate account of the similarity relation. Lewis suggests the following criteria for the comparative overall similarity between worlds:

> (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.
> (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)

Following Lewis (1979, 471), let us call a big, widespread violation of law a *big miracle* and a small, localized, simple violation of law a *small miracle*. Let us call the way of spelling out comparative overall similarity between worlds by criteria like (1)–(4) the *miracles approach*. It is worth pointing out that the miracles that feature in this approach are not themselves very

[53] The asymmetry-by-fiat approach does not entail the Limit Assumption (see footnote 44), however, since the latter is more general: the Limit Assumption applies to counterfactuals irrespective of whether their antecedents are about specific times.

[54] These criteria would read: it is of the first importance to match the actual world perfectly in particular fact until just before the antecedent-time and not to involve miracles after antecedent-time; it is of the second importance to minimize the violations of law just before the antecedent-time.

miraculous. Nothing like witchcraft is required. If a miracle (big or small) occurs in a given world, *our* laws of nature are violated at that world, not the laws of nature of that very world. (Given that laws at least entail regularities without exceptions, it is impossible for the laws of nature of a given world to be violated in that very world; see Lewis 1979: 468–469.) Lewis's criteria (1)–(4) are most naturally read as exceptionless priorities, such that the criteria higher up on the list always trump the criteria further down.[55] Thus, a world where a big miracle occurs is always less similar overall to our world than a world without big miracles. Among worlds that are on a par with respect to big miracles, a world that matches the actual world perfectly in particular fact throughout a larger region of space–time than another one is always more similar overall to the actual world than the latter. Among worlds that are on a par with respect to all of the above, a world where a small miracle occurs is always less similar overall to our world than a world without a small miracle. Among worlds that are on a par with respect to all of the above, a world that matches the actual world approximately in particular fact throughout a larger region of space–time than another one is always more similar overall to the actual world than the latter (if we opt for the 'little importance' rather than the 'no importance' reading of criterion (4)).

To see how the miracles approach rules out backtracking, let us first assume that our laws of nature are deterministic. Let us assume, that is, that any worlds where our laws of nature hold are either always alike or never alike. (I will say more below about why I make this assumption.) It seems that we never need more than a small miracle to prevent the occurrence of a given event, even if we hold history fixed until just before the time at which the event actually occurred. Take the falling of the barometer reading. We can find a world where a small miracle prevents the falling of the barometer reading, owing to a tiny malfunctioning of the barometer just before its reading actually fell, say. We can also find a world where a small miracle prevents the earlier drop in air pressure. At that world, too, the barometer reading does not fall. The two worlds are on a par with respect to criterion (1), since neither involves a big miracle. The world

[55] We can think of aggregating aspects of similarity to overall similarity as an interpretation of social choice theory: the aspects of similarity play the role of the individual preferences, and overall similarity plays the role of the collective preference. In social choice terms, the similarity aspect of whether or not there is a big miracle is *dictatorial*, because it overrides any other aspects of similarity. While it seems implausible that there are always dictatorial aspects of similarity, such aspects may be tolerated in the special case of causal contexts; see Morreau 2010 and Kroedel and Huber 2013 for further discussion.

where the barometer malfunctions just before the time at which its reading actually fell has more perfect match of particular fact with the actual world than the world where there is no earlier drop in air pressure. By criterion (2), the world with the malfunctioning barometer wins the contest for similarity with the actual world. Since the world with the malfunctioning barometer matches the actual world perfectly in particular fact until just before the time at which the reading actually fell, no backtracking ensues (except perhaps into the very near past).

How does the miracles approach yield the truth of counterfactuals about ordinary cases of counterfactual dependence, such as 'If I had not thrown the dart, then the balloon would not have burst'? Like in the barometer case, we can find worlds whose history is exactly like the actual history until just before the time at which I actually threw and where a small miracle prevents my throw. By criterion (2), these worlds would be even closer to the actual world if they also matched the actual world perfectly after the antecedent-time. According to Lewis, such 'reconvergence' requires a big miracle, however.[56] All the traces of my failure to throw – my memories of not throwing, light rays reflected from my stationary arm, etc. – need to be erased. This requires a multitude of small miracles, which add up to a big miracle. By criterion (1), the worlds with perfect reconvergence that is due to a big miracle are less similar to the actual world than worlds that diverge from the actual world after the original small miracle. One could achieve less than perfect convergence at the cost of a few additional small miracles. In particular, a world could hold fixed the balloon's bursting at the cost of one extra small miracle. By criteria (3) and (4), however, avoiding small miracles is more important than increasing approximate match of particular fact.[57] Thus, there are antecedent-worlds where a single small miracle prevents my throw and where, consequently, the balloon does not burst. By Lewis's criteria, these worlds are closer to the actual world than any

[56] See Lewis 1979. Elga (2001) and Wasserman (2006) argue that sometimes a small miracle suffices to bring about perfect reconvergence.

[57] At least this is the canonical story. It raises some tricky issues, however. First, we cannot read criterion (3) as being an all-or-nothing matter about the occurrence of *some* small miracle, since the number of small miracles matters. This in turn raises questions about how to count small miracles, which is complicated by the fact that, according to Lewis, miracles have a mereological structure similar to that of events. Second, it might seem that a world where the bottle's shattering is brought back by a small miracle has extra *perfect* match of particular fact with the actual world, namely extra perfect match in the spatiotemporal region occupied by the shattering. By criterion (2), this result would jeopardize the desired truth of our counterfactual. We can avoid the result by reading 'spatiotemporal region' in (2) as 'spatiotemporal region that is not scattered along the spatial dimension', but stipulating such a reading raises complications of its own. See Kroedel 2018 for further discussion.

worlds with perfect or imperfect reconvergence where I do not throw and the bottle still shatters. The shattering counterfactually depends on the throw, as desired.

There is some controversy about whether the miracles approach always avoids backtracking. Consider the following case. A nuclear bomb explodes in the centre of our town. The blast first destroys my house; a fraction of a second later, it destroys your house, which is further away from the centre. It might seem that it takes a big miracle to prevent the destruction of my house given that the bomb explodes, but only a small miracle (a tiny malfunction in the fuse, say) to prevent the explosion. Hence, in the closest worlds where my house is not destroyed, the nuclear bomb does not explode and your house is not destroyed either. Consequently, the counterfactuals 'If my house had not been destroyed, then the bomb would not have exploded' and 'If my house had not been destroyed, then your house would not have been destroyed' both come out true.[58] But they should not come out true, of course, if we want to avoid the result that the destruction of my house causes the destruction of your house, which would follow by our principle about causation.

There is a bold and a modest response. The bold response denies that the counterfactuals are true and claims that it merely takes a small miracle to prevent the destruction of my house. Thus, the closest worlds where my house is not destroyed match the actual world until just before the time at which my house is destroyed in the actual world. In particular, in these worlds the nuclear bomb still explodes and, later, your house is still destroyed. Why should we say that the miracle that prevents the destruction of my house (given that the nuclear explosion occurs) is a small one? According to Lewis (1986e: 55–56), what distinguishes big miracles from small ones is that big miracles are spread out more broadly and have parts (themselves small miracles) that are varied.[59] Given that the nuclear explosion occurs, certain laws need to be broken throughout a spatial volume around my house for a short while in order to shield my house from the blast. While admittedly this volume has a substantial size, it is not spread

[58] Kment (2010: 84, 107 n. 10) credits this example to Peter Lipton. Woodward (2003: 133–145) raises a similar issue. See also Jonathan Bennett 2003: 204–211. (All future references of the form 'Bennett 2003' are to Karen Bennett's 2003.)

[59] Lewis writes that '[a] big miracle consists of many little miracles together, *preferably* not all alike' (1986e: 56, my emphasis). The context of his discussion strongly suggests that the 'preferably' qualification can be dropped, however, for the following sentence states that '[w]hat makes the big miracle more of a miracle is . . . that it is divisible into many *and varied* parts, any one of which is on a par with the little miracle' (1986e: 56, my emphasis; see also Lewis 1979: 471).

out in the sense of being scattered.[60] Further, the parts of this miracle are all alike, because they all involve violations of the same laws (namely whichever laws need to be broken in order to prevent the radiation, heat, and impact of the explosion from reaching my house).

The modest response is to tolerate cases of backtracking that allegedly result from the miracles approach and to weaken our principle about causation accordingly. Instead of claiming that counterfactual dependence as assessed according to the miracles approach is sufficient for (forward-in-time) causation, we can claim that counterfactual dependence as assessed according to the miracles approach is sufficient for causation *if* the relevant counterfactual is not evaluated in a backtracking way. Perhaps in the case of the nuclear explosion the truth of 'If my house had not been destroyed, then your house would not have been destroyed' is due to backtracking according to the miracles approach. If so, no causal consequences follow. By contrast, the truth of 'If I had not thrown the dart, then the balloon would not have burst' is not due to backtracking according to the miracles approach. Hence we may infer that my throw causes the balloon to burst. We shall see that in cases of mental causation the miracles approach will not yield backtracking either, so the modest response will be available in those cases too.

As it stands, the miracles approach requires determinism to yield the right verdicts about counterfactual dependence. (By determinism, I mean the claim that the laws of a given world are deterministic in the sense defined earlier in this section.) Without determinism, perfect reconvergence would not require another miracle (big or small), because it could occur by mere chance (see Lewis 1986e: 60). Similarly, imperfect reconvergence that merely guarantees the falsity of the consequent at the world in question could come about by chance. Thus, most counterfactuals about ordinary cases of counterfactual dependence would come out false (see Hawthorne 2005, Hájek ms.). The same holds for the asymmetry-by-fiat approach.[61] If we let an antecedent-world evolve lawfully and the laws are indeterministic, the consequent of the counterfactual may come out false by chance.

I will set aside the question of whether the falsity of ordinary counterfactuals under indeterminism would be a serious problem *per se*. We would not get counterexamples to our principle about causation, because the principle says that counterfactual dependence is sufficient for causation, not that it is necessary. While the principle would lose much of its utility if

[60] By contrast, Lewis's paradigmatic big miracles, the reconvergence miracles, are spread out in the sense of being scattered (1979: 471).

[61] *Pace* Lewis, who holds that the approach 'has no need of determinism' (1986e: 62).

there were few cases of counterfactual dependence under indeterminism, there is a similar principle that we could use as a substitute. The similar principle says that what we may call *probabilistic dependence* is sufficient for causation: event *c* causes a later event *e* just in case *e* would have had a much lower chance of occurring had *c* not occurred.[62] It seems that both the asymmetry-by-fiat approach and the miracles approach would evaluate enough counterfactuals that express probabilistic dependence as true to justify using a sufficient condition for causation in terms of counterfactuals.[63] Thus, in the end not much hinges on whether or not our laws of nature are deterministic. It is mainly for simplicity that, for the remainder of this book, I shall assume that they are.

The upshot so far is that the asymmetry-by-fiat approach is simple and straightforwardly rules out backtracking. The miracles approach is less simple (albeit more generally applicable), and it is less clear that it rules out backtracking. Owing to these advantages of the asymmetry-by-fiat approach over the miracles approach, I shall – at least *ceteris paribus* – prefer the asymmetry-by-fiat approach in later arguments. When we consider mental causation under dualism, however, the miracles approach will turn out to be superior, because it lends itself to making a modal distinction between different kinds of laws (in particular, between physical laws and psychophysical laws).

Backtracking is not the only threat to the sufficiency of counterfactual dependence for causation. Another one comes from strange causal relata. Here is an example. At midnight, a bottle shatters owing to sudden external forces. At 11:59 p.m., the bottle had the property of shattering-in-one-minute (call this property *S*+). If the bottle had not had *S*+ at 11:59 p.m., it would not have shattered at midnight. But it does not seem that the bottle's having *S*+ at 11:59 p.m. causes the bottle's shattering at midnight.

The appropriate response is to rule out events that involve certain kinds of properties. We should allow only properties that are sufficiently intrinsic and temporally intrinsic, that is, roughly, properties that are about how things are with the object in question itself and about how things are with that object at the time of instantiation.[64] Property *S*+ is not about how things are with the object at the time of instantiation, but rather about how things will be with the object a minute after the time of instantiation. We can spell out the envisaged response in two ways. We can say that only

[62] See Lewis 1986d. For further discussion, see Hitchcock 2004a.
[63] The miracles approach requires some modifications for the indeterministic case; see Lewis 1986e.
[64] See Lewis 2004: 78. Cases where merely the requirement of intrinsicness is violated, not the requirement of temporal intrinsicness, are discussed in Lewis 1986b: 262–266 and 1986d: 189–193.

instantiations of properties that are sufficiently (temporally) intrinsic are genuine events. Or we can say that only counterfactual dependence between events that involve properties that are sufficiently (temporally) intrinsic suffices for causation. I have no firm opinion about how (temporally) intrinsic the properties in genuine events have to be and will therefore embrace the second option.[65]

It might seem that instances of properties that are highly temporally extrinsic still sometimes qualify as causes. Suppose that, at noon, a celebrity has the property of dying at midnight from a prolonged illness. Call the property of dying-at-midnight *D*+. Property *D*+ is highly temporally extrinsic (at least when instantiated at a time other than midnight). Having learned about the impending death, a journalist writes an obituary in the afternoon. Does the celebrity's having *D*+ cause the journalist's writing? It might seem so at first sight, but a better diagnosis of the case is that the journalist's writing and the celebrity's death have a common cause, namely the celebrity's medical condition before noon. This is consistent with citing the fact that the celebrity has *D*+ – more idiomatically, the fact the celebrity is going to die – as a reason for the journalist's writing the obituary.[66] For the fact can be a reason for the journalist to write by virtue of being the content of a belief of the journalist, which in turn is a cause of the writing; the fact need not itself be such a cause.

Lastly, alleged causes that seem far-fetched pose a threat to the sufficiency of counterfactual dependence for causation. On the street I bump into a stranger, Albert, who subsequently misses his bus. On the next bus, Albert meets his future wife. They have a child, Berta, who dies 90 years later. If I had not bumped into Albert, then Berta

[65] A related worry is about omissions. Lots of events counterfactually depend on omissions, but one might not want to accept that they are caused by those omissions. That omissions cannot be causes is more controversial than that properties like *S*+ cannot be causes. If one wants to rule out omissions as causes, one could pursue a similar strategy: restrict our sufficient condition for causation to instances of 'positive' properties and disallow instances of 'negative' properties. I will remain neutral on questions about omissions as causes, but in Section 3.4 I will discuss ways of dealing with omission in the causal modelling framework. For further discussion, see Lewis 1986d: 189–193 and McGrath 2005. Should we also demand that only instances of *natural* properties in Lewis's (1983) sense can be causes and effects? We had better not without good reason. Arbitrary disjunctions of intrinsic properties are still intrinsic (see Weatherson 2001) – not so for natural properties. Given non-reductive physicalism, mental properties turn out to be – or at least to be strictly equivalent to – long disjunctions of physical properties (see Section 2.2). Perhaps this makes mental properties somewhat unnatural, but we would still like to maintain that their instances can be causes and effects. For further discussion of disjunctive causes, see Sartorio 2006 and Beebee 2017.

[66] On the related issue of corresponding 'because'-sentences, see Jenkins and Nolan 2008 and Schnieder 2015.

would not have died.[67] But it might seem that my bumping into Albert does not cause Berta's death.

The appropriate response is to accept that my bumping into Albert does cause Berta's death and to explain away appearances to the contrary as a pragmatic phenomenon. Counterfactual dependence is sufficient for a kind of causation that is 'broad and non-discriminatory' (Lewis 1973a: 559). If event *e* counterfactually depends on an earlier event *c*, it follows that *c* is *a* cause of *e*. It does not follow that *c* is among those causes of *e* that are explanatorily relevant, and hence worth mentioning, in any given context. (*A fortiori*, it does not follow that *c* is *the* cause of *e*, if the definite article is supposed to single out the most explanatorily relevant one among *e*'s causes.) In most contexts, my bumping into Albert counts as irrelevant for a causal explanation of Berta's death. In those contexts, it would sound strange to say that my bumping into Albert causes Berta's death, but it remains true that it is among the causes of Berta's death.[68]

For completeness I should mention another type of case that is often cited in objections to Lewis's theory of causation. I write 'Larry'. By writing 'Larry', I *ipso facto* write 'rr'. Thus, if I had not written 'rr' I would not – indeed, could not – have written 'Larry'. But it sounds strange to say that my writing 'rr' causes me to write 'Larry' (see Kim 1973). This result does not follow from our principle about causation, however. The principle is restricted to cases where the putative cause occurs before the putative effect.[69] My writing 'rr' does not, however, occur before my writing 'Larry' (although it ends earlier). More generally, if two events occur at times that do not overlap and they involve only properties that are temporally intrinsic, we never get the kind of necessary connection between those events that give rise to 'Larry'-style counterexamples.[70]

[67] That is, the event of Berta's death would not have occurred, because Berta would not have existed in the first place. The example is a variation of an example from Lewis 1986d: 184. Thomson (2003) takes cases of this kind to refute the sufficiency of counterfactual dependence for causation.

[68] See Lewis 2004: 101. For a recent elaboration of this approach, see Swanson 2010. According to Sartorio 2010, there are further cases, which indirectly involve omissions, that are counterexamples to the sufficiency of counterfactual dependence for causation. Weslake 2013 argues that they can be defused by a strategy that is similar to the one used here.

[69] Thus, the principle remains neutral on whether there can be simultaneous causation. See Section 3.6 for further discussion.

[70] Lewis's own response is to restrict causal claims to cases where cause and effect are wholly distinct events, that is, events that occur in non-overlapping spatiotemporal regions (see Lewis 1986b: 259 for discussion). It is not entirely clear that cause and effect can never overlap. Perhaps the First World War caused a famine that started before the war ended. For our purposes we need not settle this issue, however.

1.6 Transference Views, Double Prevention, and Powers

A rival view of causation requires that a certain physical quantity be transferred from cause to effect. This view can be spelled out in different ways. One might remain neutral on what physical quantity is transferred (see Aronson 1971) or allow only quantities that obey a physical conservation law (see Dowe 2000, Salmon 1994), especially energy (see Fair 1979). In typical cases of causation, such a transfer indeed takes place. A thrown rock transfers energy to a bottle, thereby shattering it. A moving billiard ball transfers momentum to another ball, thereby making it roll into the pocket. A lightning strike transfers energy to a house, thereby igniting it. Transfer accounts also give the right verdicts (or at least no wrong verdicts) about cases that full-blown accounts of causation in terms of counterfactuals have found hard to cope with, namely cases of late pre-emption and overdetermination. Only Billy's throw, which actually hits the bottle, transfers energy to it; Suzy's throw, which actually arrives at the bottle's place only after it is destroyed by Billy's throw, does not. And all bullets from the firing squad transfer energy to the victim.

The cases we have just considered make it *prima facie* plausible that transfer of a physical quantity is necessary for causation. The converse claim, that transfer of a physical quantity is sufficient for causation, does not seem plausible. Let us modify the example of the rock that shatters the bottle a bit. Suppose that the rock is first heated over a fire and then thrown at the bottle, which shatters. The fire transfers a physical quantity (namely heat) to the bottle via the rock, but does not seem to qualify as a cause of the shattering in any interesting sense.[71] Since friends of explaining causation in terms of transfer are well advised not to endorse the claim that transfer of a physical quantity is sufficient for causation, by 'transference views' I shall merely mean views according to which the transfer of a physical quantity is necessary for causation.

There is a family of cases that make trouble for transference views of causation. A pillar is propped up on a rack. I kick the rack aside, and the pillar falls down.[72] A catch holds a stretched spring in position. I release the catch, and the spring accelerates (see Aronson 1971: 425). There are many more cases of this kind (see Schaffer 2000a, 2004a). They all have the following structure: something happens that would have been prevented by something else, which is itself prevented. (Such cases have become

[71] For a similar example, see Hitchcock 1995: 316. One can easily find examples of this kind where the transferred physical quantity that is seemingly irrelevant is a *conserved* quantity, such as charge.

[72] See Paul and Hall 2013: 191 for a similar example.

known as cases of *double prevention*.) By kicking the rack aside, I prevent it from preventing the pillar from falling. By releasing the catch, I prevent it from preventing the spring from accelerating.

On the face of it, my kicking the rack aside causes the pillar to fall down, and my releasing the catch causes the spring to accelerate. Our counterfactual principle about causation says so too. If I had not kicked the rack aside, the pillar would not have fallen down. If I had not released the catch, the spring would not have accelerated. By our principle, the kicking causes the falling, and the releasing causes the accelerating. (The condition that the putative effect occurs after the putative cause is satisfied here too. For simplicity, I suppress reference to the events' time order here and in what follows.) This is bad news for transference views of causation. For in cases of double prevention like the ones we have considered, no transfer takes place between what seems to be the cause and what seems to be the effect. In kicking away the rack, I do not transfer anything to the pillar. In releasing the catch, I do not transfer anything to the spring.[73] (Of course, energy was transferred on the pillar when it was propped up, and energy was transferred to the spring when it was stretched, but this is not the issue here.) Thus, our counterfactual principle about causation and transference views are in conflict, and it seems that, at least with respect to the double-prevention cases we have considered so far, the counterfactual principle has the upper hand.

We can use so-called neuron diagrams to illustrate the structure of double-prevention cases.[74] The conventions for these diagrams are as follows. Circles represent neurons; specifically, shaded circles represent neurons that fire, and non-shaded circles represent neurons that do not fire. Arrows represent excitatory connections, lines with dots inhibitory connections. A neuron that has incoming connections fires just in case it is excited by other neurons and it is not inhibited by any other neuron.[75] In the case depicted in Figure 1.1, the relation of the firing of neuron *c* to the firing of neuron *e* is that of double prevention. The firing of *c* prevents the firing of *d*, which, had it not been prevented, would have prevented the firing of *e*. Figure 1.2 shows what would have happened if *c* had not fired. In this case, *e* would not have fired either, because it would have been

[73] Perhaps in some possible realizations of these cases, I also cause the pillar to fall to the side by transferring momentum to it or cause the spring to heat up by transferring body heat from my hand to it via the catch. We can, however, easily imagine realizations that are sufficiently idealized so that no such transfers take place.

[74] See Schaffer 2000a, 2004a, and Paul and Hall 2013: 175. My neuron diagrams differ slightly in structure from those of Schaffer and Paul and Hall, but in inessential ways.

[75] For a critical discussion of the use of neuron diagrams in the philosophy of causation, see Hitchcock 2007b.

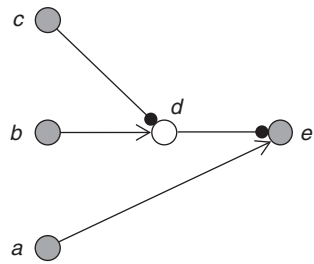

Figure 1.1. A neuron case of double prevention

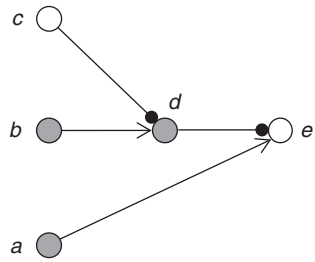

Figure 1.2. If *c* had not fired . . .

prevented by the firing of *d*. On the face of it, the firing of *c* causes the firing of *e*. Again, our counterfactual principle agrees, because the firing of *e* counterfactually depends on the firing of *c*. But again, there is no transference from the firing of *c* to the firing of *e* (if we assume that no relevant quantity passes through a neuron that does not fire), so we seem to have another counterexample to the transference view.

We can use neuron diagrams to illustrate not just causal relations between the firings of (idealized) neurons, but also causal relations between ordinary events. Thus, we can map the other cases of double prevention onto the neuron structure depicted in Figures 1.1 and 1.2. Figure 1.3 shows a neuron representation of the pillar example.[76] What used to represent the firing of neuron *c* now represents my kicking the rack aside, and what used to represent the firing of neuron *e* now represents the falling down of the pillar. The firing of neuron *b* is now the event of the rack's being in place at a time before my kicking, while the firing of neuron *d* is now the event of the rack's being in place at a time just after my kicking. What used to be the

[76] For similar representations, see Schaffer 2000a and 2004a.

https://doi.org/10.1017/9781108762717.003 Published online by Cambridge University Press

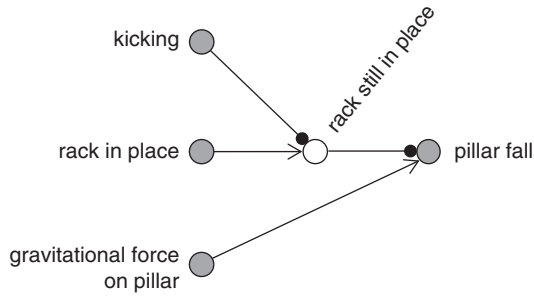

Figure 1.3. A real-life case of double prevention

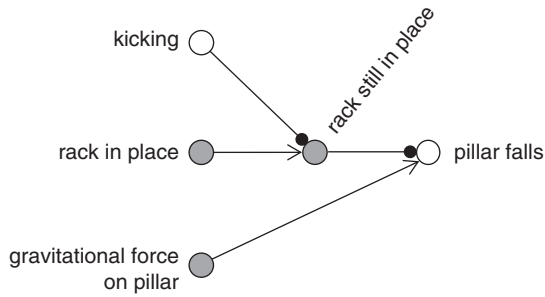

Figure 1.4. If I had not kicked away the rack . . .

firing of neuron *a* is now a (non-double-preventive) cause of the pillar's falling down, such as the presence of a gravitational force that acts on the pillar. Figure 1.4 shows what would have happened if I had not kicked away the rack. In this case, the rack would have remained in place, which would have prevented the pillar from accelerating and thus from falling down.

Our double-prevention cases involve omissions as intermediaries. My kicking is followed by the absence of the rack underneath the pillar, which is followed by the falling of the pillar. My releasing of the catch is followed by its absence from its original position, which is followed by the acceleration of the spring. The firing of neuron *c* is followed by the non-firing of neuron *d*. One might be sceptical about whether omissions can be causes or effects. But this scepticism does not make the counterexamples to transference views go away, for the counterexamples are cases where one genuine event (my kicking the rack aside / my releasing the catch) causes another genuine event (the falling of the pillar / the acceleration of the spring) (see Paul and Hall 2013: 190).

Admittedly, conflict between our counterfactual principle about causation and transference views also arises in cases of double prevention where it is more controversial how to resolve it. I shoot down an interceptor that would otherwise have shot down a bomber. The bomber destroys the target. As in the rack and spring cases, I prevent something from happening (namely the shooting down of the bomber) which in turn would have prevented something else from happening (the destruction of the target). Unlike in the other double-prevention cases, however, the different events in the story are not even continuous in space–time. Perhaps the bomber crew knew nothing about the threat to their mission, because I shot down the interceptor hundreds of miles from the bomber's course (see Hall 2004b). Our counterfactual principle says that my shooting down the interceptor causes the destruction of the target, because the target would not have been destroyed if I had not shot down the interceptor. Transference views say that my shooting down the interceptor does not cause the destruction of the target, because I do not transfer anything to it. It might seem more plausible to side with the transference views here because there is no spatiotemporal continuity between my actions and the destruction of the target.[77]

It is one peculiar feature of double-prevention cases that they allow for spatiotemporal discontinuity. Another peculiar feature is that, if they involve causation between the double preventer (that is, the event that prevents the prevention, such as my kicking away the rack) and the event whose prevention is prevented (such as the pillar's falling down), they also show that causation is not a matter of the intrinsic connection between events. The idea that causation is an intrinsic matter can be spelled out as follows. Take a case where, in the actual world, event *c* causes event *e*. Take all the events that, in the actual world, cause *e* and that occur from a certain time before *c* onwards. (These events of course include *c*.) Embed these events in an arbitrary nomologically possible situation. According to the intrinsicness idea, in the new situation *c* still causes *e* (see Paul and Hall 2013: 196–197). We can easily construct a counterexample to the intrinsicness thesis on the basis of our original neuron example if we assume, at least for the sake of the argument, that the firing of neuron *c* causes the firing of neuron *e*. In the example, the firing of neuron *b* does not seem to be among the causes of the firing of *e*. Thus, if the intrinsicness thesis holds,

[77] Hall's own diagnosis is that we should distinguish two concepts of causation and that counterfactual dependence is sufficient for only one of them, which applies in cases of double prevention (2004b). Won (2014: 215) holds that in double-prevention cases the event that prevents the preventer is not a cause at all.

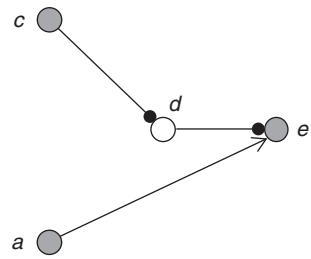

Figure 1.5. An intrinsic duplicate of the neuron case

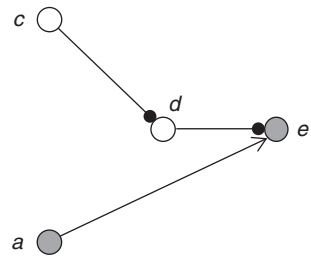

Figure 1.6. The intrinsic duplicate if *c* had not fired

duplicating our neuron structure while omitting neuron *b* should leave all the causal relations intact. Figure 1.5 shows this new structure.

Contrary to what the intrinsicness thesis predicts, in the new structure the firing of *c* does not cause the firing of *e* (Paul and Hall 2013: 196–197). This seems clear without assuming any particular theory of causation. In particular, friends of the counterfactual approach will reach this verdict. The firing of *e* no longer counterfactually depends on the firing of *c*. (Figure 1.6 shows the case where *c* does not fire.) This lack of counterfactual dependence does not by itself entail that there is no causation, because counterfactual dependence was never assumed to be necessary for causation. But, unlike in cases of pre-emption and overdetermination, where we also have causation without counterfactual dependence, it is unclear how the firing of *c* should otherwise cause the firing of *e* in the new situation where neuron *b* is absent.

We have assumed that, in the original neuron example, the firing of *c* causes the firing of *e*. In the intrinsic duplicate of the example shown in Figure 1.5, the firing of *c* does not cause the firing of *e*. Thus, if double prevention involves causation between the double preventer and the event whose prevention is prevented, then causation cannot always be a matter of

the intrinsic connection between cause and effect. More specifically, if double prevention involves causation between the two events in question that is due to counterfactual dependence, then causation that is due to counterfactual dependence cannot always be a matter of the intrinsic connection between cause and effect.

The upshot so far is this. According to our counterfactual principle about causation, cases of double prevention are cases of causation. According to transference views, they are not. On the face of it, certain cases of double prevention, such as the pillar, spring, and neuron examples, seem to be cases of causation. For other cases, such as the bomber example, it is less plausible that they are cases of causation, because there is no spatiotemporal continuity between the events in question.[78] If one thinks that causation is an intrinsic matter, one also has reason to deny that cases of double prevention are cases of causation.

What should we make of this situation? I think we can still make a very strong case for the claim that double preventers are causes, which is *ipso facto* a very strong case against transference views of causation.

First, the worries about discontinuity and failures of intrinsicness can be explained away or at least attenuated. On reflection, it does not seem so implausible that my shooting down the interceptor causes the destruction of the target. Recall that we are reading 'causes' in the sense of being *a* cause of the effect, not in the sense of being a cause that has a particular explanatory relevance for the effect or that even is *the* cause of the effect. If we had to list the causes of the target's destruction in the order of their explanatory relevance, the bombing of the target would come first. But it does not follow that my shooting down of the interceptor is not to be found on the list at all. Perhaps I get a medal for my role in the successful destruction of the target, or perhaps I get blamed for it. How could this be justified if my action is not a cause of the target's destruction? According to a standard assumption, causation is necessary (though not sufficient) for moral responsibility, so I could rightly be praised or blamed for my role in the destruction only if my action caused it.[79] That there is no spatiotemporal continuity between cause and effect in our example may be unusual, but it is hard to see why it should make causation impossible.[80] The result that causation is not a matter of the intrinsic connection between cause and

[78] Examples like the bomber case can be multiplied. See, for instance, Lewis 2004: 83–84.

[79] On the standard assumption, see Sartorio 2007. Schaffer (2000a) also uses considerations of moral responsibility in support of the claim that double prevention involves causation.

[80] Arguably, cases of 'action at a distance' also arise in cases of quantum entanglement; see Fenton-Glynn and Kroedel 2015 for discussion.

effect does not have to be regarded as especially problematic, either. Few people, I take it, have a strong intuition that intrinsicness is a non-negotiable feature of causation, especially a strong intuition that is not parasitic on prior beliefs in transference views about causation.[81]

Second, as Schaffer points out, cases of double prevention display various features that are typically associated with causation (without having to be present in all cases of causation). For instance, knowing that the double preventer occurs licenses the prediction that the event that would have been prevented if the double preventer had not occurred will occur too, and bringing about the occurrence of the double preventer is an effective strategy for bringing about the event that would have been prevented.[82] Knowing that someone kicked away the rack, say, licenses the prediction that the pillar is going to fall, and kicking away the rack is an effective strategy for making the pillar fall.

Third, there are cases that seem to be among the most paradigmatic cases of causation and yet involve double prevention. The actions of modern firearms work much like our example of the spring (except that in firearms the spring starts out compressed rather than stretched out) (see Schaffer 2000a). In a cocked gun, the sear holds back the coiled spring. When the trigger is pulled, the sear is removed. No longer held back by the sear, the spring uncoils and causes the hammer to hit the cartridge, which causes the propellant to explode, which in turn causes the acceleration of the bullet. The pulling of the trigger is related to the acceleration of the bullet by double prevention. The pulling of the trigger prevents the sear's holding back of the spring, which, if it had not been prevented, would have prevented the spring from uncoiling and thus would have prevented the bullet from accelerating. If someone kills someone else by pulling the trigger of a gun, the relation between the pulling of the trigger and the death of the victim seems as causal as it ever gets. Our counterfactual principle agrees, because the victim would not have died if the trigger had not been pulled (assuming that there are no further redundant causes). Transference views disagree, because the pulling of the trigger does not transfer anything to the victim owing to the double-prevention structure of the case.

In sum, it seems much more plausible that some double-prevention cases are cases of causation than that some double-prevention cases are not cases of

[81] Lewis finds it intuitive that causation is a matter of an intrinsic relation between cause and effect, but adds that intuitions about what is intrinsic should be mistrusted (1986d: 205). For further discussion, see Schaffer 2000a, Hall 2004a, Hawthorne 2004, Lewis 2004, and Weatherson 2007.

[82] See Schaffer 2000a: 285. Lombrozo (2010) and Woodward (2012, 2014) discuss a feature that is present in some, but not all, cases of double prevention, viz. the feature they dub 'stability'.

causation. If we treat double-prevention cases uniformly, we should therefore accept that all of them are cases of causation. Unlike our counterfactual principle about causation, transference views reach the opposite verdict. We should therefore reject transference views of causation.

The argument from double prevention also applies to certain powers theories of causation. Explaining causation in terms of powers or dispositions has become increasingly popular. One of the most detailed and influential powers theories of causation is due to Stephen Mumford and Rani Lill Anjum (2011).[83] According to their theory, there is causation when powers exercise themselves (2011: 6). More specifically, they hold that causation is the passing around of powers. For example, when the heat of a fire causes my body to warm up, it passes the power of warming things up to by body (see Mumford and Anjum 2009: 283). It does not always have to be the same power that is passed on, however. When a fragile glass is dropped and breaks, the pieces of glass have powers they did not have before, such as the power to cut (see Mumford and Anjum 2009: 284; 2011: 6–7).

According to Mumford and Anjum's powers theory, double prevention is not causation:

> Double prevention concerns the non-exercise of powers: twice over. A power is prevented from exercising when another also fails to exercise. We have, therefore, two failures of causation. Just as two wrongs do not make a right, two failures of causation do not make a cause. (Mumford and Anjum 2009: 287)

Applied to one of our examples, we can presumably locate the two failures of causation that Mumford and Anjum diagnose as follows: in the neuron example, the power of neuron *b* to simulate neuron *d* is prevented from exercising by the firing of neuron *c*, and neuron *d* fails to exercise its power to inhibit neuron *e*.

Why is there no exercise of powers in cases of double prevention? Why not say that, in the neuron example, it is still the case that neuron *c* exercises a causal power vis-à-vis neuron *e*, even though neuron *b* fails to exercise its power to stimulate neuron *d*, and *d* fails to exercise its power to inhibit neuron *e*? The deeper reason, according to Mumford and Anjum, for why *c* fails to exercise a causal power vis-à-vis *e* is that no power is passed from *c* to *e*. We can think of this passing as a kind of transference. Indeed, Mumford and Anjum hold that their theory is a transference theory, where

[83] Earlier powers theories are presented in Harré and Madden 1975, Bhaskar 1975, Cartwright 1989, Ellis 2001, Heil 2003, Molnar 2003, and Martin 2008.

what is transferred is powers and not necessarily conserved quantities (2011: 102). No transfer of causal powers takes place between a double preventer and the event whose prevention is prevented. For instance, no causal power is transferred from the firing of neuron *c* to the firing of neuron *e*. So, by Mumford and Anjum's theory, the firing of *c* does not cause the firing of *e*. Likewise for other cases of double prevention.[84]

Let us call theories of causation according to which causation requires transfer of causal powers *powers transference views*.[85] For the reasons just given, powers transference theories, just like transference views that talk about physical quantities, are committed to the claim that double prevention is not causation. Mumford and Anjum think it is a welcome result that their theory denies that double-prevention cases are cases of causation. They focus on cases like the bomber example where this denial is *prima facie* plausible. But, if we take all the considerations into account, the result is just as bad as it is for transference views in general. In particular, denying that there is causation in the bomber example comes at the cost of denying that one can cause people to die by pulling the trigger of a gun. This is too high a price to pay.[86]

In this section we have investigated transference and powers transference views of causation and discussed how they fare vis-à-vis cases of double prevention. We have seen that, all things considered, a strong case can be made for the claim that cases of double prevention are cases of causation. Transference and powers transference views of causation cannot accommodate this claim; our counterfactual principle can. This strongly speaks against transference and powers transference views, while speaking in favour of the counterfactual principle.

Proponents of transference or powers transference views might not be convinced by the argument from double prevention. Indeed, it would be surprising if they were, for philosophical debate tends to end in deadlock

[84] Mumford and Anjum do not quite make the point about the relation between the passing of powers on the one hand and double prevention on the other in these general terms, but they come close to it when they discuss double-prevention cases that involve spatiotemporal discontinuity. In this context, they write that '[c]ause and effect are to be understood as power and manifestation where one merges into another in a continuous process' (Mumford and Anjum 2009: 287).

[85] For continuity with the definition of transference views, I define powers transference views such that they merely claim that transfer of power is necessary for causation. Thus, Mumford and Anjum's view that causation *is* the transfer of powers is stronger than a mere powers transference view as defined here.

[86] Hüttemann (2013) defends a version of the powers theory of causation according to which double preventers *are* causes and thinks – rightly, by our lights – that this is a virtue of his theory. Vetter (2015) develops a theory of powers ('potentialities' in her terminology) that she claims is able to support both a Mumford-and-Anjum-style and a Hüttemann-style powers theory of causation (98–100).

rather than conversion.[87] They can, however, still read the remainder of this book as showing how far one can get in solving the problems of mental causation if one adopts our counterfactual principle about causation rather than a transference or powers transference view and as laying out the challenge for the competitor views. As we shall see, the troubles for transference and powers transference views are far from over. Difficulties, including difficulties from double prevention, will reappear in the context of mental causation.

1.7 Conclusion

This chapter has introduced different theories about the nature of mind, in particular reductive physicalism, non-reductive physicalism, and dualism, including naturalistic dualism. It has argued that, in the context of mental causation, causal relata are best conceived of as particular events. It has also argued that, in that context, the best account of particular events is the strong Kimian account, according to which events are constituted by triples of an object, a property, and a time, and according to which events have their identity (including their trans-world identity) determined by the object, the property, and the time. The chapter has introduced the truth-conditions for counterfactual conditionals and some of the logical peculiarities of these conditionals. It has defended a principle that states a sufficient condition for causation in terms of counterfactual dependence: an event causes a later event if the second event would not have occurred had the first event not occurred. In order for this principle to defy some *prima facie* problems, certain assumptions need to be made about how the relevant counterfactuals are evaluated. In particular, backtracking readings of those counterfactuals must be ruled out. This can be done either by following the asymmetry-by-fiat approach or by following the miracles approach. The principle also needs to be restricted to instances of properties that are sufficiently intrinsic and temporally intrinsic. In a different application of the notion of intrinsicness, cases of double prevention showed causation by counterfactual dependence not to be a matter of an intrinsic connection between cause and effect. These cases also showed the principle about causation in terms of counterfactual dependence to be in conflict with transference views and powers transference views of causation: according to the principle, cases of double prevention are cases of causation; according to transference and powers transference views, they are not. A strong case can be made for resolving the conflict in favour of the principle in terms of counterfactuals.

[87] At least that is what Lewis thought: see Lewis 2000: 102.