CHAPTER 3

Mental Causation by Causal Modelling

3.1 Introduction

The world of modelling is glamorous. This holds even within philosophy, where for the past two decades or so causal modelling has been one of the most successful approaches to the study of causation. In particular, causal modelling theories of token causation[1] (that is, causation between token events) have been able to solve numerous problems that had plagued earlier theories.

Causal modelling theories of token causation can be regarded as descendants of earlier counterfactual theories such as Lewis's (1973a). Their central tools are causal models, that is (roughly), structures that represent events and counterfactuals about these events. The advantage of causal modelling theories is that they can represent more complex counterfactual structures than those earlier theories could. The earlier theories were limited to counterfactuals whose antecedents and consequents talked about the occurrence or non-occurrence (typically the non-occurrence) of individual events, such as 'If I had not thrown the dart, the balloon would not have burst.'[2] Causal modelling theorists, by contrast, can invoke counterfactuals about what would have been the case if a certain event had or had not occurred while such-and-such other events had or had not occurred.

To illustrate the approach of taking into account more complex antecedents, consider a case of so-called early pre-emption.[3] An apprentice

[1] Among such theories that have been advanced following Pearl 2000 and Spirtes *et al.* 2000 are Hitchcock 2001, Woodward 2003, Halpern and Pearl 2005, Hall 2007, Hitchcock 2007a, Halpern and Hitchcock 2010, and Halpern and Hitchcock 2015.

[2] The earlier theories could achieve more complexity by formulating conditions on causation in terms of chains of such counterfactuals (see Lewis 1973a; compare also Section 2.4).

[3] The following is a simplified version of a case from Hitchcock 2004b. The difference from late pre-emption that was introduced in Chapter 1 is that in early pre-emption the backup cause is pre-empted before the chain from the actual cause to the effect is completed.

assassin, called Apprentice, shoots the victim, called Victim, who dies. There was a risk that Apprentice would lose his nerve and fail to fire. To complete the assassination in that case, an expert assassin, called Expert, was present as a back-up: if Apprentice had not fired, Expert would have fired instead. We would like to say that the firing of Apprentice causes Victim to die, but Victim's death does not counterfactually depend on Apprentice's firing, for if Apprentice had not fired, Expert would have fired instead, in which case Victim would have died anyway. If Apprentice had not fired *and Expert had not fired either*, however, then Victim would not have died. Causal modelling theorists can capitalize on this insight. They can claim that what matters for causation is that one event depends on another if we hold certain other facts fixed (in our example, the fact that Expert did not fire). This seems to be an intuitive diagnosis of why Apprentice's firing causes Victim to die. The diagnosis can in turn be formulated neatly in terms of causal models.[4]

On the face of it, causal modelling theories of causation, and the counterfactuals with complex antecedents that they employ, are ideally suited for being applied to (putative) cases of mental causation, where not only a mental event and its (putative) effect are in play, but also the physical realizer of the mental event and earlier physical goings-on. It is thus somewhat surprising that the application to mental causation has only been investigated for one causal modelling theory in the literature so far, namely the interventionist theory by James Woodward (2003). Whether interventionism can accommodate mental causation has turned out to be controversial, so there is reason to be suspicious about the ability of other causal modelling theories to do so.[5]

This chapter applies causal modelling theories to the case of putative mental causes and argues that these theories can, after all, explain mental causation. We shall see that this holds for other causal modelling theories as well as for interventionism. These different causal modelling theories can explain mental causation, I will argue, although this requires some unorthodoxy in how the relevant causal models are built. The chapter does not merely provide a vindication of mental causation by causal modelling, however. It also uses causal modelling to provide a solution

[4] Lewis's theory offers a diagnosis too, by claiming that there is chain of counterfactual dependence from the apprentice assassin's firing to the victim's death, but this diagnosis seems less intuitive.

[5] Gebharter (2017) investigates the applicability to mental causation both for certain probabilistic causal modelling theories and for Woodward's interventionism. His results are negative in both cases, however, which strengthens the suspicion that causal modelling theories in general cannot accommodate mental causation.

to the problem of overlapping realizers from Section 2.3. Recall that the problem was that our principle about causation in terms of counterfactuals commits us to the claim that an aluminium ladder's opacity causes my electrocution when I hold it against a power line, because the realizers of opacity overlap with the realizers of conductivity. It will turn out that conditions on causation in terms of causal modelling that avoid the commitment to the ladder's opacity being a cause are problematic. But causal modelling will allow us to formulate a criterion for explanatory relevance that can solve the problem.

One can be more or less ambitious in using causal modelling, trying to find necessary and sufficient or merely sufficient conditions for causation. As in the previous chapter, in this chapter I will consider only sufficient conditions for causation, not necessary and sufficient conditions. Another respect in which one can be ambitious in causal modelling is to aim not merely at formulating conditions on causation in terms of causal models, but also at providing new foundations for counterfactuals in terms of causal models.[6] In this respect, too, I choose the modest side, because this will allow us to continue working with the account of counterfactuals that was established in Chapter 1.[7]

The plan for this chapter is as follows. Section 3.2 introduces the causal modelling framework. Section 3.3 presents a causal model that represents the counterfactual structure involving a mental event and various physical events. The design of the model is somewhat unorthodox, but it satisfies a simple sufficient condition for causation in terms of causal models. Section 3.4 shows that mental causation survives possible refinements of this simple sufficient condition for causation. Section 3.4 applies the causal modelling framework to the problem of overlapping realizers. Section 3.5 addresses several objections according to which the model for mental causation is in some respect inappropriate. Section 3.7 discusses how interventionism fares vis-à-vis mental causation and argues that it can accommodate mental causation in a similar way to other causal modelling theories.

3.2 Causal Models

The point of causal models is to represent complex patterns of counterfactuals about events, from which conclusions about causal relations

[6] This approach is advocated in Galles and Pearl 1998, Halpern 2000, and Hiddleston 2005.

[7] Briggs (2012) argues that accounts of counterfactuals in terms of causal models face significant formal limitations. See also Halpern 2013 and Huber 2013.

between those events may be drawn. In this section I will introduce causal models by way of an example which does not involve mental events. I will follow the framework presented by Christopher Hitchcock (2001, 2007a). This framework deals with counterfactuals about, and causation between, token events; Section 3.7 will discuss interventionism, which uses a slightly different framework.

The example I will use is this:

> *Lightning Strike*. Lightning strikes my house, which subsequently catches fire. The sprinkler system is activated and extinguishes the fire. If the lightning had not struck, then the sprinkler system would not have been activated. If the sprinkler system had not been activated, then the house would have burnt down.

The counterfactuals that are true in *Lightning Strike* can be captured by a causal model. Formally, a *causal model* is an ordered pair $\langle \mathbf{V}, \mathbf{E} \rangle$ of a set of variables $\mathbf{V}$ and a set of equations $\mathbf{E}$ (also known as *structural equations*) that involve these variables. In causal modelling theories of token causation, the variables represent the occurrence or non-occurrence of token events, or different ways in which a token event can occur. In the *Lightning Strike* example it suffices to use binary variables that represent whether or not a given event occurs. In a given case there are typically various options for the choice of the variables. In our example, it seems reasonable to use the variables L, S, and B with the following interpretation:

> L = 1 if the lightning strikes, 0 otherwise
> S = 1 if the sprinkler system is activated, 0 otherwise
> B = 1 if the house burns down, 0 otherwise

(In this chapter, italic capital letters, which were used for properties in previous chapters, stand for variables. Some letters double as variables and as names for properties.) Let us also introduce a further variable, U, to represent the initial conditions before the lightning strike:

> U = 0 or U = 1 depending on the initial conditions

The purpose of U is to determine whether or not the lightning strike takes place. Introducing a variable like U is not mandatory in our example, but it will prove useful when modelling mental causation (see also Halpern and Pearl 2005: 856).

The equations in $\mathbf{E}$ represent counterfactuals about how the values of the variables in $\mathbf{V}$ depend on one another. As in previous chapters, I will

https://doi.org/10.1017/9781108762717.005 Published online by Cambridge University Press

assume Lewis's truth-conditions for counterfactuals. I stipulate that all the counterfactuals that we are dealing with be read in a non-backtracking way and I will continue to assume determinism. Some of the counterfactuals that are relevant for our example are already contained in the description of *Lightning Strike*, for instance, the claim that if the lightning had not struck, then the sprinkler system would not have been activated. Other counterfactuals are true in the example without being mentioned in the description, for instance, the claim that if the lightning had not struck while the sprinkler system had been activated anyway, then the house would not have burned down.

The equations of a causal model represent the counterfactuals that are true in a given case as follows. Each variable appears on the left-hand side of exactly one equation, which is called the equation *for* that variable. The equations are read from right to left: for any assignment of values to the variables on the right-hand side, the equation says that if the variables on the right-hand side had assumed those values, the variable on the left-hand side would have assumed the value that results from the function on the right-hand side.[8] This function may simply be identity, or it may be more complex, as in the equation for B below. The equations must not contain any redundant elements. If the value of a given variable does not make a difference to the value of the target variable over and above other variables, it must be eliminated from the equation of the target variable.

In the *Lightning Strike* example, the equations are as follows:

$$\begin{array}{ll} LS & U \Leftarrow 1 \\ & L \Leftarrow U \\ & S \Leftarrow L \\ & B \Leftarrow L \cdot (1 - S) \end{array}$$

(In what follows, I will use names such as '*LS*' to refer both to a causal model as a whole and to the model's equations.) U is the only variable whose value is given and not determined by any other variables. Such variables are called *exogenous*; the variables that are not exogenous (in our example, L, S, and B) are called *endogenous*. Typically, the equations for the exogenous variables specify the actual values of variables. This is what the equation for U (that is, the first equation) does here, which says that (the value of) U is 1. But, as we shall see shortly, when a modification of

[8] Some authors, including Hitchcock, use the identity sign to write the equations of a model instead of an arrow, but even when they are thus written, the equations do not express identities, but the non-symmetric relation described here.

https://doi.org/10.1017/9781108762717.005 Published online by Cambridge University Press

an original causal model is used to evaluate a counterfactual in the original model, the equations for exogenous variables can have the role of counterfactual suppositions.

I said earlier that the equations in a causal model represent counterfactuals. Admittedly, it is somewhat idiosyncratic to say that the equation for U, which merely sets U to 1, represents a counterfactual, but at any rate the equations for the endogenous variables represent genuine counterfactuals. The equation for L (that is, the second equation) represents two counterfactuals: if U had been 0, then L would have been 0; and if U had been 1, then L would have been 1. Similarly, the equation for S (that is, the third equation) represents two counterfactuals: if L had been 0, then S would have been 0; and if L had been 1, then S would have been 1. The equation for B represents four counterfactuals: if both L and S had been 1, then B would have been 0; if L had been 1 while S had been 0, then B would have been 1; if L had been 0 while S had been 1, then B would have been 0; lastly, if both L and S had been 0, then B would have been 0. We can of course translate these counterfactuals back into natural language. For instance, the penultimate counterfactual represented by the equation for B is the counterfactual mentioned three paragraphs back that is among those not contained in the description of *Lightning Strike*, namely the claim that if lightning had not struck while the sprinkler system had been activated anyway, then the house would not have burned down.

A set of equations can represent counterfactuals besides those that are represented by the individual equations. For instance, the equations in LS represent that if the sprinkler system had not been activated, then the house would have burned down. This is so, roughly, because if we set S to 0, the equations for the remaining variables yield that B is 1. More formally, we can define the truth of a counterfactual in a causal model as follows. Let X_1, X_2, . . ., Y be variables in a causal model $\langle \mathbf{V}, \mathbf{E} \rangle$. Then the *counterfactual* 'If X_1 had been x_1, X_2 had been x_2, . . ., then Y would have been y' is *true in* $\langle \mathbf{V}, \mathbf{E} \rangle$ if and only if, in the new causal model $\langle \mathbf{V}, \mathbf{E}' \rangle$ that we get by replacing the equations for X_1, X_2, . . . with the equations $X_1 \Leftarrow x_1$, $X_2 \Leftarrow x_2$, . . ., Y assumes value y. (This definition subsumes the truth of counterfactuals that are represented by individual equations.)

To see how the formal definition applies to the example of the claim that if the sprinkler system had not been activated, then the house would have burned down, the claim first needs to be rephrased in terms of variables. Thus rephrased, it says that if S had been 0, then B would have been 1. In order to evaluate the claim in our model LS, we need to replace the equation for S in that model, which reads '$S \Leftarrow L$', with an equation that

reads '$S \Leftarrow 0$'. In other words, we need to replace the original equation for S with an equation that sets S to the value specified in the antecedent of our counterfactual. This change of the equation of S yields the following new model:

$$\begin{array}{ll} LS' & U \Leftarrow 1 \\ & L \Leftarrow U \\ & S \Leftarrow 0 \\ & B \Leftarrow L \cdot (1 - S) \end{array}$$

In the new model LS', we can calculate the value of B by successively substituting specific values for variables: the equation for U sets U to 1; substituting '1' for 'U' in the equation for L sets L to 1 as well; the new equation for S sets S to 0; which, together with the other specific values gathered to far, allows us to substitute '$1 \cdot (1 - 0)$' for the right-hand-side of the equation for B. Thus, in LS', B assumes value $1 \cdot (1 - 0)$, that is, value 1, which is what the consequent of our counterfactual says; therefore, the counterfactual is true in our original model LS.

The procedure of determining the value of a given variable from the equations might not succeed if we get stuck in a circle. Therefore, it is common to impose the requirement that the sets of equations be *acyclic*, that is, that the equations can be ordered such that no variable appears on the right-hand side of any equation after it has appeared on the left-hand side. The equations in LS, for instance, are acyclic, since we can thus order them (in the reverse order of their listing). Non-acyclic sets of equations may still allow the procedure for evaluating counterfactuals to work, however; we will return to this issue in the following section.

Recall the standard definition of counterfactual dependence between events from Section 1.4: an actually occurring event e counterfactually depends on an actually occurring event c if and only if e would not have occurred if c had not occurred. A notion of counterfactual dependence between *variables* can be defined in the causal modelling framework. Thus, for variables X and Y in a causal model $\langle \mathbf{V}, \mathbf{E} \rangle$ that have the actual values x and y respectively, let us say that *Y counterfactually depends on X* in $\langle \mathbf{V}, \mathbf{E} \rangle$ if and only if there are non-actual values x' and y' of X and Y, respectively, such that the counterfactual 'If X had been x', then Y would have been y'' is true in $\langle \mathbf{V}, \mathbf{E} \rangle$.

In LS, for example, the variable for the sprinkler system's being activated, S, counterfactually depends on the variable for the lightning strike,

https://doi.org/10.1017/9781108762717.005 Published online by Cambridge University Press

L, since the actual values of *S* and *L* are both 1 and the counterfactual 'If *L* had been 0, then *S* would have been 0' is true in our causal model. By contrast, the variable for the house's burning down, *B*, does not counterfactually depend on *L*, for the counterfactual 'If *L* had been 0, then *B* would have been 1' is not true in our model, and no other counterfactuals with non-actual values of both *L* and *B* are available in it.

What is the relation between counterfactual dependence between variables in a causal model and counterfactual dependence between events? If variables *X* and *Y* are binary and their actual values stand for the occurrence of events (as opposed to the occurrence of omissions), we get standard counterfactual dependence between events from the counterfactual dependence between variables *X* and *Y*. More precisely, if *X* and *Y* are binary and their actual values – let *x* and *y* be these actual values – stand for the occurrence of events, then it follows from the counterfactual dependence of variable *Y* on variable *X* that the event represented by $Y = y$ counterfactually depends on the event represented by $X = x$ in the standard sense. For example, in *LS*, variable *S* counterfactually depends on variable *L*; both variables are binary; and their actual values stand for the occurrence of events, namely $L = 1$ for the lightning strike and $S = 1$ for the activation of the sprinkler system. It follows that the activation of the sprinkler system counterfactually depends on the lightning strike in the standard sense of counterfactual dependence between events.

In cases of multi-valued variables and omissions, we might have counterfactual dependence between variables without counterfactual dependence between events. For instance, suppose that a multi-valued variable *Y* counterfactually depends on a multi-valued variable *X*. It is consistent with this counterfactual dependence of *Y* on *X* that it is merely the case that the event represented by *Y* would have occurred slightly differently if the event represented by *X* had occurred slightly differently while the event represented by *Y* would still have occurred (albeit perhaps slightly differently) if the event represented by *X* had not occurred. Or suppose that variables *X* and *Y* are binary, *Y* counterfactually depends on *X*, and the actual value of *X* stands for an omission: if omissions are not events, no counterfactual dependence between events follows.

In previous chapters, we took counterfactual dependence to suffice for causation between events. (At least where the dependent event occurs later than the event it depends on – I suppress that qualification until Section 3.6.) We may continue to do so without further ado by taking counterfactual dependence between variables in corresponding cases to suffice for causation. When putative mental causes are concerned, the

new task will be to establish the counterfactual dependence claims in the framework of causal modelling.

Cases of counterfactual dependence between variables that might not correspond to cases of counterfactual dependence between events have to be treated with a bit of caution. I will set aside until Section 3.4 cases of counterfactual dependence between variables whose actual values stand for omissions. As for counterfactual dependence involving multi-valued variables, we should distinguish cases where the variable for the putative effect remains binary from cases where the variable for the putative effect is itself multi-valued. If the variable for the putative effect is binary and it counterfactually depends on a multi-valued variable that stands for the putative cause, it seems straightforward to infer a causal relation. Consider the following non-technical analogue: irrespective of what would have happened had I not thrown the dart, if the balloon would not have burst had I thrown the dart differently, we can still infer that my (actual) throw caused the bursting. If the variable for the putative effect is multi-valued, things are not quite so straightforward. Consider again a non-technical analogue: suppose that event *e* would merely have occurred slightly differently if event *c* had not occurred or if event *c* had occurred slightly differently. Should we infer that *c* caused *e*? A positive answer seems plausible, but it is not obvious.[9] Fortunately, we need not settle the issue here. The important cases of multi-valued variables for putative causes in this chapter will be cases where the variable for the putative effect is still binary. For simplicity, I will ignore the potential complication from multi-valued effect variables and assume that counterfactual dependence between variables (multi-valued or not) is sufficient for causation, at least when their actual values do not stand for omissions.

Let us return to the tools of causal modelling. The equations from a causal model can be used to construct a *causal graph*. The causal graph of a given model contains all the variables of that model. An arrow is drawn from a variable to another if and only if the first variable appears on the right-hand side of the equation for the second variable. Figure 3.1 shows the causal graph of the *Lightning Strike* example.

While causal graphs often make the structure of the equations in a causal model easier to grasp, we should not overestimate the amount of information it carries. There can be arrows that do not correspond to counterfactual dependence. In our example, there is an arrow from *L* to *B* although, as we

[9] The issue is closely related to Lewis's (2004) theory of causation as influence. For critical discussion of this theory, see Collins 2000 and Kvart 2001.

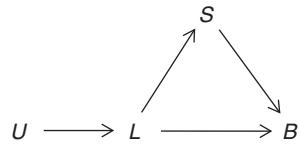

Figure 3.1. The causal graph of *LS*

saw, *B* does not counterfactually depend on *L*.[10] We have not pronounced any verdicts about which events in our example fail to stand in a causal relation. But it seems at least doubtful that the lightning strike causes the house *not* to burn down.[11] If there turns out to be no causal relation between what is represented by the actual values of *L* and *B*,[12] our case also illustrates that there can be an arrow in a causal graph without a corresponding causal relation.

We can define a graph as *acyclic* if and only if one can never double back on the same variable by following a path along the direction of the arrows. A graph is acyclic if and only if the corresponding equations are acyclic.

For further illustration of the causal modelling framework, let us model the assassination example from the previous section. This example will also allow us to anticipate the strategy for formulating conditions on causation in causal models that will be employed later in this chapter. In the example, Apprentice fires and kills Victim. If Apprentice had not fired, Expert would have fired and killed Victim instead. If neither Apprentice nor Expert had fired, Victim would not have died. In order to model the example, we need three binary variables that represent the actions of Apprentice and Expert and the fate of Victim. Let us use the variables *A, E*, and *D* with the following interpretations:

A = 1 if Assassin fires, 0 otherwise
E = 1 if Expert fires, 0 otherwise
D = 1 if Victim dies, 0 otherwise

[10] Variable *B potentially* counterfactually depends on variable *L*, however, for *B* would counterfactually depend on *L* if *S* were held fixed at value 0.

[11] Strictly speaking, the house's not burning down is an omission, but this is inessential, for we would reach the same verdict if we replaced *B* with a variable that stands for the *bona fide* event of the house's (still) standing at a later time: it seems at least doubtful that the lightning strike causes that event.

[12] The case is similar to the counterexample to the transitivity of causation that was discussed in Section 1.4. It is even more similar to an example by Hartry Field from an unpublished lecture, which is also often cited as a counterexample to the transitivity of causation. In Field's example, someone places a bomb outside Smith's door; Smith sees the bomb, disarms it and survives. For a discussion of the case, see Paul and Hall 2013: 215–231.

(One could add an exogenous variable U, as in the model for *Lightning Strike*, but I omit such a variable here for simplicity.) The counterfactual claims that are true in the example include the following: if A had been 0, then E would have been 1 (if Apprentice had not fired, then Expert would have fired instead); if A had been 0, then D would have been 1 (if Apprentice had not fired, then Victim would still have died); if A had been 0 and E had been 0, then D would have been 0 (if neither Apprentice nor Expert had fired, then Victim would not have died). The equations for the case are as follows:

$$\begin{array}{ll} AS & A \Leftarrow 1 \\ & E \Leftarrow 1 - A \\ & D \Leftarrow \mathrm{Max}\{A, E\} \end{array}$$

It can easily be verified that the equations in AS represent the above counterfactuals (and further true counterfactuals about the case). For instance, we can verify the truth of 'If A had been 0 and E had been 0, then D would have been 0' as follows: replace the equation for A with the equation '$A \Leftarrow 0$', which sets A to the value specified in the antecedent. Similarly, replace the equation for E with the equation '$E \Leftarrow 0$'. In the resulting model, substitute the values of A and E in the equation for D. Thus, D assumes value Max{0, 0}, that is, value 0. This is the value specified in the consequent of our counterfactual, so the counterfactual is true in our original model AS.

The counterfactual 'If A had been 0, then D would have been 0' is false in our model AS, because Victim would still have died if Apprentice had not fired. Thus, variable D does not counterfactually depend on variable A, for the only non-actual value that variables A and D can assume is 0. Apprentice's firing (that is, the event represented by $A = 1$), causes Victim's death (the event represented by $D = 1$). Since D does not counterfactually depend on A, this causal relation cannot obtain because of counterfactual dependence between the two variables. We saw in the previous section, however, that it seems plausible that what underlies the causal relation in this case is that Victim would not have died if neither Apprentice nor Expert had fired. In variable-talk, what underlies the causal relation seems to be the truth of the counterfactual 'If A had been 0 and E had been 0, then D would have been 0.' The causal graph of our model, which is shown in Figure 3.2, suggests a reason for the relevance of this counterfactual. There are two ways of travelling from A to D by following the direction of the arrows in the graph. One can travel from A to D directly, or one can

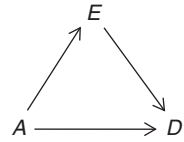

Figure 3.2. The causal graph of *AS*

travel via *E*. One of these routes, the direct route, yields dependence of *D* on *A* if we hold the variable that is not on this route, *E*, fixed at its actual value. To say that there is dependence between two variables when certain other variables are held fixed at their actual values is simply to say that a certain counterfactual with a complex antecedent is true. In our case, that counterfactual is 'If *A* had been 0 and *E* had been 0, then *D* would have been 0.'

The strategy behind this kind of reasoning is that the causal graph of a model provides information about routes or paths in the model, which in turn allows us to formulate conditions on causation in the form of counterfactuals with complex antecedents; these antecedents say that, first, the variable for the candidate cause assumes a different value while, second, certain off-path variables are held fixed at their actual values. In Section 3.5, I will investigate how this strategy can be used to deal with the problem of overlapping realizers. First, however, let us turn to mental causation.

3.3 A Model for Mental Causes

This section applies the causal modelling framework to mental causation. The case of non-reductive physicalism will be considered first; at the end of this section we will turn to super-nomological dualism. The plan for the non-reductive physicalist case is to start by selecting the variables for a causal model for mental causation. Then we will draw up an inventory of various true counterfactuals involving those variables. A set of equations that represents these counterfactuals will complete the causal model for mental causation. The model will be shown to have the consequence that there is mental causation because a simple sufficient condition for causation in terms of causal models is satisfied. We shall see that, *mutatis mutandis*, the results from the non-reductive physicalist case also hold for the dualist case.

Assume that we are dealing with a specific instance of a mental property and a later physical event that is the (putative) effect of the mental

property-instance. Without loss of generality, let my headache be the instance of the mental property, and let my hand's moving towards the aspirin be the later physical event. For modelling the case, we need a binary variable *M* that represents the instantiation or non-instantiation of the property of having a headache (equivalently, the occurrence or non-occurrence of the corresponding strong Kimian event that is constituted, *inter alia*, by the property of having a headache). We need a variable *P* that represents the instantiation or non-instantiation of the various realizers of headaches. Variable *P* should be multi-valued, so that it can represent the instantiation of the actual realizer of the mental property, the instantiation of alternative realizers, and the non-instantiation of all realizers. (In Section 3.6 I will address objections to modelling the mental event and its realizers in the way I have just suggested.) We need a binary variable *P** that represents the occurrence or non-occurrence of my hand's moving towards the aspirin, the putative physical effect of the headache. Lastly, let us use an exogenous variable *U* to represent initial conditions again. We can think of *U* as representing the physical goings-on in my body and perhaps in my immediate environment just before the mental event occurs. The purpose of *U* is to at least partially determine the value of *P*; it should therefore be multi-valued as well. We can specify what the different values of the variables represent as follows:

U = 0 or 1 or 2 . . . depending on the initial conditions
P = 0 if no realizer of headaches is instantiated, 1 or 2 or . . . otherwise, depending on which realizer is instantiated
M = 1 if the property of having a headache is instantiated, 0 otherwise
*P** = 1 if my hand moves towards the aspirin, 0 otherwise

If one would like to flesh out the different positive values of *P*, one can let *P* = 1 represent my having firing c-fibres, let *P* = 2 represent my having firing x-fibres, let *P* = 3 represent my having an active semiconductor network of a certain kind in my head, etc. Correspondingly, we can think of *U* = 1 as representing a state of my body and perhaps my immediate environment that is standardly followed by a c-fibre firing of mine, *U* = 2 as representing a state that is standardly followed by an x-fibre firing of mine, *U* = 3 as representing a state that is standardly followed by semiconductor activity of a certain kind in my head, etc. Since *P* = 0 represents that no realizer of headaches is instantiated, we can think of *U* = 0 as representing a state that is standardly followed by no instantiation of a realizer of headaches. For our purposes, however, all that is matters is that the positive values of *P* represent instantiations of the different

realizers of headaches and that the positive values of U represent corresponding earlier states that standardly lead up to them.

What are the counterfactual relationships between the different values of our variables that our causal model should capture? Let us start with the relationship between P and P^*. Part of this relationship is straightforward. As we saw in the previous chapter, if I had not instantiated any realizer of headaches, my hand would not have moved towards the aspirin. Expressed in terms of variables, this counterfactual reads:

(1) If P had been 0, then P^* would have been 0.

It is also straightforward that if I had had firing c-fibres, my hand would have moved towards the aspirin. Indeed, this counterfactual is automatically true given that in fact I have firing c-fibres and my hand moves towards the aspirin. (I'll present a counterfactual in terms of variables that subsumes this counterfactual in a moment.)

What is not so straightforward is what would have been the case if I had instantiated the alternative realizers of headaches. At first sight, it might seem that my hand would still have moved towards the aspirin if I had instantiated any such realizer. But a moment's reflection shows that the latter claim is dubious at best. Many of the alternative realizers are rather exotic. Implanting them in my body is likely to be a pretty disruptive procedure that yields behavioural effects (or a lack thereof) of the alternative realizers that differ drastically from the behavioural effects of firing c-fibres. In other words, it seems likely that the instantiations of many alternative realizers would not be followed by my hand's moving towards the aspirin. It will turn out in the next chapter that this result has some positive repercussions for solving the exclusion problem, but it would complicate our task of model-building considerably (not least because the list of alternative realizers is open-ended). In this chapter, I will therefore assume, for simplicity, that all the alternative realizers have uniform behavioural consequences. I will assume, that is, that the instantiation of any alternative realizer of headaches would have been followed by my hand's moving towards the aspirin. If we combine this assumption with the straightforward claim that the actual realizer of headaches would be (indeed, is) followed by hand's moving towards the aspirin, we get the following claim in terms of variables:

(2) If P had been p, with $p \neq 0$, then P^* would have been 1.[13]

[13] Strictly speaking, (2) is not a counterfactual but a schema, or perhaps a counterfactual that is in the scope of a universal quantifier that ranges over the possible values of P. For simplicity I will treat (2)

Claim (2) says that, for any non-zero value of P, if P had assumed that value, then P^* would have assumed value 1. Without the assumption of (2) for simplicity, the arguments from this chapter would still go through, at least *mutatis mutandis*.[14]

Let us consider the relationship between the realizer-variable, P, and the variable for the headache, M. A number of counterfactuals about how the values of M and P are related can be read off from the relation between the mental property and its realizing properties. Recall that, according to non-reductive physicalism, mental properties strongly supervene on physical properties. Recall that from this strong supervenience it follows that the instantiation of a mental property strictly implies, and is strictly implied by, the instantiation of a realizer of that mental property. Applied to our case, we get the following two strict conditionals:

(i) Necessarily, if the property of having a headache is instantiated, then a realizer of headaches is instantiated.
(ii) Necessarily, if a realizer of headaches is instantiated, then the property of having a headache is instantiated.

Claims (i) and (ii) express the consequence of strong supervenience in terms of properties. If we rephrase these claims in terms of variables, we get:

(3) Necessarily, if M is 1, then P is not 0.
(4) Necessarily, if P is p, with $p \neq 0$, then M is 1.

By contraposition and the assumption that, necessarily, M is either 1 or 0, we get:

as a counterfactual, however (similarly for claim (8) below). As an alternative to (2), we could formulate a separate counterfactual for each possible value of the antecedent-variable: we could use the claims $P = 1 \mathrel{\square\!\!\rightarrow} P^* = 1$, $P = 2 \mathrel{\square\!\!\rightarrow} P^* = 1$, etc. instead. This would make our set of equations rather cumbersome, although the excess complexity would not be as bad as in the case of the alternative model MC^* that will be discussed in Section 3.6.

[14] It is an option to change the framework and to let variable P^* be multi-valued, with different non-zero values of P^* standing for different variations of my post-headache behaviour, instead of assuming (2) with binary P^*. A problem with this option is that, for all we know, different alternative realizers result in exactly the same variation from my actual behaviour, though we do not know which of the many realizers do. In response, one could try to err on the side of proliferation of values, give P^* as many values as P has, and stipulate that, for any value x of P, if P had been x, then P^* would have been x too. But then P^* would have to assume the different values that correspond to a given specific behavioural variation at once, which is impossible (on constraints on the values of variables; see Hitchcock 2007a: 502 and Halpern and Hitchcock 2010, §4.3).

(5) Necessarily, if P is 0, then M is 0.
(6) Necessarily, if M is 0, then P is 0.[15]

Strict conditionals entail the corresponding counterfactuals (see Section 1.4). Thus, (3)–(6) entail, respectively, the following counterfactuals:

(7) If M had been 1, then P would not have been 0.[16]
(8) If P had been p, with $p \neq 0$, then M would have been 1.
(9) If P had been 0, then M would have been 0.
(10) If M had been 0, then P would have been 0.

Note that these counterfactuals do not backtrack, since the mental event and its possible realizers occur at the same time. (Whether any causal relation between the mental event and its realizers follows from such counterfactuals will be discussed in Section 3.6.)

Variable U was introduced to represent initial conditions that should at least contribute to determining the value of P. On the face of it, it might seem that we can simply say that which realizer (if any) is instantiated depends merely on the initial conditions, such that P would assume whatever value U had. But things are more complicated, since, by (7) and (10), whether a realizer occurs also depends on whether or not the mental event occurs. Moreover, since the dependence expressed by (7) and (10) derives from a metaphysically necessary connection between the mental property and its realizers that is expressed by (3) and (6) respectively, this connection should trump any contingent connection between the initial conditions and the instantiation of a realizer.

For the case in which U and M are both 0, claim (6) and the idea that P assumes the value of U pull in the same direction; in this case P would be 0 too:

(11) If U had been 0 and M had been 0, then P would have been 0.

Put less technically, (11) says that if I had been in a state that is standardly followed by no instantiation of a realizer of headaches

[15] I have not defined the truth of claims about metaphysical modality such as (3)–(6) in a causal model; indeed, these claims are not supposed to be true *in* the causal model that we are about to construct in the strict sense. I am using variables in these claims merely to facilitate the derivation of certain counterfactuals which, in contrast to (3)–(6), will be true in our model in the strict sense.

[16] Counterfactuals like (7) that are not about variables' assuming specific values are not covered by our definition of the truth of a counterfactual in a causal model. The definition can easily be extended, however; see Halpern and Pearl 2005: 851–852. For a discussion of counterfactuals with disjunctive antecedents, see Briggs 2012.

and I had not had a headache, then I would not have instantiated any realizer of headaches. Similarly, in a case in which U and M are both not 0 (such that M is 1), claim (3) and the idea that P assumes the value of U agree in predicting that P would assume the value of U:

(12) If U had been u, with $u \neq 0$, and M had been 1, then P would have been u.

In intuitive terms, (12) says that if I had been in a state that is standardly followed by the instantiation of a certain realizer of headaches and I had had a headache, then I would have instantiated the realizer.

If one of U and M had been 0 while the other had not been 0, however, there is a conflict between the idea that P assumes the value of U on the one hand, and (3) and (6) on the other, which must be resolved in favour of (3) and (6). Thus, we get:

(13) If U had not been 0 and M had been 0, then P would have been 0.

The intuitive gloss of (13) says that if I had been in a state that is standardly followed by the instantiation of a realizer of headaches, but I had not had a headache, then I would not have instantiated a realizer of headaches. Parallel reasoning yields:

(14) If U had been 0 and M had been 1, then P would not have been 0.

Intuitively, (14) says that if I had been in a state that is standardly followed by no instantiation of a realizer of headaches, but had had a headache, then I would have instantiated a realizer of headaches. For definiteness, let us assume that P would have been 1 if U had been 0 and M had been 1:

(14′) If U had been 0 and M had been 1, then P would have been 1.

Given the suggestion about how to flesh out the different positive values of the realizer-variable P, (14′) says that if I had been in a state that is standardly followed by no instantiation of a realizer of headaches, but had had a headache, then I would have had firing c-fibres. The assumption of (14′) will facilitate constructing our causal model, but nothing hinges on it.

Here is a set of equations that represents the counterfactuals we have established so far:

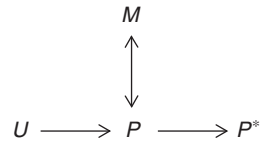

Figure 3.3. The causal graph of *MC*

$$
\begin{aligned}
MC \quad & U \Leftarrow 1 \\
& P \Leftarrow 1 \text{ if } U \text{ is } 0 \text{ and } M \text{ is } 1, M \cdot U \text{ otherwise} \\
& M \Leftarrow \mathrm{Min}\{P, 1\} \\
& P^* \Leftarrow \mathrm{Min}\{P, 1\}
\end{aligned}
$$

Figure 3.3 shows the causal graph corresponding to *MC*.

The graph of *MC* is not acyclic, since we can go back and forth on the double-arrow between *P* and *M*. Correspondingly, the equations in *MC* are not acyclic either. Generally, a set of equations that fails to be acyclic may have no solutions or more than one solution. It can easily be verified that the equations in *MC* have two solutions. On one solution, *U*, *P*, *M*, and *P** are all 1; on the other solution, *U* is 1 while *P*, *M*, and *P** are 0.

That the equations in our causal model fail to be acyclic does not preclude us from evaluating counterfactuals in this model, however. At least it does not preclude us from evaluating the counterfactuals that are most interesting for our purposes, namely those that are about the relation between *M* and *P**. This is so because the truth-conditions for a counterfactual in a given causal model that were given in the previous section draw on what is the case in a certain modification of that model. In our case, this modification has a unique solution even though the original model does not.[17]

The following counterfactual is the most interesting one for our purposes:

(15) If *M* had been 0, then *P** would have been 0.

Counterfactual (15) says that my hand would not have moved towards the aspirin if I had not had a headache. In order to evaluate counterfactual (15) in our model *MC*, we have to consider the new model we get from *MC* by

[17] It is also possible to give a general definition for the truth of counterfactuals for non-acyclic equations; see Halpern and Pearl 2005: 883–884.

https://doi.org/10.1017/9781108762717.005 Published online by Cambridge University Press

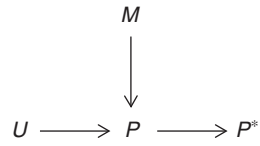

Figure 3.4. The causal graph of *MC′*

replacing the equation for *M* with an equation that sets *M* to the value specified in the antecedent of (15), that is, to 0. Here is this new model:

$$
\begin{aligned}
MC' \quad & U \Leftarrow 1 \\
& P \Leftarrow \text{if } U \text{ is 0 and } M \text{ is 1, } M \cdot U \text{ otherwise} \\
& M \Leftarrow 0 \\
& P^* \Leftarrow \text{Min}\{P, 1\}
\end{aligned}
$$

Figure 3.4 shows the causal graph corresponding to the new model *MC′*.

The equations in *MC′* and the corresponding graph are acyclic. This can easily be seen from the graph of *MC′*, where one cannot trace a path that contains a variable twice by following the direction of the arrows. Since the equations in *MC′* are acyclic, they have a unique solution. According to this solution, *M*, *P*, and *P** are all 0, while *U* is 1. Since *P** is 0 according to this solution, counterfactual (15) is true in our original model *MC*. (By parallel reasoning we get that 'If *M* had been 1, then *P** would have been 1' is true in our original model *MC*.)

We took certain counterfactuals as a starting-point and constructed a causal model, *MC*, whose individual equations represented those counterfactuals. It turned out that counterfactual (15) is also true in *MC*. While it will be seen that this is good news for the project of accommodating mental causation in the causal modelling framework (a project we will resume in a moment), I should forestall a possible misunderstanding of the result. Equations that individually represent true counterfactuals may well collectively represent false counterfactuals. It may happen, in other words, that the individual equations of a model represent counterfactuals that are in fact true while a counterfactual that is in fact false is also true in that model.[18] So the mere fact that our model *MC* was constructed to yield

[18] Take, for instance, a model (call it *TR*) that consists in the set of binary variables {*X, Y, Z*} and the set of equations {$X \Leftarrow 1$, $Y \Leftarrow X$, $Z \Leftarrow Y$}. The second equation represents (*inter alia*) the counterfactual (i) 'If *X* had been 0, then *Y* would have been 0'; the third equation represents (ii) 'If *Y* had been 0, then *Z* would have been 0.' The counterfactual (iii) 'If *X* had been 0, then *Z* would have been 0' is

certain true counterfactuals and that (15), too, is true in *MC* should not be taken to demonstrate that (15) is in fact true. But we can make a stronger case for (15). Counterfactual (15) follows logically from the counterfactuals that we built into our model. Specifically, (15) follows from (1), (9), and (10). (The inference has the same form as the inference from (16)–(18) to (19) in Section 2.5 had.) Thus, our model does not by itself prove (15), but since (15) can be established independently, it is a virtue of the model that (15) is true in it.[19]

The two solutions to the equations in *MC* leave it open whether the mental event, a realizer-instance, and the later physical event actually occur – according to one solution, they do; according to the other, they do not.[20] But we may stipulate that they do all actually occur, such that *U*, *M*, *P*, and *P** are all 1.[21] We may stipulate, in other words, that I am in a state that is standardly followed by a c-fibre firing, that later I have a headache and firing c-fibres, and that later still my hand moves towards the aspirin. Then it follows from (15), by the definition of counterfactual dependence between variables from the previous section, that variable *P** counterfactually depends on variable *M*. As for the relation between *P** and *P*, counterfactual (1) says that if *P* had been 0, then *P** would have been 0 (if I had not instantiated a realizer of headaches, then my hand would not have moved towards the aspirin). That (1) is true was built into our model *MC*. Given that the actual values of *U*, *M*, *P*, and *P** are all 1, by (1) *P** counterfactually depends on *P* in addition to counterfactually depending on *M*.

We saw in the previous section that we may draw causal conclusions from counterfactual dependence between variables, provided their actual values do not correspond to omissions. The actual values of *M*, *P*, and *P** do not correspond to omissions. Thus, we may infer from the

true in *TR* without being represented by an individual equation. Since counterfactuals are not generally transitive (see Section 1.4) it might be that in fact (i) and (ii) are true while (iii) is false.

[19] Hitchcock (2001: 287, 2007a: 502–503) takes the requirement that no counterfactuals be true in a given model that are in fact false to be a criterion for the appropriateness of that causal model. Another requirement he imposes, namely that the variables correspond to events that are sufficiently independent, will be discussed in Section 3.6.

[20] This is consistent with our assumption of determinism. It may well be that the initial conditions represented by *U* are rich enough for it to follow from the actual laws of nature and the assumption that *U* is 1 that *P*, *M*, and *P** are all 1. That there are two solutions to our equations that involve the same value of *U* while differing in the values of *P*, *M*, and *P** merely shows that the lawful connection between *U* and the variables that represent later events can be broken in a counterfactual situation where *M*'s being 0 forces *P*'s being 0.

[21] Instead of this stipulation, we could build the actual values of all our variables into the model; see Briggs 2012: 144.

counterfactual dependence of variable P^* on variable M that (the event represented by) $M = 1$ causes (the event represented by) $P^* = 1$. Similarly, from the counterfactual dependence of P^* on P we may infer that $P = 1$ causes $P^* = 1$. We may infer, that is, that my headache causes my hand to move towards the aspirin and that my c-fibres' firing causes my hand to move towards the aspirin.

Thus, causal models can capture the patterns of counterfactual dependence and causation that hold in cases involving supervenient mental properties. This is good news for those who are sceptical about the ability of causal modelling theories to accommodate mental causation in light of the controversy about interventionism (which will be discussed in Section 3.7). To the disinterested, it may seem that causal models have not given us anything that we did not already have. Our simple argument in Section 2.2 establishes the same as the comparatively complicated model *MC* does, namely that some actually occurring physical events counterfactually depend on actually occurring mental events. We can feed this result into our old sufficient condition for causation in terms of counterfactual dependence between events irrespective of how we arrived at the result. This is true, but causal modelling can do more. First, some theorists have suggested replacing simple counterfactual dependence as a sufficient condition for causation with more sophisticated sufficient conditions that are not entailed by simple counterfactual dependence. We shall see in the following section that these more sophisticated conditions still apply in cases involving mental events like the one discussed in this section. Thus, causal modelling theories can still accommodate mental causation when they move beyond the simple conditions for causation that were used by standard counterfactual theories. Second, we saw in Section 2.3 that one option to solve the problem of overlapping realizers is to qualify the simple sufficient condition for causation within the causal modelling framework. We shall see that, eventually, causal modelling is better suited to formulating a sufficient condition for explanatory relevance than to formulating a sufficient causation for causation in order to solve the problem of overlapping realizers. But the condition for explanatory relevance, too, essentially involves the resources of the framework, especially the structure of a causal model, which can be read off from the model's graph.

Before turning to the refinements of our sufficient condition for causation, let me note that our causal model *MC* can be used by super-nomological dualists as well as non-reductive physicalists. We took as a starting-point the strict conditionals (3)–(6), which are false if dualism is true, but we might equally well have started from the corresponding counterfactuals (7)–(10), which are true if super-nomological dualism is. According to super-

nomological dualism, the psychophysical laws could not have failed so easily as the ordinary laws of nature. The relation between the mental property and its physical bases is a matter of psychophysical laws, whereas the relation between physical goings-on at different times is a matter of ordinary laws of nature. Thus, as in the case of non-reductive physicalism, the relation between variable P and variable M trumps that between variable U and variable P in case there is a conflict. As a result, we get counterfactuals (11)–(14)/(14′), the model *MC* and the corresponding causal results, just as we did in the non-reductive physicalism case.

3.4 Defaults and Normality

Our old counterfactual principle about causation from Section 1.4 said that if an actually occurring event e counterfactually depends on an actually occurring event c, then c causes e. (As I said earlier, the temporal qualification that e occurs later than c will be suppressed until Section 3.6.) An analogous principle in terms of causal modelling says that if, in a causal model, variable X is x, variable Y is y, and Y counterfactually depends on X, then (the event represented by) $X = x$ causes (the event represented by) $Y = y$. One might endorse the old principle without endorsing the causal modelling principle if one thinks that omissions are not events or that omissions cannot be causes or effects. In this case, a refinement of the causal modelling principle is called for. I will not pronounce a verdict on whether there really is a need for such a refinement. It will suffice for our purposes to show that, if there is, the resulting principle about causation can still accommodate mental causation.

I will illustrate the refinements of the causal modelling principle with the following case:[22]

> *Omission*. Assassin poisons Victim's drink. Bodyguard possesses an antidote that would neutralize the poison, but does not administer it. Victim dies of poisoning, but would not have died if the drink had not been poisoned or if the antidote had been administered.

We can model *Omission* by using variables A, G, and D with the following interpretation:

> A = 1 if Assassin poisons Victim's drink, 0 otherwise
> G = 1 if Bodyguard administers the antidote, 0 otherwise
> D = 1 if Victim dies, 0 otherwise

[22] My presentation of the case follows Hitchcock 2007a: 504–505.

https://doi.org/10.1017/9781108762717.005 Published online by Cambridge University Press

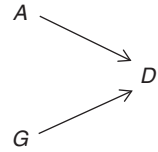

Figure 3.5. The causal graph of *OM*

The description of *Omission* tells us that the actual value of A is 1, and that the actual value of G is 0. It also tells us that D would have been 0 if A had been 0 (Victim would not have died had Assassin not poisoned the drink), and that D would have been 0 if A and G had both been 1 (Victim would not have died if the drink had been poisoned and the antidote administered). Thus, we get the following equations:

$$\begin{aligned} OM \quad & A \Leftarrow 1 \\ & G \Leftarrow 0 \\ & D \Leftarrow \text{Min}\{A, 1 - G\} \end{aligned}$$

Figure 3.5 shows the causal graph corresponding to *OM*.

Variable D counterfactually depends on variable G in the model *OM*, for if we replace the equation for G with one that sets G to the non-actual value 1, the equations yield that D assumes the non-actual value 0. That is, the equations tell us that Victim would not have died if Bodyguard had administered the antidote. If we take counterfactual dependence between variables to suffice for causation, it follows that $G = 0$ causes $D = 1$. It follows, in other words, that Bodyguard's failure to administer the antidote causes Victim's death.

If one finds this result implausible,[23] one can pursue different strategies for qualifying the sufficiency of counterfactual dependence for causation within the causal modelling framework. This section discusses two such strategies. The first strategy adds the qualification that the causal model in which counterfactual dependence obtains be of a certain kind. The second strategy instead adds the qualification that only counterfactuals with certain features should be regarded as indicative of causation.[24] In what follows, I shall present both strategies and argue that either strategy still

[23] For a recent discussion of causation and responsibility in cases of omission, see Moore 2009.

[24] The first strategy is pursued in Hitchcock 2007a, the second in Halpern and Pearl 2005, Hall 2007, Halpern 2008, Halpern and Hitchcock 2010, and Halpern and Hitchcock 2015. For critical discussion of the first strategy, see Wolff 2016.

allows mental causation, or at any rate accommodates mental causation at least as well as physical-to-physical causation.

Following Hitchcock (2007a), we can formulate a new sufficient condition according to which counterfactual dependence is sufficient for causation in causal models where the connection between the putative cause and the putative effect is of a certain kind. Hitchcock introduces the distinction between *default* and *deviant* values of variables. Roughly, default values are those that correspond to states of a system that persist in the absence of outside influence, while deviant values are those that correspond to states that do not thus persist. According to the present suggestion, we may take counterfactual dependence to suffice for causation in a given model if in that model the connection between the putative cause and the putative effect is such that default values of previous variables always yield default values of later variables.

This idea can be made more precise by using the following terminology.[25] Let a *predecessor* of a variable be a variable that occurs on the right-hand side of the equation for that variable, and let a *path* be a sequence of variables such that the first variable is a predecessor of the second variable, the second variable is a predecessor of the third variable, . . ., and the penultimate variable is a predecessor of the last variable. Let a *path from X to Y* be a path that contains X and Y as first and last elements of such a sequence, respectively. Let a path be *acyclic* if and only if it does not contain any variables twice, *cyclic* otherwise. In a causal graph, the predecessors of a variable X are those variables that have an arrow pointing towards X, and a path from X to Y can be traced by following the direction of arrowheads. Let the *network* connecting variable X to variable Y be the set of variables that are on some path or other between X and Y. We can now define the notion of a self-contained network as follows: a network connecting X and Y is *self-contained* if and only if each variable Z in this network takes its default value if all its predecessors in the network (if any) take their default value while all its predecessors outside of the network (if any) assume their actual values. Following Hitchcock, one might claim that the counterfactual dependence of variable Y on variable X is sufficient for $X = x$ to cause $Y = y$ in a given causal model if the network connecting X to Y is self-contained.[26] (In case variables have multiple default values, 'its

[25] My terminology differs slightly from Hitchcock's in order to be applicable to sets of equations that are not acyclic.

[26] Hitchcock (2007a: 511–512) takes counterfactual dependence to be necessary as well as sufficient for causation for the values of variables that are connected by a self-contained network, or at least

default value' and 'their default value' should be replaced by 'a default value' in the definition.)

This sufficient condition for causation no longer implies that in *Omission* Bodyguard's inaction causes Victim's death. We may assume that the default state of Assassin is not to poison Victim's drink, that the default state of Bodyguard is not to administer the antidote, and that the default state of Victim is not to die. Correspondingly, the default values of *A, G*, and *D* are all 0. The network connecting *G* to *D* contains just these two variables themselves. This network is not self-contained, for if *G* assumes its default value 0 while *A* assumes its actual value 1, then *D* assumes value 1, which is not its default. By contrast, the network connecting *A* to *D* is self-contained, for if *A* assumes its default value 0 while *G* assumes its actual value 0, then *D* assumes its default value 0. Thus, the new sufficient condition for causation rules that Assassin's poisoning the drink causes Victim to die, but it remains silent on whether Bodyguard's failure to administer the antidote causes the death.

The new sufficient condition also remains silent on whether there is causation in the cases of double prevention that we discussed in Sections 1.6 and 2.6.[27] Take our neuron example of double prevention, where the firing of *c* prevents the firing of *d*, which, had it not been prevented, would have prevented the firing of *e* (see Figure 1.1 on p. 51).

To model the example, let us use the variables *A, B, C, D*, and *E*, where *A* = 1 if neuron *a* fires, 0 otherwise; *B* = 1 if neuron *b* fires, 0 otherwise; etc. The following equations give us the counterfactuals that are true in the example:

$$
\begin{array}{ll}
DP & A \Leftarrow 1 \\
 & B \Leftarrow 1 \\
 & C \Leftarrow 1 \\
 & D \Leftarrow B \cdot (1 - C) \\
 & E \Leftarrow A \cdot (1 - D)
\end{array}
$$

Figure 3.6 shows the causal graph corresponding to *DP*.

In the model *DP*, the counterfactual 'If *C* had been 0, then *E* would have been 0', the technical equivalent of 'If neuron *c* had not fired, then neuron *e* would not have fired', is true. This can easily be verified by replacing the equation for *C* by the equation '$C \Leftarrow 0$', which sets *C* to the value specified

necessary and sufficient for us to be inclined to judge that there is causation. For our purposes only the sufficiency for causation is relevant.

[27] Hitchcock alludes to this result at 2007a: 513.

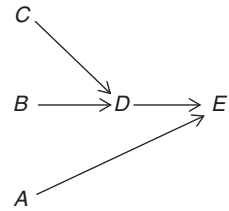

Figure 3.6. The causal graph of *DP*

in the antecedent of the counterfactual; substituting the specific values of *A, B*, and *C* in the remaining equations yields value 1 for *D* and, eventually, value 0 for *E*, which is the value specified in the consequent. Hence, in the model *DP*, variable *E* counterfactually depends on variable *C*. But the new qualification, namely that the network connecting the putative cause to the putative effect be self-contained, is not met. The network connecting *C* to *E* is the set {*C*, *D*, *E*}. This network fails to be self-contained irrespective of whether 0 or 1 is the default value for our variables, that is, irrespective of whether non-firing or firing is the default state of our neurons. Suppose that *A* and *B*, which are outside of the network connecting *C* and *E*, assume their actual values. Suppose further that the default value for our variables is 0. Then if *C* assumes the default value 0, *D* assumes the deviant value 1, so {*C*, *D*, *E*} is not self-contained. Suppose, on the other hand, that the default value for our variables is 1. Then, given that *A* and *B* both have their actual value 1, if *C* assumes the default value 1, *D* assumes the deviant value 0, so again {*C*, *D*, *E*} is not self-contained.

While the new principle about causation no longer implies that omissions and double preventers are causes, it still yields the verdict that, in our model *MC*, the headache is a cause of my hand's movement. What are the default and deviant values of the variables in that model? It turns out that we can leave this question open as long as we assume some degree of uniformity in what counts as a default and deviant value for our variables. Whether or not the default state for my headache, which is represented by variable *M*, is one where it occurs, it seems reasonable that occurrence is the default state for the headache if and only if the default states for its realizers, which are represented by variable *P*, are such that some realizer is instantiated. Similarly, it seems reasonable that occurrence is the default state for the movement of my hand, which is represented by variable P^*, if and only if the default state for the realizers is one where some realizer is instantiated.

https://doi.org/10.1017/9781108762717.005 Published online by Cambridge University Press

Thus, we get two cases: in case (i), the default values of *P, M*, and *P** are all 0; in case (ii), the default values of *P, M*, and *P** are all not 0; more precisely, any non-zero value of *P* is a default value, and the default values of *M* and *P** are both 1.[28]

The network connecting *M* to *P** is the set {*M, P, P**}.[29] Assume that the default and deviant values of our variables are as specified in case (i). Variable *P** has only one predecessor, namely *P*, which is in the network. By the equation for *P**, if *P* assumes its default value 0, then *P** assumes its default value 0. Variable *P* has one predecessor in the network, namely *M*, and one predecessor outside of the network, namely *U*. By the equation for *P*, if *M* assumes its default value 0, then *P* assumes its default value 0 regardless of the value of *U; a fortiori, P* assumes its default value 0 if *M* assumes its default value 0 while *U* assumes its actual value 1. Variable *M* has one predecessor, namely *P*, which is in the network. By the equation for *M*, if *P* assumes its default value 0, then *M* assumes its default value 0. Thus, in case (i), the network connecting *M* to *P** is self-contained.

Assume now that the default and deviant values of our variables are as specified in case (ii). In this case, if *P* assumes a default value, that is, a non-zero value, then *P** assumes its default value 1. If *M* assumes its default value 1, then *P* assumes a default value, that is, *P* is not 0, regardless of the value of *U; a fortiori, P* assumes a default value if *M* assumes its default value while *U* assumes its actual value 1. If *P* assumes a default value, that is, a non-zero value, then *M* assumes its default value 1. Thus, in case (ii) the network connecting *M* to *P** is self-contained as well.

In sum, the network connecting *M* to *P** is self-contained irrespective of how we assign default and deviant values to our variables (provided that we do it uniformly). Hence, if we do not simply take counterfactual dependence between variables to suffice for causation but restrict this sufficient condition for causation to cases where the network between the variable for the putative cause and the variable for the putative effect is self-contained, it still follows that *M* = 1 causes *P** = 1. Since the network that connects *P* to *P** is a subset of the network that connects *M* to *P**, it is self-contained as well. Therefore, *P*= 1 also still counts as a cause of *P** = 1.

[28] In case (i) *P* has multiple deviant values (namely all non-zero values), and in case (ii) *P* has multiple default values.

[29] Since *M* is a predecessor of *P* while *P* is also a predecessor of *M* in our causal model, there are many paths from *M* to *P** besides the path ⟨*M, P, P**⟩, for instance, ⟨*M, P, M, P, P**⟩, ⟨*M, P, M, P, M, P, P**⟩, etc. Still, the network connecting *P* to *P** just is the set {*P, M, P**}; multiple occurrences of *M* and *P* on cyclic paths do not make a difference to the identity of the network.

The second strategy of qualifying the sufficient condition for causation in terms of counterfactual dependence imposes no restriction on the network between the putative cause and the putative effect within a causal model, but instead restricts the conditions under which a counterfactual that expresses counterfactual dependence may be taken to entail causation. The idea is that, in order to entail causation, such a counterfactual must not take us to situations that are too abnormal.

To make this idea more precise, say that a setting of values of all the variables of a given model *manifests* the counterfactual dependence of variable Y on variable X if and only if this setting is a setting of variables we get in a model that is the result of replacing the equation for X with '$X \Leftarrow x'$', where x' is a non-actual value of X, and Y assumes a non-actual value in that model. Intuitively, a setting of variables that manifests the counterfactual dependence of Y on X represents a situation that would have been the case if Y had differed along with X.[30]

For instance, in the *Omission* example, there is one setting of the variables in the corresponding model *OM* that manifests the counterfactual dependence of D on G. In this setting A is 1, G is 1, and D is 0: Assassin poisons the drink; Bodyguard administers the antidote; and Victim survives. One might think that actions are always less normal than omissions. Since the variable setting that manifests the counterfactual dependence of D on G involves two actions (namely Assassin's and Bodyguard's) while the actual setting involves merely one (Assassin's), one might take this to show that the setting is too abnormal for Bodyguard's omission to be a cause of Victim's death (see Halpern and Hitchcock 2015: 439–441).

While it no longer seems to follow that Bodyguard's omission causes Victim's death, the new sufficient condition for causation in terms of counterfactual dependence plus normality, unlike the new condition in terms of self-contained networks, still seems to rule that double preventers are causes. If actions are less normal than omissions, presumably firings of neurons are less normal than non-firings. In the model for the double-prevention case, *DP*, the counterfactual dependence of E on C is manifested by a setting where C and E are 0 while A, B, and D are 1. In the actual situation, all variables except D are 1. In other words, if c had not fired, there would have been three firings and two non-firings, while there are four firings and one non-firing in the actual situation. Cumulatively, the situation that would have been the case if c had not fired is more normal

[30] Manifestation as defined here is a simplified version of what Halpern and Hitchcock (2015: 436) call being a 'witness' for a causal relation.

than the actual situation, which *prima facie* seems pretty normal itself. Since the counterfactual dependence of E on C is manifested by a comparatively normal situation, we may infer that $C = 1$ causes $E = 1$.

Admittedly, more can be said about the notion of normality and how it depends on the default or deviant values of individual variables.[31] The details of how normality is assessed need not concern us for the purposes of assessing the efficacy of mental events in the model *MC*, however. We shall see that the variable settings that manifest the counterfactual dependence of my hand's moving on my headache coincide with those that manifest the counterfactual dependence of the hand's moving on the actual realizer-instance. Thus, the headache is on a par with its actual realizer as far as normality considerations are concerned. Since the realizer seems to be a *bona fide* cause of the later physical event,[32] mental causation remains unscathed.

We saw in the previous section that P^*, the variable for my hand's moving towards the aspirin, counterfactually depends both on M, the variable for my headache, and on P, the variable for the realizers of headaches, because the following counterfactuals are true:

(15) If M had been 0, then P^* would have been 0.
(1) If P had been 0, then P^* would have been 0.

Counterfactuals (15) and (1) represent the only ways in which these counterfactual dependences can come about in our model. For (15), this is clear since M and P^* are binary and (we assumed) both have the actual value 1. While P is not binary, the only non-actual value that P can assume such that the value of P^* varies along with P is 0. Thus, the variable settings that manifest the counterfactual dependence of P^* on M and on P are both unique. Moreover, these settings are identical. The setting that manifests the counterfactual dependence of P^* on M can be calculated by replacing the equation for M with '$M \Leftarrow 0$' in *MC* (which yields the equations of model *MC′*); then the remaining equations yield that P and P^* are both 0. The setting that manifests the counterfactual dependence of P^* on P can be calculated by replacing the equation for P with '$P \Leftarrow 0$' in *MC*; then the

[31] See Halpern and Hitchcock 2010, §5; 2015: 433–436 for further discussion. One aspect that we will briefly return to in the following section is that of statistical normality. Notice that the normality dimension is standardly taken to be different from the dimension of overall similarity between worlds that is in play in Lewis's (1973b, 1979) semantics for counterfactuals. In particular, that the actual world is most similar overall to itself is standardly taken to be consistent with there being worlds that are more normal than the actual world.

[32] In the previous chapter, we saw that one might doubt that instances of realizers can in principle be causes or effects. These doubts will be addressed in Section 4.4.

remaining equations yield that M and P^* are both 0.[33] Thus, the setting where M, P, and P^* are all 0 is the one setting that manifests both the counterfactual dependence of P^* on M and the counterfactual dependence of P^* on P.[34] Since the causation of $P^* = 1$ by $P = 1$ does not seem problematic, this setting should be regarded sufficiently normal to license the inference that $P = 1$ causes $P^* = 1$. Consequently, the same setting should be regarded as sufficiently normal to license the inference that $M = 1$ causes $P^* = 1$. So the strategy of qualifying the sufficiency of counterfactual dependence for causation by requiring that the counterfactual dependence not take us to situations that are too abnormal poses no threat to my account of mental causation.

In this section we have considered two qualifications to the principle that counterfactual dependence between variables in a causal model suffices for a causal relation between (what is represented by) the values of these variables. One qualification is that the network between the variables in question be self-contained. The other qualification is that the counterfactual dependence be manifested in worlds that are not too abnormal. The unqualified principle rules certain omissions to be causes. With either qualification, this result no longer follows. The unqualified principle also rules certain cases of double prevention to be causes. With the self-containment qualification, this result no longer follows; with the normality qualification, it still follows. The most important result for our purposes is that either qualification still yields the result that mental events can have physical effects. Thus, mental causation can be robustly accommodated within the causal modelling framework.

[33] It might be objected that it is an artefact of our definition of manifestation that M assumes the value specified by its equation, despite being in an intuitive sense off the path from P to P^*. If we allowed a setting where M is 1 while P and P^* are 0 to manifest the counterfactual dependence of P^* on P, however, this setting is unlikely to count as more normal than one where M, P, and P^* are all 0. Non-reductive physicalists would deem such a setting metaphysically impossible. It would seem dubious if metaphysically impossible settings could be more normal than metaphysically possible ones even if we allowed the normality dimension of worlds to differ from the overall similarity dimension. Dualists would deem the setting metaphysically possible. But they might follow the lead of those theorists who take inactions to make for increased normality and actions to make for decreased normality by claiming that the non-occurrence of an event makes for increased normality and that the occurrence of an event makes for decreased normality. In sum, the occurrence of the mental event that is represented by a setting where M is 1 while P and P^* are 0 makes that setting *less* normal, or at any rate not more normal, than one where all three variables are 0.

[34] In this setting, variable U assumes its actual value 1, but even if we allowed U to vary, yielding different settings that manifest our counterfactual dependences, the set of such settings that manifest the counterfactual dependence of P^* on P would coincide with the set of settings that manifest the counterfactual dependence of P^* on M.

What lesson should we draw from the different verdicts about double prevention that the two qualifications of the sufficient condition for causation yield? We saw in Section 1.6 that a good case can be made for the claim that double preventers are causes. And we saw that, if there is no causation by double prevention, there is no human agency, because if there is no causation by double prevention, then there is no causation of muscle contraction by neural impulses. Friends of the self-containment qualification could try to argue that the default and deviant states of muscle fibres are somehow different from those of the idealized neurons of our double-prevention case. Or they could use a different sufficient condition for causation to accommodate the causation involved in muscle physiology. But they are not forced to pursue either strategy. Having advocated merely a sufficient condition for causation, they are not committed to denying a causal relation in cases of double prevention.

3.5 Overlapping Realizers Redux

This section investigates how causal modelling can help us solve the problem of overlapping realizers from Section 2.3. Recall the example that made trouble for the unqualified sufficient condition for causation in terms of counterfactual dependence between events: I hold an aluminium ladder against a power line and subsequently get electrocuted. Being made of aluminium, the ladder is both conductive and opaque. If the ladder had not been conductive, I would not have been electrocuted. But if the ladder had not been opaque, I would not have been electrocuted either. For if the ladder had not been opaque, presumably it would have been made of some standard transparent material that would have been non-conductive, in which case I would not have been electrocuted. Thus, by the unqualified sufficient condition for causation, the opacity-instance causes my electrocution. This result seems implausible. In this section I shall elaborate on two suggestions from Section 2.3 about how to respond to the problem of overlapping realizers. The first suggestion is to replace the unqualified sufficient condition for causation with a condition in terms of causal models that allows us to draw causal conclusions only if the actual values of certain variables are held fixed. We shall see that this suggestion is initially attractive, but ultimately problematic, because of difficulties in selecting an appropriate causal model. The second suggestion is to use the condition from the first suggestion not as a condition for causation, but as a condition for explanatory relevance. Used as a condition for explanatory relevance, the condition will allow us to formulate what seems, overall, the

best response to the problem of overlapping realizers, namely that the opacity-instance is a cause of the electrocution, but not one that is typically considered explanatorily relevant.

Both suggestions exploit the asymmetry between what would have been the case if the ladder had been opaque but not conductive and what would have been the case if the ladder had been conductive but not opaque that we observed in Section 2.3: if the ladder had been opaque but not conductive, then I would not have been electrocuted, whereas, if the ladder had been conductive but not opaque, then I would still have been electrocuted. We saw that, in order to elaborate this idea, we need a rationale for identifying the events to be held fixed in these comparisons. Such a rationale is needed because, for virtually any pair of events that are related by counterfactual dependence, we can find other events that actually occur and whose holding fixed makes no difference to the occurrence or non-occurrence of the dependent event, while we can also find other events that actually occur and whose holding fixed does make a difference to the occurrence or non-occurrence of the dependent event. For instance, if I had not thrown the dart and there had been just as many grains of sand on Mars as there actually are, then the balloon would not have burst, but if I had not thrown the dart and the dart had been on its actual trajectory a second later, then the balloon would still have burst. Intuitively, the relevant events to be held fixed seemed to be those that are not on a causal path between the putative cause and the putative effect; the dart's being on its actual trajectory a second after the time at which I actually threw it seems to lie on such a causal path from my throw to the balloon's bursting. We shall see that the causal modelling apparatus enables us to spell this idea out by using the definition of a path from the previous section.

But, first, let us construct a causal model for the electrocution example; call this model *EL*. Let *C*, *O*, and *E* be binary variables that stand for the ladder's being conductive, the ladder's being opaque, and my being electrocuted, respectively. Let *P* be a multi-valued variable that represents the microphysical makeup of the ladder. As in our model for mental causation, let us introduce an exogenous variable *U* that contributes to determining the value of *P* by representing initial conditions. Like *P, U* should be multi-valued. Let us use the following interpretation of the variables:

U = 0 or 1 or 2 or 3, depending on the initial conditions
P = 0 if the ladder instantiates a realizer of both conductivity and opacity (e.g., aluminium), 1 if the ladder instantiates a realizer of

https://doi.org/10.1017/9781108762717.005 Published online by Cambridge University Press

conductivity that is not also a realizer of opacity (that is, an exotic transparent conductor), 2 if the ladder instantiates a realizer of opacity that is not also a realizer of conductivity (e.g., wood), 3 if the ladder instantiates neither a realizer of conductivity nor a realizer of opacity (e.g., if the ladder is made of transparent plastic)

C = 1 if the ladder is conductive, 0 otherwise

O = 1 if the ladder is opaque, 0 otherwise

E = 1 if I am electrocuted, 0 otherwise

Given that my electrocution counterfactually depends both on the ladder's conductivity and on the ladder's opacity, the following counterfactuals should be true in our causal model for the case:

(16) If C had been 0, then E would have been 0.

(17) If O had been 0, then E would have been 0.

As we saw, the case is not symmetric between C and O. If the ladder had been conductive but not opaque, then I would still have been electrocuted:

(18) If C had been 1 and O had been 0, then E would have been 1.

On the other hand, if the ladder had been opaque but not conductive, then I would not have been electrocuted:

(19) If O had been 1 and C had been 0, then E would have been 0.

What is the counterfactual relation between C and O on the one hand, and P on the other? From the way in which P was characterized, we can easily read off the values that C and O would have assumed if P had assumed a specific value. For instance, C would have been 0, but O would have been 1 if P had been 2 (that is, the ladder would have been opaque but not transparent if it had been made of a material like wood). Conversely, specific values of C and O rule out certain values of P. For instance, if O had been 0, then P would not have been 0 or 2 (that is, if the ladder had not been opaque, then it would not have been made out of a material like aluminium or a material like wood). Indeed, in light of the results from Section 2.3, we can be more specific. We saw that it seems plausible that if the ladder had not been opaque, it would have been made out of a middle-of-the-road, non-conductive transparent material like transparent plastic rather than an exotic transparent conductor. Thus, we have:

(20) If O had been 0, then P would have been 3.

Similarly, it seems plausible that if the ladder had not been conductive, it would have been made out of a material like wood rather than a material like transparent plastic:

(21) If C had been 0, then P would have been 2.

As in the model MC, the value of U should set the value of P, provided there would be no conflict with the values of C and O. Thus, we can think of $U = 0$ as representing a state that is standardly followed by the ladder's being made of aluminium, of $U = 1$ as representing a state that is standardly followed by the ladder's being made of an exotic transparent conductor, etc. If there is a conflict between U on the one hand and C and O on the other, it must be resolved in favour of C and O. For instance, if U had been 1, then P would have been 1, but if U had been 1 and C had been 0, then P would have been 2. That is, if the initial state had been one that is standardly followed by the ladder's being made out of an exotic transparent conductor, then the ladder would have been made out of an exotic transparent conductor, but if the initial state had been one that is standardly followed by the ladder's being made out of an exotic transparent conductor and the ladder had failed to be conductive, then the ladder would have been made out of a material like wood.

For our purposes, we need not write down the equations of EL; drawing the graph will suffice. The graph looks like this: there is an arrow to E from P, but from no other variable, since the value of P makes a difference to the value of E, but no other variable makes a difference to E over and above the difference made by P.[35] There is an arrow from P to C. There is also an arrow from C to P, owing to the truth of (21).[36] Similarly, there is an arrow from P to O, and there is an arrow from O to P, owing to the truth of (20). In sum, we get the graph depicted in Figure 3.7.

The graph of EL shows that there is one acyclic path from C to E, which does not include O (namely the path $\langle C, P, E\rangle$), and one acyclic path from O to E, which does not include C (namely the path $\langle O, P, E\rangle$).[37]

[35] Difference-making is supposed to be captured by counterfactuals that are non-vacuously true. The fact that, say, $P = 0 \,\&\, C = 0 \,\square\!\!\rightarrow E = 0$ is vacuously true while $P = 0 \,\square\!\!\rightarrow E = 0$ is false is not a reason for including C in the equation for E.

[36] If one finds (21) implausible, one could still justify the inclusion of C in the equation for P, and hence the arrow from C to P in the causal graph, from the truth of the counterfactual $C = 0 \,\square\!\!\rightarrow P = 2 \vee P = 3$.

[37] These acyclic paths overlap with various cyclic paths with the same starting-points and ends. For instance, the acyclic path $\langle C, P, E\rangle$ overlaps with the cyclic path $\langle C, P, C, P, E\rangle$.

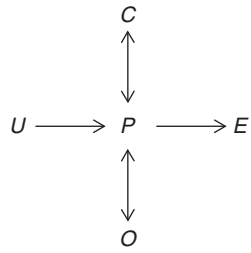

Figure 3.7. The causal graph of *EL*

The counterfactual dependence of *E* on *O* as well as the counterfactual dependence of *E* on *C* was built into our model *EL*, because we stipulated that counterfactuals (17) and (16) be true in the model. The actual values of *C*, *O*, and *E* all stand for the occurrence of genuine events that are not omissions, namely the ladder's being opaque, the ladder's being conductive, and my electrocution. Thus, it follows from the simple sufficient condition for causation in terms of counterfactual dependence between variables that the opacity-instance as well as the conductivity-instance causes the electrocution. This result does not come as a surprise, of course, but merely mirrors the situation outside of the causal modelling context that we investigated in Section 2.3. The question is whether we can use the resources of the causal modelling framework in order to avoid the result that the opacity-instance causes the electrocution. The graph of *EL* and its path-structure corroborates the idea that what matters is not counterfactual dependence as such, but counterfactual dependence when off-path variables are held fixed at their actual values. However, before discussing in detail a qualification for our sufficient condition for causation along these lines, let us briefly investigate how the qualifications discussed in the previous section deal with the electrocution example.

The first qualification was that counterfactual dependence is sufficient for causation only in a self-contained network. Variable *E* counterfactually depends both on *C* and on *O*. The network connecting *C* to *E* is the same as the network connecting *O* to *E*, namely the set {*C*, *P*, *O*, *E*}. If this network is self-contained, the suggestion yields that both $C = 1$ and $O = 1$ are causes of $E = 1$. If the network is not self-contained, the suggestion remains neutral with respect to either putative cause. Thus, the suggestion does not help us to discriminate between *C* and *O*.[38]

[38] It seems that in fact the network {*C*, *P*, *O*, *E*} is not self-contained. It is a bit unclear how to assign default and deviant values to the variables, but presumably 1 is the deviant value for *E*, while the

The second qualification was that counterfactual dependence suffices for causation only if the situations that manifest the counterfactual dependence do not take us to worlds that are too abnormal. If the ladder had not been conductive, then (i) it would have been made out of a material like wood, in which case it would still have been opaque and I would not have been electrocuted. If the ladder had not been opaque then (ii) it would have been made out of a material like transparent plastic, in which case it would not have been conductive and I would not have been electrocuted. Given that there are few transparent plastic ladders while wooden ladders abound, case (ii) counts as less normal than case (i), at least statistically speaking. This might explain why we take the conductivity-instance to be a 'better' cause than the opacity-instance.

Given the normality qualification, the opacity-instance still qualifies as a good enough cause, however. We saw in the discussion of the *Omission* example that the normality qualification takes actions to be less normal than omissions. Although it might sound a bit odd to say that my electrocution is an action, it seems that my electrocution should similarly count as less normal than my non-electrocution. If so, the aspect of case (ii) in which it is less normal than the actual situation, namely the ladder's being made out of a material like transparent plastic, would be offset by the aspect in which it is more normal than the actual situation, namely my failure to be electrocuted. Overall, case (ii) should come out about as normal as the actual situation, so nothing prevents us from taking the opacity-instance to be a cause of the electrocution. Moreover, there are cases where an opacity-instance is a *bona fide* cause of a later event, so the situation that would have obtained if the ladder had not been opaque cannot by itself be too abnormal. Suppose that, in our set-up, the ladder casts a shadow. Had it not been opaque, it would not have cast a shadow; hence, the ladder's being opaque causes the shadow. There seems to be nothing wrong with this reasoning, but here too the counterfactual dependence is manifested in a situation where the ladder is made out of a material like transparent plastic.

Let us turn to the suggestion of formulating a sufficient condition for causation in terms of paths. While E counterfactually depends both on C and on O, C still makes a difference to E if we hold O fixed at its actual value, but O does not make a difference to E if we hold C fixed at its actual value. The causal graph of *EL* suggests a rationale for taking these counterfactuals to indicate that the conductivity-instance causes the electrocution,

actual values are (among the) default values of C, P, and O. Thus, the actual distribution of values in the network is a counterexample to its being self-contained.

while not taking them to indicate that the opacity-instance causes the electrocution: *E* varies along with *C* if we hold the variables that are off the acyclic path from *C* to *E* fixed at their actual values, while *E* does not vary along with *O* if we hold the variables that are off the acyclic path from *O* to *E* fixed at their actual values. What matters is whether a variable makes a difference to another if we hold the off-path variables fixed at their actual values; if it does, we may infer a causal relation.

Thus, we can formulate a new sufficient condition for causation as follows: let *X* and *Y* be binary variables in an appropriate causal model and let the actual values of *X* and *Y* (assume they are both 1) represent the occurrence of property-instances such that the property-instance represented by '$Y = 1$' occurs later than the property-instance represented by '$X = 1$'. Let both '$Y = 1$' and '$X = 1$' represent (strong Kimian) events that are not omissions. If '$X = 0 \; \Box\!\!\rightarrow Y = 0$' is true in that model and '$X = 0$ & FIX $\Box\!\!\rightarrow Y = 0$' is true in that model,[39] where FIX stands for the claim that all variables that are not on any acyclic path from *X* to *Y* are held fixed at their actual values, then the property-instance represented by '$X = 1$' causes the property-instance represented by '$Y = 1$'.[40] (Henceforth, when talking about off-path variables, I shall mean variables that are not on any *acyclic* path between the variables in question.)

The new sufficient condition for causation not only seems to handle the electrocution example well, it also still allows us to establish mental causation in the model *MC*. In that model, there is one acyclic path from the variable for my headache, *M*, to the variable for my hand's moving towards the aspirin, *P**, namely the path $\langle M, P, P^* \rangle$. The only variable in the model that is not on this path is variable *U*, which represents the initial conditions. The actual values of the variables in the model are all 1. Thus, when we apply the condition, FIX is the claim that *U* has value 1. It is true in the model that if *M* were 0, then *P** would be 0; this is just the counterfactual

[39] Since ϕ & $\chi \; \Box\!\!\rightarrow \psi$ does not logically imply $\chi \; \Box\!\!\rightarrow \psi$ (because ψ can be true in the closest ϕ-&-χ-worlds without being true in the closest χ-worlds *simpliciter*), '$X = 0$ & FIX $\Box\!\!\rightarrow Y = 0$' can be true while '$X = 0 \; \Box\!\!\rightarrow Y = 0$' is false; hence '$X = 0 \; \Box\!\!\rightarrow Y = 0$' is not redundant here.

[40] Hitchcock 2001 defends the view that it is a necessary and sufficient condition for causation that there is an *active causal route* between the putative cause variable and the putative effect variable. In our terminology a route is a kind of path. For there to be an active causal route between variables *X* and *Y* it is necessary and sufficient that the value of *X* makes a difference to the value of *Y* if all variables that are on other paths between *X* and *Y* are held fixed at their actual values (Hitchcock 2001: 286–287). Hitchcock's condition is different from the new sufficient condition for causation under discussion, for the latter does not require variables on alternative paths from *X* to *Y* to be held fixed, and the former does not require variables that are *not* on any path from *X* to *Y* to be held fixed. Hitchcock's condition is similar to the spirit behind our informal diagnosis of the example involving the apprentice and expert assassins in Section 3.2, however.

dependence of variable P^* on variable M. It is also true in the model that if M were 0 and FIX were the case, then P^* would be 0. Thus, it follows from the new sufficient condition that $M = 1$ causes $P^* = 1$; less technically, it follows that my headache causes my hand to move towards the aspirin.

Unfortunately, the new sufficient condition for causation has a number of disadvantages. First, although it delivers the verdict that the conductivity-instance and the mental property-instance are causes while not delivering the verdict that the opacity-instance is a cause, it no longer delivers the verdict that the realizer-instance is a cause of the electrocution and the later physical event. There are several reasons for this. To start with, the realizer-variable P in MC and EL fails to be binary, but the new sufficient condition can be applied only to binary variables for the putative causes and effects. Technically, this can easily be rectified: we can allow multi-valued cause and effect variables X and Y and simply demand that there be *some* non-actual value of X which would have yielded *some* non-actual value of Y, holding the off-path variables fixed at their actual values. Now, in EL, we cannot vary P at all while holding the off-path variables fixed at their actual values. The only acyclic path from P to E is $\langle P, E\rangle$. By the characterization of P, it is metaphysically impossible for P to vary while C and O are held fixed at their actual values. The counterfactual '$P = p$ & FIX $\Box\!\!\rightarrow E = 0$' is *vacuously* true for some non-actual value p of P, but taking this to imply a causal relation is certainly not in the spirit of the present suggestion. In MC, the only acyclic path from P to P^* is $\langle P, P^*\rangle$. Holding M fixed at its actual value 1 while varying P is metaphysically possible for some variations of P, but those variations do not yield a change in the value of P^*. Setting P to 0 would change P^* to 0, but again we run up against a metaphysical impossibility, namely that of P's being 0 while M is 1.

We can modify the new sufficient condition so that these metaphysical impossibilities no longer bar the efficacy of the realizer-instance of O/C and M. We can stipulate that off-path variables need to be held fixed only if their values are not necessitated by the value of the (putative) cause-variable.[41] Then $P = 1$ still comes out as a cause of $P^* = 1$ in MC, and $P = 1$ still comes out as a cause of $E = 1$ in EL. But even if we did not make this modification, the situation would at most be unfortunate, not untenable. Our new condition that requires off-path variables to be held fixed still is merely a sufficient condition for causation. It may remain silent on whether a case involves causation as long as it does not

[41] See Woodward 2015 for a similar suggestion. Alternatively, one could let P^* be multi-valued, but, even setting the problem from note 14 aside, this would have the desired result only in MC, not in EL.

Figure 3.8. The causal graph of *EL′*

diagnose causation where there is none. That the realizer-instances of the mental property and of conductivity/opacity are causes is plausible in any case and does not require a principled argument.

The second and more pressing problem with the new sufficient condition is that it uses the notion of an *appropriate* causal model. By itself, this is not an unusual requirement. Causal models are standardly required to be appropriate in the sense of satisfying certain minimal standards of model-building. In the new sufficient condition, however, the notion of appropriateness carries a lot of weight – indeed, too much weight.

There are several alternative causal models of the electrocution case where, by the new sufficient condition, the opacity-instance causes the electrocution. For instance, take a simple model, call it *EL′*, that includes only the variables *O* and *E*. Figure 3.8 shows the causal graph of *EL′*.

In *EL′*, there are no variables that are off the path from *O* to *E*. Hence it is trivially true that the value of *O* makes a difference with respect to the value of *E* if all off-path variables are held fixed at their actual values. Similarly for a model, call it *EL″*, that is like the original model *EL* but does not contain *C*. As Figure 3.9 shows, there is an off-path variable in *EL″*, namely *U*, but *E* still varies along with *O* if that off-path variable is held fixed.

It might seem that the result that the opacity-instance causes the electrocution can be avoided as soon as we include a variable for the ladder's conductivity in our model, but this is not the case. Take a model, call it *EL‴*, that is just like the original model *EL* but does not contain a variable for the physical realizer. In *EL‴* the value of *O* makes a difference to the value of *E*, but this difference is nothing over and above the difference made by the value of *C*. So *C* is on the path from *O* to *E* (see Figure 3.10), and again *E* varies along with *O* if the off-path variable *U* is held fixed.[42]

[42] The double-arrow between *C* and *O* in the causal graph of *EL‴* comes about as follows. Given the truth of (20), we have $O = 0 \; \Box\!\!\rightarrow P = 3$. By the characterization of variable *P*, we have $\Box[P = 3 \supset C = 0]$. From these two claims, $O = 0 \; \Box\!\!\rightarrow C = 0$ follows logically (see Section 1.4). Similarly, given the truth of (21), we have $C = 0 \; \Box\!\!\rightarrow P = 2$. Together with $\Box[P = 2 \supset O = 1]$ we get $C = 0 \; \Box\!\!\rightarrow O = 1$. Thus, *O* occurs in the equation for *C* and vice versa. The derivations of

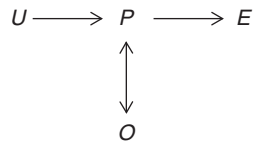

Figure 3.9. The causal graph of *EL″*

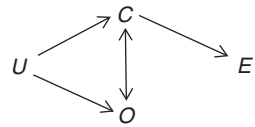

Figure 3.10. The causal graph of *EL‴*

Thus, in order to avoid the result that the opacity-instance causes the electrocution, we have to read the requirement that the causal model be appropriate such that a causal model for the case it not appropriate unless it includes a variable for the realizer-instance *and* a variable for the conductivity-instance.

These are very strong requirements of appropriateness for our case, and they smack of being *ad hoc*, of being tailored to avoiding the result that the opacity-instance causes the electrocution. (To avoid misunderstanding: I think that it *is* appropriate to include a variable for the realizers of supervenient properties. Indeed, I will defend this claim in the following section. What strikes me as too strong is the claim that it is *in*appropriate *not* to represent the realizers or, for that matter, the conductivity.)

Could we formulate the new sufficient condition for causation without invoking this overly strong notion of an appropriate causal model? We could say that it suffices for causation that the value of one variable makes a difference to the value of another if we hold fixed all off-path variables in *some* model. This suggestion is clearly a non-starter, for it would still yield the result that the opacity-instance causes the electrocution owing to the condition's being satisfied by $O = 1$ and

$O = 0 \,\square\!\!\rightarrow C = 0$ and $C = 0 \,\square\!\!\rightarrow O = 1$ that I just gave used variable *P*, which is not contained in *EL‴*, as a logical intermediary. This is harmless, because the realizers of conductivity and opacity can play a role in deriving certain true counterfactuals which do not themselves talk about those realizers. Among these counterfactuals, those that talk about the relation between *C* and *O* should be true in *EL‴*. An alternative way of establishing $O = 0 \,\square\!\!\rightarrow C = 0$ is via claim (5-O) from Section 2.3.

https://doi.org/10.1017/9781108762717.005 Published online by Cambridge University Press

$E = 1$ in simple alternative models like *EL'*. Alternatively, we could say that it suffices for causation that the value of X makes a difference to the value of Y if we hold fixed all off-path variables in *all* models with variables corresponding to X and Y. Then it would no longer follow that the opacity-instance causes the electrocution-instance, because the condition fails to be satisfied by $O = 1$ and $E = 1$ in the original model *EL*. It would, however, be virtually impossible to apply the condition in practice. So far, causal modelling theorists have constructed *a* model for a given case and investigated what is true in it. Investigating what is true in *all* models that contain a certain pair of variables is a task of a very different kind, and *prima facie* it does not look very promising. Lastly, we could say that it suffices for causation *in a given model* if the value of one variable makes a difference to the value of another if we hold fixed all off-path variables *in that model*. Then, however, causation would be model-relative and, it seems, cease to be an objective feature of the world. This seems to be too high a price to pay for solving the problem of overlapping realizers.[43]

Owing to its troublesome model-relativity, the condition of dependence while holding off-path variables fixed fails to solve the problem of overlapping realizers when it is used as a sufficient condition for causation. It can be used in a different and more promising way, however. Recall the final suggestion for solving the problem of overlapping realizers from Section 2.3: according to this suggestion, the opacity-instance is a cause of the electrocution, but one that, unlike the conductivity-instance, has little explanatory relevance in our context. The causal modelling framework allows us to spell out this suggestion in more detail by formulating principles about explanatory relevance. As was the case with causation, these principles will fall short of constituting a full-blown theory of explanatory relevance, but we shall see that together they are still strong enough to solve our problem. The basic idea, which I will elaborate in the remainder of this section, is that standard counterfactual dependence is a defeasible sufficient condition for explanatory relevance among whose defeaters is the failure to satisfy the condition of dependence while holding off-path variables fixed.[44]

[43] Van Fraassen (1980: Ch. 5) holds that causation is a context-dependent relation. Newen and Čuplinskas (2002) advocate an account of (mental) causation that draws on an interest-relative notion of *events*. Hitchcock (2003) holds that causation is an objective matter, but that there is no unique causal relation. Menzies (2004) holds that causation is relative to a causal model, but he uses 'causal model' in a sense that is different from the present one. For further discussion, see Price and Corry 2007 and Ismael 2016.

[44] The same general strategy, viz. that of distinguishing causation from explanatory relevance in a causal modelling framework, is advocated by Woodward (2010).

Suppose that we are dealing with two strong Kimian events (that are not omissions) such that one event occurs later than the other. Suppose that we have a causal model that accurately represents the counterfactuals that are true about the two events; let '$X = 1$' represent the occurrence of the earlier event in the model and '$Y = 1$' the occurrence of the later event, where X and Y are binary variables. Suppose, lastly, that variable Y counterfactually depends on variable X in the model. In such a situation, $X = 1$ causes $Y = 1$ by our simple sufficient condition for causation in the causal modelling framework. (Unlike the sufficient condition that required holding off-path variables fixed, the simple sufficient condition does not introduce any problematic model-relativity, because the counterfactual that underlies the counterfactual dependence of Y on X is true in any model that accurately represents the counterfactuals that are true about the corresponding events.) That such a situation obtains, I suggest, is also a defeasible sufficient condition for its being the case that $X = 1$ is explanatorily relevant to $Y = 1$. Typically, counterfactual dependence is indicative not only of causation but also of explanatory relevance.[45] My throwing the dart, for instance, is not merely a cause of the balloon's bursting; it is also a cause that explains its effect.

In some cases, however, other factors defeat counterfactual dependence as a sufficient condition for explanatory relevance (though not as a sufficient condition for causation). I will not attempt to give an exhaustive list of such factors, but one of them is excessive temporal distance in the absence of a thing or feature that persists. My bumping into Albert, for instance, is among the causes of Berta's death, because her death, like her birth, counterfactually depends on it. But the bumping is too far removed in time. And it does not create a thing or feature that persists until the death, unlike, say, the actions of an artist whose paintings both cause and explain the viewers' delight despite the fact that the artist died long ago.

Another factor that defeats counterfactual dependence as a sufficient condition for causation, I suggest, is the failure to satisfy the condition of dependence while off-path variables are held fixed. Thus, in the electrocution case as it is represented in the model *EL*, the opacity-instance causes the electrocution owing to the counterfactual dependence of the electrocution-variable on the variable for the ladder's opacity. But the opacity-instance does not explain the electrocution if model *EL* is used, because the

[45] See Swanson 2010 and note 16 of Chapter 2. Ney (2012) holds that we ordinarily classify something as causation on the basis of counterfactual dependence, but denies that this is in conflict with 'thick' notions of causation such as transfer accounts. In light of the double-prevention cases discussed in Section 1.6, however, this denial is hard to defend.

electrocution-variable no longer depends on the opacity-variable if the variable for the ladder's conductivity is held fixed at its actual value.

Whether the condition of dependence while holding off-path variables fixed is satisfied depends on the causal model that is being considered. Although relativity to a model is implausible for conditions on causation, it is not a problem for conditions on explanatory relevance, for explanation is sensitive to context, and the context is in turn partly constituted by the causal model that is being considered.[46] Thus, when we do not consider the ladder's conductivity and model the electrocution example by the causal model *EL′*, say, which represents only the ladder's opacity and the electrocution, the opacity-instance qualifies as both a cause of the electrocution and a cause that explains the effect. (I concede that it is difficult to get into the mindset of not considering the conductivity once one has considered it, as we have. This is a common phenomenon for certain kinds of context-shift, however.) When we consider the ladder's conductivity as well as its opacity in the original model *EL*, the opacity-instance is still a cause of the electrocution, but opacity's claim to explanatory relevance is defeated by the failure of the electrocution variable to depend on the opacity variable when the conductivity variable is held fixed at its actual value. By contrast, if model *EL* is considered, the conductivity-instance is not simply a cause, but one that is explanatorily relevant, because conductivity's claim to explanatory relevance, which is due to the counterfactual dependence of the electrocution variable on the conductivity variable, is not defeated by the failure of the off-path condition.[47]

Mental causes remain explanatorily relevant if the condition of dependence while holding off-path variables fixed is used as a criterion of explanatory relevance. For we saw that their physical effects counterfactually depend on them, and still depend on them when the off-path variables are held fixed. Thus, using the causal modelling framework to formulate conditions for causation as well as conditions for explanatory relevance allows us

[46] For the context to be thus constituted, one need not consider a causal model *per se* (that is, under the mode of presentation of variables, equations, etc.) as long as one represents the relevant counterfactual structure.

[47] The model *EL‴* yields a *prima facie* difficulty for the suggestion that the condition of dependence while holding off-path variables fixed is a condition on (model-relative) causal explanation. In *EL‴*, both conductivity and opacity are represented and both satisfy the condition of dependence while holding off-path variables fixed. If *EL‴* represents the context, the opacity should count as an explanatorily relevant cause of the electrocution despite the fact that the conductivity is also in play. One way to respond is to say that a context in which both the opacity and the conductivity are relevant and their counterfactual relationship is assessed accurately is most likely also a context where the overlapping realizers of opacity and conductivity are salient, so the model that should be used is *EL*, not *EL‴*.

to accommodate mental causes and their explanatory relevance, while also offering an attractive solution to the problem of overlapping realizers.

3.6 Objections and Replies

Let us return to our model for mental causation, *MC*. We have seen that, if we choose *MC* as a causal model, we can capture how mental events can have physical effects, either by using straightforward counterfactual dependence as a sufficient condition for causation or by using a more complex sufficient condition that invokes the default/deviant distinction or a normality criterion. It might be objected, however, that *MC* is in some respect inappropriate as a causal model. In this section, I address three such objections: the objection that having a variable for the headache as well as a variable for its realizers violates a constraint on the independence of variables in a model; the objection that the different values for the realizer-variable do not represent versions of the same event; and the objection that the role of the exogenous variable in our model is dubious.

Here is the first objection. Given non-reductive physicalism, the connection between a mental event and its realizers is metaphysically necessary. Specifically, by claims (3)–(6) from Section 3.3, necessarily, variable M is 0 if and only if variable P is 0, because, necessarily, there is no headache just in case no realizer of headaches is instantiated. Similarly, it is necessary that M is not 0 if and only if P is not 0, because, necessarily, there is a headache just in case a realizer of headaches is instantiated. It is sometimes claimed by causal modelling theorists that there should be no metaphysically necessary connections between the values of different variables.[48] If we do not impose this constraint, they hold, we get spurious cases of causation. In our model, for instance, by (9) and (10), M would have been 0 if P had been 0, and P would have been 0 if M had been 0. Given that the actual values of our variables are all 1, it follows that M counterfactually depends on P and that P counterfactually depends on M. But it seems that we should not say that $P = 1$ causes $M = 1$ or that $M = 1$ causes $P = 1$. In other words, it seems that we should not say that the instance of the actual realizer of my headache, the c-fibre firing, causes my headache or that the headache causes the c-fibre firing.

If we deny any causal relation between the property-instances represented by M and P, we need to impose some kind of restriction. One possible restriction is a ban on causal models where some values of different variables are related by metaphysical necessity (call such variables *metaphysically*

[48] See, for instance, Hitchcock 2007a: 502 and Halpern and Hitchcock 2010, §4.3.

dependent). But there are two alternatives. When the principle about causation in terms of counterfactual dependence was introduced outside of the causal modelling context in Section 1.4, it was restricted to cases where the putative effect occurs after the putative cause. We can impose the same restriction on the corresponding principle about counterfactual dependence in a model. The other alternative is to restrict the principle about counterfactual dependence in a model to putative cause and effect variables that are metaphysically independent. (Both alternatives could be implemented similarly if one endorsed one of the qualifications that were discussed in Section 3.4.)

Imposing either of these alternative restrictions on the sufficient condition for causation allows us to represent the structure of our case more perspicuously than using models that have been purged of metaphysically dependent variables would. We could in principle ban either *M* or *P* from our model and preserve the counterfactual dependence of *P** on *P* and on *M*, respectively, but the new models would be much impoverished.[49] So we should restrict the sufficient condition for causation rather than ban models with metaphysically dependent variables.[50]

Faced with the choice between the restriction of the principle about causation to metaphysically independent variables and the restriction that the putative cause variable represent an event that occurs earlier than the putative effect variable, we should choose the temporal restriction. Given non-reductive physicalism, the temporal restriction entails the restriction to (putative) cause and effect variables that are not metaphysically dependent. At least it does if, as in Section 1.5, we confine ourselves to instances of properties that are temporally intrinsic, for there cannot be a metaphysically necessary connection between properties that are temporally intrinsic and instantiated at different times. Thus, the temporal restriction achieves what we want in the case of non-reductive physicalism. It has the advantage of also dealing nicely with the dualist case. If dualism is true, variables *M* and *P* are no longer metaphysically dependent, since dualists take it to be metaphysically possible for the mental event to occur without any of its bases, and metaphysically possible for any such base to occur without the mental event. Still, as we saw in Section 3.3, dualists can endorse all the counterfactuals that are true in our model *MC* if

[49] Hitchcock demands that an appropriate causal model 'include enough variables to capture the essential structure of the situation being modeled' (2007a: 503). For further discussion of variable choice, see Woodward 2016.

[50] Another restriction on variables that one could demand would specify admissible and inadmissible total settings of the endogenous variables in a given model (see Halpern and Pearl 2005: 869–870); one could thus declare metaphysically impossible combinations of values inadmissible. This would not help with the present problem, however, because the settings that manifest the counterfactual dependence between *M* and *P* are metaphysically possible.

they adopt super-nomological dualism. In particular, they can endorse counterfactuals (9) and (10), which express the counterfactual dependence of *P* on *M*, and vice versa. The mental property and its base are instantiated at the same time, so restricting the sufficient condition for causation to (variables that represent) events that occur one after the other avoids commitment to a simultaneous causal relation between the instances of the mental property and the instances of its base.[51] (The same result could be achieved by banning variables that represent simultaneous events from models instead of restricting our sufficient condition for causation, but as Max Kistler (2013: 73) points out, such a ban would have the disadvantage of disallowing models where the variable for a later event depends on the variables for two earlier events which are simultaneous yet mutually counterfactually and causally independent.)

The second objection to using the causal model *MC* to model mental causation concerns the multi-valued variable *P* from the model. The model *MC* uses different non-zero values of *P* to represent the instantiations of different realizers of the mental property. It might be objected that this violates a constraint on causal modelling, namely the constraint that the various values of non-binary variables represent different versions of the same event and not the occurrence of entirely different events.[52] The different realizers of our mental event are very dissimilar to one another. In Section 3.3, we illustrated the different values of *P* by assuming that $P = 1$ represents my having firing c-fibres, that $P = 2$ represents my having firing x-fibres, that $P = 3$ represents my having an active semiconductor network of a certain kind in my head, etc. It might be held that there is no single event of which all these different realizer-instantiations are versions. (Our objector might concede that all the realizers of the mental property fall under the description 'being a realizer of such-and-such a mental property', but deny that this description corresponds to a property whose instances are genuine events.)

As we saw in Section 1.3, the individuation of events is a complicated matter. Therefore I will not try to refute the claim that the different values of *P* do not represent versions of a single event. Instead I will show that we could modify our causal model such that it no longer involves any suspicious non-binary variables. It will turn out that the new model comes at a price that does not justify the benefits, however, so we are better off with the original model.

[51] For further discussion of simultaneous causation, see Fenton-Glynn and Kroedel 2015. Kistler (2013) argues against the existence of causal relations in cases where the values of different variables are related by synchronic (classical) association laws.

[52] See Hitchcock 2007a: 499. For further discussion of the relation between variables and events, see Hitchcock 2012a.

Instead of using one non-binary variable to represent all the realizers, we can use a binary variable for each realizer. Thus, we get:

P_1 = 1 if a c-fibre firing occurs, 0 otherwise
P_2 = 1 if an x-fibre firing occurs, 0 otherwise
P_3 = 1 if a certain semiconductor network is active, 0 otherwise
⋮

If we want to mimic the counterfactual relations between the single realizer-variable P and the other variables that held in our old model *MC*, we can formulate the following equations:

$$
\begin{aligned}
MC^* \quad & U \Leftarrow 1 \\
& P_1 \Leftarrow 1 \text{ if } U \text{ is } 0 \text{ and } M \text{ is } 1,\ M \cdot U \text{ if is } 1,\ 0 \text{ otherwise} \\
& P_2 \Leftarrow (M \cdot U)/2 \text{ if } U \text{ is } 2,\ 0 \text{ otherwise} \\
& P_3 \Leftarrow (M \cdot U)/3 \text{ if } U \text{ is } 3,\ 0 \text{ otherwise} \\
& \vdots \\
& M \Leftarrow \mathrm{Max}\{P_1, P_2, P_3, \ldots\} \\
& P^* \Leftarrow \mathrm{Max}\{P_1, P_2, P_3, \ldots\}
\end{aligned}
$$

Figure 3.11 shows the causal graph that corresponds to the equations *MC**.

Although the new model *MC** still verifies counterfactual (15) and thus still establishes that P^* counterfactually depends on M, it has at least two disadvantages. First, it is much more complex than the old model *MC*. That model has four variables; *MC** has as many variables as there are possible realizers of the mental event, plus another three. Correspondingly, while *MC* has four equations, *MC** has as many equations as there are possible realizers of the mental property, plus another three. Second, there is massive metaphysical dependence between the variables for the different realizers. Presumably, it is metaphysically impossible for something to be a c-fibre firing and also to be an x-fibre firing, metaphysically impossible for something to be an x-fibre firing and also to be the activity of a semiconductor network, etc. Thus, settings of variables where P_1 and P_2 are both 1, settings where P_2 and P_3 are both 1, etc. are metaphysically impossible. Hence it is metaphysically necessary that P_2 is 0 if P_1 is 1, metaphysically necessary that P_3 is 0 if P_2 is 1, etc. So there is metaphysical dependence between P_1 and P_2, between P_2 and P_3, etc.

We saw above that metaphysically dependent variables are not problematic *per se* and should be tolerated if they are necessary to represent the

https://doi.org/10.1017/9781108762717.005 Published online by Cambridge University Press

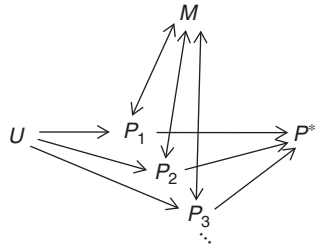

Figure 3.11. The causal graph of MC^*

structure of a case. Nevertheless, we should not multiply them beyond necessity. Our original model *MC* has just two metaphysically dependent variables; in MC^* they are legion. Moreover, MC^* contains multiple pairs of metaphysically dependent variables irrespective of whether non-reductive physicalism or dualism is assumed, since there is metaphysical dependence among the different realizer-variables. (Given non-reductive physicalism, there is also metaphysical dependence between *M* and each of the realizer-variables, as was the case in the original model *MC*.) So whatever the benefit of banning multi-valued variables that fail to represent different versions of the same event, it is outweighed by the added complexity and massive metaphysical dependence between variables that we need to accept if we implement this ban.

The third and final objection claims that complexity considerations in fact tell against our model *MC*. This model contains an exogenous variable *U*. It might be held that variable *U* is dispensable. Moreover, it might be held that *U* is to blame for the failure of our equations to be acyclic.

We can indeed simplify our model by removing *U*. This can be done in two obvious ways, depending on the new role that is assigned to *P*. Both ways, however, yield new problems that are not worth the gain in simplicity.

The first way of removing *U* delegates the status of being an exogenous variable to *P* instead. Thus, *P* is now set independently of the other variables. Given that the actual value of *P* is 1, we get the following equations:

$$\begin{aligned} MC\backslash U_{\mathrm{EX}} \quad & P \Leftarrow 1 \\ & M \Leftarrow \mathrm{Min}\{P, 1\} \\ & P^* \Leftarrow \mathrm{Min}\{P, 1\} \end{aligned}$$

Figure 3.12 shows the causal graph corresponding to $MC\backslash U_{\mathrm{EX}}$.

https://doi.org/10.1017/9781108762717.005 Published online by Cambridge University Press

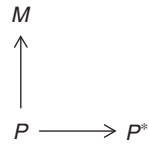

Figure 3.12. The causal graph corresponding to $MC\backslash U_{EX}$

Although the model $MC\backslash U_{EX}$ is simpler than MC, it no longer allows P^* counterfactually to depend on M. Nor does it allow P counterfactually to depend on M. In particular, the counterfactual (i) 'If M had been 0, then P^* would (still) have been 1' is true in $MC\backslash U_{EX}$, as is the counterfactual (ii) 'If M had been 0, then P would (still) have been 1.' In fact, (i) and (ii) are false, however, for they contradict counterfactuals (15) and (10), respectively, which were established in Section 3.3. Thus, some counterfactuals are true in $MC\backslash U_{EX}$ while being in fact false. This makes $MC\backslash U_{EX}$ inappropriate as a causal model for our case.[53]

The second way of removing U from our model continues to treat variable P as endogenous. If we modify the original equation for P minimally to accommodate this change, we get the following equations:

$$\begin{aligned} MC\backslash U_{END} \quad & P \Leftarrow 0 \text{ if } M \text{ is } 0, P \neq 0 \text{ otherwise} \\ & M \Leftarrow \text{Min}\{P,\ 1\} \\ & P^* \Leftarrow \text{Min}\{P,\ 1\} \end{aligned}$$

Figure 3.13 shows the causal graph corresponding to $MC\backslash U_{END}$.

It can easily be checked that all the counterfactuals (except of course the counterfactuals involving U) that were true in our original model MC are also true in the model $MC\backslash U_{END}$. So unlike $MC\backslash U_{EX}$, $MC\backslash U_{END}$ is a genuine alternative to MC. It has one drawback, however. The equations from our original model MC had two solutions. The equations $MC\backslash U_{END}$ have as many solutions as there are possible realizers of the mental event.[54] Like metaphysically dependent variables, multiple solutions should be accepted if they cannot be avoided. But like metaphysically dependent variables, they

[53] Our sufficient condition for causation allows exogenous variables to be causes. Sometimes causation is defined only for endogenous variables, however (see Halpern 2008, Halpern and Hitchcock 2010). Proponents of such a restrictive definition have an additional reason to reject $MC\backslash U_{EX}$ if they want variable P to be at least a candidate cause. Similarly for the third way of simplifying MC by making M an exogenous variable, which will be discussed in note 55.

[54] This is due to the second case in the equation for P. For definiteness one might want to stipulate that P simply be 1 in this case, but then P would become a *de facto* binary variable, and we could no longer capture multiple realizability.

https://doi.org/10.1017/9781108762717.005 Published online by Cambridge University Press

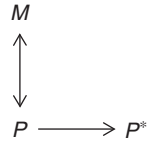

Figure 3.13. The causal graph corresponding to $MC\backslash U_{END}$

should not be multiplied beyond necessity. The gain in simplicity we get by removing U is not worth having a multitude of solutions instead of just two.[55]

3.7 Interventionism

James Woodward (2003) advocates an interventionist theory of causation. Interventionism is a member of the causal modelling family. Its distinctive feature is that it emphasizes the importance of interventions, that is (roughly), isolated manipulations of variables. Whether interventionism can accommodate mental causation on the assumption of non-reductive physicalism has been a matter of controversy.[56] This section argues that interventionism can accommodate mental causation along the lines suggested in previous sections, although this requires modifications of the original theory.

The framework of the debate over interventionism and mental causation differs somewhat from that of our discussion, as it is primarily about causal relations between variables themselves, and not primarily about causal relations between token events that are represented by specific values of variables. For variables X and Y, the basic idea of Woodward's interventionism is that X causes Y if and only if it is possible to manipulate the value of X for at least some individuals that possess X such that this manipulation changes the value of Y for those individuals, given the satisfaction of certain appropriate conditions (see Woodward 2003: 40). More specifically, Woodward characterizes the relations of being a direct cause and being a contributing cause as follows:

[55] A third way of simplifying MC by removing U – though not one that seems particularly obvious – is to make M an exogenous variable, with equations $M \Leftarrow 1$, $P \Leftarrow \text{Min}\{M, 1\}$, and $P^* \Leftarrow \text{Min}\{P, 1\}$. While (10) is true in this model, it still faces problems analogous to those with $MC\backslash U_{EX}$, because (9) is false in it.

[56] See Woodward 2008, Baumgartner 2009, 2010, and Woodward 2015, 2017. Further discussions of the applicability of interventionism to mental causation include Campbell 2007, Shapiro and Sober 2007, Raatikainen 2010, Shapiro 2010, Hoffmann-Kolss 2014, and Weslake 2017.

> (**M**) A necessary and sufficient condition for X to be a (type-level) *direct cause* of Y with respect to a variable set **V** is that there be a possible intervention on X that will change Y [. . .[57]] when one holds fixed at some value all other variables Z_i in **V**. A necessary and sufficient condition for X to be a (type-level) *contributing cause* of Y with respect to variable set **V** is that (i) there be a directed path from X to Y such that each link in this path is a direct causal relationship; that is, a set of variables $Z_1 \ldots Z_n$ such that X is a direct cause of Z_1, which is in turn a direct cause of Z_2, which is a direct cause of . . . Z_n, which is a direct cause of Y, and that (ii) there be some intervention on X that will change Y when all other variables in **V** that are not on this path are fixed at some value. (Woodward 2003: 59; first emphasis added)

Admittedly, more needs to be said (and is said by Woodward) about what an intervention is. For our purposes, however, we can assimilate interventions to (non-backtracking) counterfactuals. Thus, we can read 'there is a possible intervention on X that will change Y when one holds fixed at some value all other variables Z_i' as 'there is a true (non-backtracking) counterfactual with a possible antecedent according to which the value of Y would have changed if the value of X had changed while the values of the Z_i had been so-and-so'.[58]

Michael Baumgartner (2009, 2010) claims that a mental variable M cannot qualify either as a direct cause or as a contributing cause of a physical variable P^* as characterized in (**M**). He reasons as follows. Given non-reductive physicalism, the relation between a mental event and its realizer is not causal. Hence M is not a direct cause of P. Hence P is not on a causal path from M to P^* (where a *causal path* is a sequence of variables related by direct causation). Hence P is to be held fixed in any interventions that test whether M is a contributing cause of P^*. Being distinct from M, P also has to be held fixed in order to determine whether M is a direct cause of P^*. By non-reductive physicalism, it is impossible to intervene on M while holding P fixed, however. Therefore, M is neither a direct cause nor a contributing cause of P^*.

According to this reasoning, M is not a direct cause of P because, owing to non-reductive physicalism, the relation between M and P is not causal. This is too quick, however. Assume that the variables and their possible values are as in our model *MC*. Then there *is* a possible intervention

[57] The omitted phrase concerns the probability distribution of Y, which, given our assumption of determinism, we can ignore for present purposes.

[58] For a detailed discussion of the relation between interventions and counterfactuals, see Woodward 2003: 94–151.

on M that changes the value of P while we hold the remaining variables fixed, namely an intervention that changes the value of M from 1 to 0. Owing to the necessity of the relation between M and P, such an intervention changes the value of P from 1 to 0 even if the values of the remaining variables (that is, U and P^*) are held fixed at their actual values (or at any other values, for that matter). It follows that M *is* a direct cause of P as characterized in (**M**). Assuming that P is a direct cause of P^*, it follows that P is on a causal path from M to P^*. So we need *not* hold P fixed when intervening on M in order to assess whether M is a contributing cause of P^* as characterized in (**M**). Indeed, intervening on M by changing its value from 1 to 0 while holding merely U fixed, P^* changes its value from 1 to 0. Hence M *is* a contributing cause of P^* as characterized in (**M**).

Thus, following the letter of Woodward's characterizations in (**M**), interventionism, far from ruling out mental causation, in fact entails it. But this result comes at the price of predicting too many mental causes. While most theorists (including Woodward himself; see Woodward 2008, 2015) would welcome the result that M is a contributing cause of P^*, few would be happy to call M a direct cause of P. So mental causation still makes trouble for interventionism, although the source of the trouble is not where Baumgartner locates it.

Interventionists could avoid the result that M causes P by restricting causal relations to variables that are not related by metaphysical necessity. More specifically, they could proceed as follows: they could rename 'direct cause' in (**M**) to (say) 'difference-maker', rename 'causal path' to (say) 'difference-making path', define a *direct cause*$_{\text{new}}$ of Y as a difference-maker of Y that is not related to Y by metaphysical necessity, and define a *contributing cause*$_{\text{new}}$ of Y as a contributing cause of Y as characterized in (**M**) that is not related to Y by metaphysical necessity. Thus, M would not be a direct cause$_{\text{new}}$ of P, while being a difference-maker for P; P would be on a difference-making path from M to P^*; and M would be a contributing cause$_{\text{new}}$ as well as a contributing cause of P^*. We saw in Section 3.6 that other theories of causal modelling also require a restriction along the lines suggested here in order not to count the mental event as a cause of its realizer. So *prima facie* interventionism seems no worse off than those theories as far as mental causation is concerned.[59]

[59] It might seem that there is a further problem, namely that interventionism predicts too few physical causes, because intervening on P while holding M fixed at either 0 or 1 does not change the value of P^*. This result is merely an artefact of our simplifying assumption of claim (2) from Section 3.3, however. If one found this assumption intrinsically plausible, one could allow that we need not

The upshot is that interventionism, like the causal modelling theories of token causation discussed earlier in this chapter, can accommodate mental causation. While interventionism faces certain problems that arise from supervenient mental properties, the problem is that it predicts too much mental causation rather than too little. The problem can be solved, however. In particular, interventionists can forestall the conclusion that mental property-instances cause their realizers to be instantiated if they restrict their conditions on causation to variables that are not related by metaphysical necessity.

3.8 Conclusion

Causal modelling can explain mental causation. Starting from counterfactuals about a mental event, its possible realizers, and a later physical event that are true by the lights of non-reductive physicalism (and supernomological dualism) we can construct a causal model that represents all these counterfactuals. This model is unorthodox in that it fails to have acyclic equations. Nonetheless, it establishes that the variable for the later physical event counterfactually depends on the variable for the mental event. On a simple view, this counterfactual dependence suffices to establish that the mental event causes the later physical event. More sophisticated views that invoke the default/deviant distinction or a normality ordering of worlds still rule the mental event to be a cause of the later physical event. In the causal modelling framework, we can give a precise formulation to the idea that one event depends on another when off-path events are held fixed. One can attempt to use this formulation in a new sufficient condition for causation in order to solve the problem of overlapping realizers, but the resulting sufficient condition is problematic, because it makes causation model-relative. It is promising to use dependence while off-path variables are held fixed as a criterion for explanatory relevance, however, and to solve the problem of overlapping realizers by denying not the efficacy, but the explanatory relevance of one of the supervenient property-instances. Various objections against the specifics of the causal model for mental causation can be met. In particular, it is possible to forestall the conclusion that there is a causal relation between the mental event and its realizer. Interventionism licenses that conclusion,

always hold fixed metaphysically dependent variables during interventions. Woodward suggests the latter strategy in response to Baumgartner's criticism; see Woodward 2015: 327–335.

but can solve this difficulty analogously to other causal modelling approaches.

One can take the results of this chapter to show that causal modelling theories can repeat the success of the simpler counterfactual account of explaining mental causation under non-reductive physicalism and super-nomological dualism. Indeed, one can take causal modelling theories to exceed the success of the simpler counterfactual account, since they allow us to spell out a solution to the problem of overlapping realizers. This positive assessment is the attitude I recommend. Alternatively, however, one can take the results to exacerbate the exclusion problem for non-reductive physicalism and super-nomological dualism. One might think that, the more firmly we have established the existence of mental causation given either view about the nature of mind, the more pressing the worry that the physical effects of mental causes are overdetermined becomes. It is time to look at the exclusion problem in more detail.