4
Exclusion Excluded

Brad Weslake

1 Introduction

I take the exclusion problem to be the problem of providing a principled reason to reject at least one of the following inconsistent claims[1]:

- **Non-Reductionism**. Mental properties are distinct from, though metaphysically necessitated by, physical properties.
- **Completeness**. Every event has a complete causal explanation in terms of physical properties.
- **Mental Causation**. There exist causal explanations of events in terms of mental properties.
- **Exclusion**. If an event has a complete causal explanation in terms of one set of properties, then it has no causal explanation in terms of any other properties.

In this paper, I examine the prospects for a principled rejection of **Exclusion**. Following Horgan (1997, 166) and Bennett (2003, 473; 2008, 283), I will refer to this position as *compatibilism*.[2] I will refer to the conjunction of **Non-Reductionism**, **Completeness**, and **Mental Causation** as *non-reductive physicalism*.

Compatibilism is a popular position.[3] However, it has frequently been defended in the absence of an independently justified general framework for thinking about causation and causal explanation. That began to change after the development of a justly influential theory of causation and explanation by James Woodward

[1] While this initial formulation of the problem involves the causal explanation of events in terms of properties, everything I say below could be reformulated depending on your preferred view of the causal relata. Sometimes **Completeness** is weakened, so that it does not presuppose that all events have complete causal explanations. I employ the stronger principle for simplicity, as it will not make any difference to my argument. I assume throughout that 'explanation' is a factive term, and that in a causal explanation all explanans properties are causes of the explanandum. If you prefer not to formulate the problem as involving explanation at all, but rather as involving complete or sufficient causes, be patient: explanation will not appear in the final formulation of the problem I reach in Section 3.5.

[2] Bennett restricts the term to those who say that mental causation is possible without causal overdetermination. I use the term in Horgan's more general sense.

[3] Bennett (2003) cites Goldman (1969), Blackburn (1991), Pereboom and Kornblith (1991), Yablo (1992), Burge (1993), Mellor (1995, 103–104), Horgan (1997), Noordhof (1997), and Yablo (1997), to take just a few of the more prominent adherents.

Brad Weslake, *Exclusion Excluded* In: *Levels of Explanation*. Edited by: Katie Robertson and Alastair Wilson,
Oxford University Press. © Brad Weslake 2024. DOI: 10.1093/oso/9780192862945.003.0005

(2003), which has come to be referred to as *interventionism*. The development of interventionism generated a robust debate concerning whether an interventionist is entitled to reject **Exclusion**, and it is this question I explore in what follows.[4] My central claim is that there is a significant blind spot in the existing discussion, concerning the nature of the relationship between physical and mental properties. Attention to this blind spot reveals that while the best formulation of the interventionist theory of causation entails the falsity of the exclusion principle, it does so at the cost of revealing a weakness in the interventionist theory itself.

The structure of the paper is as follows. In Section 2 I introduce interventionism. In Section 3 I consider how to formulate the exclusion problem in interventionist terms, addressing each component of the problem in turn. In Section 4 I turn to arguments for **Exclusion**. In Section 4.1 I introduce a principle, *subvenience sufficiency*, concerning the relationship between physical and mental properties. The existing discussion has universally accepted the principle, thereby accepting a position I call *internalism*. I consider exclusion arguments from that standpoint in Section 4.2. In Section 4.3 I formulate an exclusion argument under the assumption that subvenience sufficiency is false, a position I call *externalism*. I argue that while interventionism has a response to the argument, it is one that reveals a limitation in the interventionist theory itself. I conclude in Section 5.

2 Interventionism Introduced

Central to the interventionist framework is the notion of a causal model.[5] A causal model is a representational device for encoding counterfactual relationships between variables. Counterfactual relationships are represented by equations which specify the way in which the value of a single variable on the left-hand side would change as a function of changes to the values of the variables on the right-hand side. More formally then, a causal model is an ordered pair $\langle \mathcal{V}, \mathcal{E} \rangle$, where $\mathcal{V}$ is a set of variables and $\mathcal{E}$ a set of equations, and every variable appears on the left-hand side of exactly one equation.

For example, a model $\mathcal{M}_1$ might contain equations representing that variable Y depends on variables X_2 and X_3, that variable X_2 depends on variable X_1, and that variables X_1 and X_3 took values 1 and 0 respectively:

$$Y := X_2 \vee X_3$$

[4] For arguments broadly sympathetic to interventionism on this score, see Shapiro and Sober (2007), Shapiro (2010), Raatikainen (2010), Polger, Shapiro, and Stern (2018), Woodward (2008a, 2015a, 2017), and Stern and Eva (2023). For arguments broadly critical, see Baumgartner (2010, 2009, 2013) (see also Baumgartner 2018), Hoffman-Kolss (2014), and Gebharter (2017b).

[5] Here I present just enough to set up the discussion that follows. For more extended introductions to causal models, see Hitchcock (2009, 2023).

$$X_2 := X_1$$

$$X_1 := 1$$

$$X_3 := 0$$

Here '$\vee$' should be interpreted as a function returning 1 if either side is 1 and 0 otherwise. Equations such as the last two, which simply assign a specific actual value to a variable, are *exogenous*. Equations such as the first two, which assign values as a function of other variables, are *endogenous*. I will assume that the equations are all deterministic, in which case the equations for a model entail the actual values of all variables in the model. In $\mathcal{M}_1$ for example, the equations entail that X_2 and Y both took value 1.

Variables in a causal model must represent entities capable of being changed by interventions, but the framework is otherwise consistent with a range of different metaphysical views concerning the nature of the causal relata. For simplicity, I will sometimes say that variable values represent properties and sometimes say that they represent events. All of this could be translated into whatever view of the causal relata is correct.[6] I will refer to a possible assignment of values to a set of variables as a *state* of that set, and I will talk freely of actual and possible variable values, changes to variable values, states, and changes of state of models. I will also talk about causal relations obtaining between variables and values of variables. This sort of talk should be interpreted throughout as reflecting corresponding actual or possible changes in, and causal relations obtaining between, what is represented by the model. I will assume throughout that a causal model must be veridical, in the sense that every counterfactual relationship specified by the model is true.

The counterfactuals represented by causal models concern *interventions*. An intervention is an exogenous change to the value of a variable in a model, in the sense that the values of the other variables in the model are not themselves causes or effects of the change, unless they are effects of the variable intervened on. Moreover, it is required that interventions be *surgical*, in the sense that the usual causes of the variable in question are suspended, so that the value of the variable depends only on the intervention. I will consider the nature of interventions in more detail in Section 4.2.

In the literature on causation it has been common to distinguish between type-causal relations and token-causal relations. An analogous distinction can be made between causal relations between variables and causal relations between variable

[6] For the complexities that this simplification evades, see Schaffer (2016, Section 1) and Gallow (2022, Section 1).

values. While the terminology is slightly misleading, I will follow Woodward (2003) and refer to causal relations between variables as *type-level* causal relations, and between values of variables as *token-level* causal relations.[7]

In the remainder of this section I introduce the definitions in the interventionist framework that will be important for what follows.[8]

First we need the type-level notion of a *direct cause* (Woodward 2003, 55):

- **DC.** X is a *direct cause* of Y in model $\mathcal{M}$ *iff* there is a possible intervention on X that would change Y when all other variables in $\mathcal{M}$ besides X and Y are held fixed at some combination of values by interventions.[9]

It is a necessary and sufficient condition for X to be a direct cause of Y in $\mathcal{M}$ that X appear on the right hand side of the equation for Y in $\mathcal{M}$. So for example in model $\mathcal{M}_1$, X_1 is a direct cause of X_2, and X_2 and X_3 are direct causes of Y.

Second, we need the type-level notion of a *directed path* (ibid., 42). This can be defined in terms of the properties of graphs associated with causal models. A *directed graph* for $\mathcal{M}$ is an ordered pair $\langle \mathcal{V}, \mathcal{E} \rangle$ where $\mathcal{V}$ is a set of vertices that correspond to the set of variables in $\mathcal{M}$ and $\mathcal{E}$ is a set of *directed edges* connecting these vertices, where there is a directed edge from vertex X to vertex Y *iff* X directly causes Y in $\mathcal{M}$. The definition is then:

- **P.** A sequence of variables $\{V_1 \ldots V_n\}$ is a *directed path* from V_1 to V_n in $\mathcal{M}$ *iff* for all $i(1 \le i < n)$ there is a directed edge from V_i to V_{i+1} in the directed graph for $\mathcal{M}$.

From here on, *path* should be read as equivalent to *directed path*. A path is simply a sequence of direct causes, but the graph-theoretic definition is useful because paths in a model can be easily discerned by constructing a diagram with the same structure as the associated directed graph. When presenting diagrams of this sort, I will follow the usual convention of using circles to represent vertices (variables) and arrows to represent directed edges (direct causes). So for example, by inspecting the diagram for $\mathcal{M}_1$ in Figure 4.1, it is easy to see that X_1 is a direct cause of X_2, that X_2 and X_3 are direct causes of Y, and that there is a path from X_1 to Y, from X_2 to Y, and from X_3 to Y.

[7] For a discussion of the relationship between type-causal relations, token-causal relations, and causal relations between variables, see Hausman (2005).

[8] While I provide references to Woodward throughout, the precise formulations I give are sometimes simplified or expanded, and sometimes make use of definitions introduced in this paper. One important simplification is that I am setting aside the generalisation to the case of probabilistic causation, on which see Fenton-Glynn (2021).

[9] In interpreting the condition in this way I agree with Baumgartner (2009). Woodward (2015a) confirms that this was his intended interpretation.

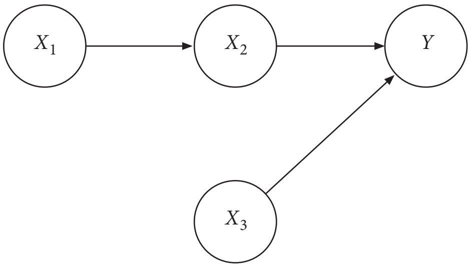

Figure 4.1 Diagram for $\mathcal{M}_1$.

I will provide diagrams of this sort when they are helpful. However, it is important to keep in mind that not all of the information relevant to causation in the interventionist framework can be read off these diagrams. In particular, to know whether the next two definitions are satisfied, you need to know the particular equations that relate the variables.

Third we need the type-level notion of a *contributing cause* (ibid., 59):

- **CC.** X is a *contributing cause* of Y in model $\mathcal{M}$ *iff* for some path P from X to Y in $\mathcal{M}$, there is an intervention on X that will change Y when all variables in $\mathcal{M}$ not on P are held fixed at some combination of values by interventions.

In model $\mathcal{M}_1$ for example, X_1, X_2, and X_3 are all contributing causes of Y. When $X_3 = 0$, an intervention setting X_1 from 0 to 1 would result in Y changing from 0 to 1. Likewise for X_2. And when $X_1 = 0$ and $X_2 = 0$, an intervention setting X_3 from 0 to 1 would result in Y changing from 0 to 1.

Finally, we need the token-level notion of an *actual cause*. The precise way to define actual causation in the interventionist framework remains a matter of lively debate. However, as I show in Weslake (unpublished), many of the proposed definitions can be formulated as instances of the following schema:

- **AC.** $X = x$ is an *actual cause* of $Y = y$ relative to model $\mathcal{M}$ *iff*:
  - **ACT.** The actual value of $X = x$ and the actual value of $Y = y$.
  - **PATH.** There exists a path P from X to Y in $\mathcal{M}$ for which an intervention on X would change the value of Y, when all variables $V_1 \ldots V_n$ in $\mathcal{M}$ that are not on P are held fixed at some combination of values satisfying <*conditions specifying permissible values* $v_1 \ldots v_n$ *for* $V_1 \ldots V_n$>.

The conditions specifying permissible values can be thought of as specifying the set of possible values of the off-path variables relative to which an intervention constitutes a test for actual causation along that path. All definitions of this form in the literature agree that *one* such permissible set is that in which all off-path variables have their actual values. So they all agree that a sufficient condition for $X = x$ to be an actual cause of $Y = y$ is for there to be a path from X to Y such that, holding all off-path variables fixed at their actual values, there is an intervention setting $X = x'$ where $x \neq x'$ that would result in $Y = y'$ where $y \neq y'$. In effect, that is, these theories agree that counterfactual dependence (of this sort) is sufficient for causation. Fortunately, for the purposes of the arguments I make below the differences between the various theories of actual causation on offer do not make any difference. So I will work with the following definition of actual causation, which also takes counterfactual dependence (of this sort) to be necessary for causation:[10]

- **AC**$_A$. $X = x$ is an *actual cause* of $Y = y$ relative to model $\mathcal{M}$ *iff*:
  - **ACT**. The actual value of $X = x$ and the actual value of $Y = y$.
  - **PATH**$_A$. There exists a path P from X to Y in $\mathcal{M}$ for which an intervention on X would change the value of Y, when all variables $V_1 \ldots V_n$ in $\mathcal{M}$ that are not on P are held fixed at their actual values.

In model $\mathcal{M}_1$ for example, $X_1 = 1$ and $X_2 = 1$ are actual causes of $Y = 1$, but $X_3 = 0$ is not. When we hold fixed $X_3 = 0$, an intervention setting X_1 from 1 to 0 would result in Y changing from 1 to 0. Likewise for X_2. But when we hold fixed $X_1 = 1$ and $X_2 = 1$, an intervention setting X_3 from 0 to 1 would not result in Y changing value from 1.

There are several consequences of these definitions that will be important in what follows. First, notice that if $X = x$ is an actual cause of $Y = y$ in $\mathcal{M}$, then X is a contributing cause of Y in $\mathcal{M}$. Second, notice that there may be more than one path that satisfies **PATH**$_A$. When **AC**$_A$ is satisfied in virtue of **PATH**$_A$ being satisfied by path P, I will say that $X = x$ is an actual cause of $Y = y$ *along path P*. Third, notice that each of these definitions is relativised to a causal model. The corresponding de-relativised definitions are as follows:[11]

- X is a *contributing cause* of Y *simpliciter iff* there exists a model in which X is a contributing cause of Y; and
- $X = x$ is an *actual cause* of $Y = y$ *simpliciter iff* there exists a model in which $X = x$ is an *actual cause* of $Y = y$.

I will return to the relationship between the relativised and de-relativised definitions in Section 4.3. However, because it will be important later, note that the

[10] The definition is equivalent (modulo some irrelevant differences) to Woodward's (AC) (2003, 77), and the definition of causation defined in terms of 'Act' in Hitchcock (2001, 286–287). In Weslake (unpublished), I argue against this and all other theories that fit the schema, but the arguments that follow also work for the theory I prefer.

[11] Here I follow Hitchcock (2007, 503) and Woodward (2008b). For an argument that interventionist definitions should not be de-relativised in this way, see Statham (2018).

de-relativised definitions do *not* require that for two variables to be causally related in either of these senses, *every* model containing those variables must represent them as causally related. One is enough.[12]

Because it will simplify the discussion later, I will also introduce the following de-relativised definition here:

- **Causal Chain.** There is a *causal chain* containing X and Y *iff* there exists a model in which X and Y are members of the same path.

This outline is sufficient to exhibit some of the key features of interventionism. First, the theory does not provide an analysis or reduction of causation but rather an explication of causal claims in terms of interventions. The concept of an intervention is itself clearly causal in character, and in the interventionist framework it is explicitly defined in causal terms. What is important for present purposes is that the truth of causal claims can be established independently of any such analysis or reduction—it is whether or not it is true that mental properties sometimes causally explain physical events that is at issue in the exclusion problem, not whether these explanations can be grounded in a reductionist account of causation. Second, this is a kind of counterfactual account of causation—causal claims involve what *would* happen given some particular intervention, not what *actually* or *will* happen. Third, causal claims are model-relative in the sense that they are only well defined with respect to the variables in a particular model. However, as should be clear, this is not a version of causal anti-realism. Causal claims are not made true or false by causal models, they are made true or false by the counterfactuals regarding experimental interventions that are represented by those models.[13] Moreover, because the counterfactuals are explicitly formulated in terms of interventions, it is typically transparent how they can be tested empirically. Nevertheless, as is clear from the definitions above, interventionism does entail that necessarily, if some causal claim is true, then there exists a model in which it is so represented.

3 Exclusion Reformulated

In the interventionist setting, the exclusion problem can be initially formulated as follows:

- **Non-Reductionism$_i$.** The values of mental variables are distinct from, though metaphysically necessitated by, the values of physical variables.

[12] In the terms employed by Stern and Eva (2023), interventionism so understood adopts the *Weak Causation Principle* but not the *Strict Causation Principle*.

[13] This may seem obvious, but the following mistake is routinely made (in this case, by a Nobel Prize winner): 'A model is in the mind. As a consequence, causality is in the mind' (Heckman 2005, 2).

- **Completeness$_i$**. For every event, there exists a causal model containing only physical variables which specifies a complete explanation of that event.
- **Mental Causation$_i$**. There exists a causal model in which a mental variable explains an event.
- **Exclusion$_i$**. If there exists a causal model specifying a complete explanation for an event, there exists no other causal model containing distinct variables specifying an explanation for that event.

In the remainder of this section I clarify and refine these notions in turn, and then more precisely reformulate the exclusion problem in the interventionist setting.

3.1 Non-Reductionism

Non-Reductionism$_i$ requires clarification of the distinction between mental and physical variables. It also requires clarification of the notion of a variable value being distinct from another variable value.

It is a standard presupposition in the debate over exclusion that there is a distinction between physical properties and mental properties in the sense required to generate the problem. I take no stand on how this distinction should be drawn. But to keep the relationship between models and what they represent clear, I assume that corresponding to it is a distinction between sets of variables that represent physical properties and sets of variables that represent mental properties. I will refer to these sets of variables as involving *vocabularies*, where a vocabulary is a set of variables with variable values that all represent a single property type. So a *physical vocabulary P* contains only variables with values representing physical properties, and a *mental vocabulary M* contains only variables representing mental properties.

In order for **Non-Reductionism$_i$** to occupy the proper role in the exclusion problem it needs to express a claim about the world, not about our ways of representing the world. So I will assume that variable values are distinct if and only if they represent distinct properties:

- **Value Distinctness**. Variable values are distinct *iff* they represent distinct properties.

In addition, note that there is a necessary condition on two variables appearing in the same model that follows from the definition of direct causation provided in Section 2. Recall that whether X is a direct cause of Y in $\mathcal{M}$ depends, according to **DC**, on whether there exists an intervention on X that will change Y when all other variables in $\mathcal{M}$ are held fixed at some combination of values by interventions. This implies an independence condition on variables coexisting in a model: if X is a direct cause of Y in $\mathcal{M}$, then there must be possible values x and x' for X such that an intervention on X from x to x' is possible when all other variables except Y in $\mathcal{M}$ are set to some combination of possible values by independent interventions. There is a natural generalisation of this independence condition standardly assumed to hold in causal models, which can be motivated by the idea that for any set of variables appearing together in a model it must be possible to non-trivially *test*, for every pair, whether **DC** holds. According to Woodward (2003, Section 3.5; 2015a), the relevant sense of possibility here is *at a minimum* metaphysical possibility. The corresponding independence condition on variables coexisting in a model $\mathcal{M}$ is this:

- **Independent Manipulability**. It is metaphysically possible that every proper subset of the variables in $\mathcal{M}$ be set to every combination of their possible values by independent interventions.[14]

Independent Manipulability reflects the natural idea that it is only variables not related by metaphysical necessity that are candidates for being related causally. It is well known that counterfactual theories of causation are inadequate if we allow dependencies between events that are related by metaphysical necessity (Kim 1973), and **Independent Manipulability** can be seen as the constraint that implements this restriction in the interventionist framework. When I refer to the interventionist theory of causation in what follows, I will take it to include all of the definitions provided in Section 2, as well as **Value Distinctness** and **Independent Manipulability.**[15]

My final formulation of **Non-Reductionism** is therefore:

- **Non-Reductionism**$_j$. Mental variables are distinct from physical variables in the sense that they are drawn from distinct vocabularies M and P, and the values of the M-variables are metaphysically necessitated by the values of the P-variables.

3.2 Completeness

Completeness$_i$ requires clarification of the notion of a complete explanation for an event. The exclusion problem is often framed in terms of causal sufficiency rather

[14] Baumgartner (2009, 167) calls a related condition *Fixability*, and Woodward (2015a, 316; 2017, 255) a related condition *Independent Fixability*. See Hoffmann-Kolss (this volume) for an additional line of argument for imposing the condition.

[15] For more detailed discussion of the reasons for imposing constraints of these sorts, and proposals for further necessary conditions on variables, see Hitchcock (2001, 2004, 2012), Halpern and Hitchcock (2010), Halpern (2016), Woodward (2016), Blanchard and Schaffer (2017), McDonald (forthcoming, 2023), and Hoffmann-Kolss (this volume).

than completeness, so any defensible notion of completeness must bear some close relationship to a notion of causal sufficiency. To begin, note that this should not be interpreted as being equivalent to the claim that every event can be completely explained in terms of some fundamental physical theory. This is so for at least two reasons. First, it is an open question whether reasonable candidates for fundamental physical theories should be interpreted causally.[16] Second, even if reasonable candidates for fundamental physical theories should be interpreted causally, it is not the case that the structure of fundamental physical theories is identical to the structure of causal models.[17] So for example, sufficiency is often defined along the following lines:

- **Physical Sufficiency**. An event *A* is physically sufficient for an event *B* *iff* the occurrence of *A* and the laws of physics together guarantee that *B* will occur (or fix a probability for *B* such that there are no further events conditioning on which would change the probability of *B*).[18]

However, **Physical Sufficiency** makes no reference to causal models, and it is not clear how it should be translated into those terms. Since I am proceeding under the interventionist assumption that causation is to be defined with respect to models, this notion is therefore inadequate for formulating the exclusion problem.[19] In making this point I do not mean to weaken the support that causal completeness assumptions rightly draw from the promise of complete explanations of events in terms of fundamental physical theories. My point is simply that there is an inference involved from the success of fundamental physics to the existence of a complete causal model in the sense required to formulate the exclusion problem in the interventionist setting.

Having clarified what **Completeness**$_i$ does not say, let us examine what it does say. There are a number of different notions of causal sufficiency that can be discriminated within the interventionist framework, only one of which, I will argue, is suitable for formulating the exclusion problem.

[16] See Russell (1913), Field (2003), and the essays in Price and Corry (2007).

[17] One reason is that causal models do not allow the representation of continuous processes (Strevens 2007, 242–244). Strevens puts the point by saying that interventionist causal models 'represent less of causal reality than is actually out there' (243), but an interventionist may consistently claim both that every interventionist model omits some causal truth, and that all causal truths are represented by some interventionist model or other (Woodward 2008b, 210–211).

[18] See, for example, Papineau (2001, 8; 2002, 17). Note that the relevant notion of event here must be liberal enough to allow events involving all physical properties instantiated across the entire cross-section of a light-cone in spacetime, if any events are going to turn out to be physically sufficient for any others (Field 2003; Ismael 2009, 2011).

[19] See Yates (this volume) for critical discussion of a set of principles closely related to **Physical Sufficiency**.

Consider first the following definition:

- **Sufficiency in the Circumstances in a Model.** A cause is sufficient in the circumstances for an effect in a model *iff* it is an actual cause of the effect in that model.

Since **Sufficiency in the Circumstances in a Model** collapses the notion of a sufficient cause and the notion of an actual cause, it is clearly too weak to play a role in formulating an appropriate causal closure principle. As a step in the right direction, consider next:

- **Weak Sufficiency in a Model.** Call an actual cause of an effect in a model a *weakly sufficient actual cause iff* it is an actual cause of the effect along path P, and there is no possible combination of interventions on variables not on P that would change the effect, if the actual cause were held fixed to its actual value by an intervention. A cause is weakly sufficient for an effect in a model *iff* it is a *weakly sufficient actual cause* of that effect in that model.[20]

Notice that this definition, like the preceding one, is a model-internal one, in the sense that sufficiency is defined with respect to a single causal model, and makes no reference to facts apart from those represented by that model.[21] This makes trouble. For suppose that we have a model $\mathcal{M}_P$ framed in variables drawn from physical vocabulary P which specifies a cause that is weakly sufficient for some effect. That is consistent with supposing that there exists some model $\mathcal{M}_{PM}$ constructed by adding variables from mental vocabulary M to $\mathcal{M}_P$, in which the M -variables specify an actual cause for the effect and the P-variables do not specify a cause that is weakly sufficient for the effect. What this possibility reveals is that the model-internal definitions of sufficiency do not adequately capture the idea, central to any closure principle, that when one class of properties is causally closed with respect to another class the latter do not make any *additional* causal difference. I conclude that an adequate closure principle must be at least as strong as:

- **Weak Sufficiency in a Weakly Closed Model.** Call a model $\mathcal{M}_F$ framed in variables drawn from vocabulary F *weakly closed* with respect to variables drawn from vocabulary G with respect to an effect *iff* $\mathcal{M}_F$ contains a weakly sufficient

[20] Related but distinct notions are *sustenance* in Pearl (2000, Section 10.2), *switch* in Woodward (2003, 96–97), *strongly causes* in Halpern and Pearl (2005, 855), and *sufficient condition* in McDermott (1995, 533; 2002, 96–97). To keep the formulations as simple as possible, I give definitions on which only values of single variables can be sufficient for others. The generalisation to multiple variables is obvious and does not make any difference to the arguments that follow.

[21] To re-emphasise a point I made in Section 2, remember that a notion being defined in a model-internal way does not imply that the corresponding fact is in any way model-dependent.

actual cause of the effect, and there exists no model $\mathcal{M}_{FG}$, constructed by adding variables from G to $\mathcal{M}_F$, in which any weakly sufficient causes in $\mathcal{M}_F$ are not also weakly sufficient causes in $\mathcal{M}_{FG}$. A cause is weakly sufficient for an effect in a weakly closed model $\mathcal{M}_F$ with respect to G *iff* it is weakly sufficient for the effect in $\mathcal{M}_F$.

Notice however that **Weak Sufficiency in a Weakly Closed Model** is compatible with the actual values of $\mathcal{M}_P$ specifying a weakly sufficient actual cause of an effect, and yet it being the case that no *alternative* values of the P-variables would have specified weakly sufficient causes of *alternative* values of the effect. That is, it is compatible with the actually instantiated physical properties sufficing for some event, while any alternatively instantiated physical properties would not have sufficed for any alternative event. So **Weak Sufficiency in a Weakly Closed Model** does not yet capture the sort of closure the successes of our scientific theorising typically license us to endorse, where for some class of properties proprietary to a theory, *whichever of those properties were instantiated* would have sufficed for all outcomes of a certain type. I conclude that the closure principle appropriate to formulating the exclusion problem in the interventionist setting is:

- **Strong Sufficiency in a Strongly Closed Model**. Call a cause *strongly sufficient* for an effect in a model *iff* it is weakly sufficient for the effect, and all alternative values of the cause would also be weakly sufficient for the value of the effect in any possible state of the model. Call a model $\mathcal{M}_F$ framed in variables drawn from vocabulary F *strongly closed* with respect to variables drawn from vocabulary G with respect to an effect *iff* $\mathcal{M}_F$ contains a strongly sufficient actual cause of the effect, and there exists no model $\mathcal{M}_{FG}$, constructed by adding variables from G to $\mathcal{M}_F$, in which any strongly sufficient causes in $\mathcal{M}_F$ are not also strongly sufficient causes in $\mathcal{M}_{FG}$. A cause is strongly sufficient for an effect in a strongly closed model $\mathcal{M}_F$ with respect to G *iff* it is strongly sufficient for the effect in $\mathcal{M}_F$.

It is important to note an immediate consequence of this definition. If a cause is strongly sufficient for an effect in a strongly closed model $\mathcal{M}_F$ with respect to G, then in any model $\mathcal{M}_{FG}$, constructed by adding variables from G to $\mathcal{M}_F$, there are no paths from any variables in G to the effect.

Strong Sufficiency in a Strongly Closed Model is a model-external definition but still a relative one, in the sense that a cause could be strongly sufficient in a strongly closed model with respect to one set of variables, but not with respect to a different set of variables.[22] While it would not make a difference to the argument below if we strengthened our understanding of completeness yet again, so that it

[22] This is also true of the closely related probabilistic conception of completeness in Sober (1999a, 139).

involved the idea of a model strongly closed with respect to *all* other variables, I prefer the present formulation. This is because understanding the problem in this way captures the great variability in the way completeness assumptions are formulated. Sometimes the worrying complete or sufficient explanation is supposed to be provided by physics, sometimes by biology, sometimes by neuroscience, sometimes by (at least the 'syntactical' explanations appearing within) cognitive science.[23] In my view the exclusion problem can be posed in terms of these different sciences precisely because it is reasonable to believe that there exist strongly sufficient causes, in models framed in variables drawn from the vocabularies of each of these sciences, which are strongly closed with respect to the M-variables.[24] If I had all of the physical information relevant to you, my knowledge of what you will and would do would not be increased by knowing any further mental information about you—and likewise if I had all of the biological information, or all of the neuroscientific information, or all of the ('syntactic') cognitive scientific information.[25] Moreover, once we understand completeness in the way I have suggested, it can be seen that the exclusion problem generalises—the physical causal model is strongly closed with respect to the variables of the biological causal model, the biological causal model is strongly closed with respect to the variables of the neuroscientific causal model, and so on up the hierarchy of the sciences and never vice versa.[26] And so if the exclusion problem arises for mental variables it also arises for any variables not appearing in some maximally strongly closed causal model.[27]

Because my central interest is in **Exclusion**, in what follows I will not defend these claims, and will simply proceed under the assumption that a closure principle concerning physical and mental variables formulated in terms of **Strong Sufficiency in a Strongly Closed Model** is true.[28] My final formulation of **Completeness** is therefore:

- **Completeness**$_j$. For every event, there exists a causal model with variables drawn from P, which is strongly closed with respect to M, in which there is a strongly sufficient actual cause P_1 for that event.

[23] The emphasis on physical causes is familiar from Kim (1998, 2005). As is made clear in Kim (1989), Kim was generalising from an argument initially formulated by Malcolm (1968) in terms of a 'neurophysiological theory'. The emphasis on syntactic causes is familiar from Field (1978) and Stich (1983).

[24] For a historical survey of how completeness in physics and biology became compelling, see Papineau (2001). For a comparison with the assumptions that generated earlier problems with mental causation, see Patterson (2005).

[25] See Loewer (2008, 2009). Note that this is not to say that my *explanations* would not be improved by the possession of this information. Indeed, I think they would be (Weslake 2010).

[26] It is a *hierarchy* in part *because* this relation is asymmetric in this way.

[27] Here I side with Bontly (2002) against Kim (1997, 1998).

[28] There are two options available to someone who accepts **Non-Reductionism**$_j$ but wishes to deny **Completeness**$_j$. One is to deny **Physical Sufficiency**, and with it **Completeness**$_j$. In my view the

3.3 Mental Causation

My initial formulation of **Mental Causation** is straightforward:

- **Mental Causation**$_j$. There exists a causal model with variables drawn from M in which a mental variable M_1 is an actual cause of an event.

Three comments on this formulation, before I introduce a revision in the following section.

First, interventionism is attractive not only as a theory of causation generally, but as a theory of mental causation specifically. In particular, it has been defended as providing a good framework for understanding causal explanation in psychology (Campbell 2007; see also Rescorla 2018; Kaiserman 2020), in psychiatry (Campbell 2008; Kendler and Campbell 2009), and in folk psychology (Menzies 2010). If interventionism is true, there is no special problem in understanding how a mental variable can be a cause.

Second, **Mental Causation**$_j$ might be granted, and yet it might be argued that only a model containing physical variables *really* represents causes, and that models containing mental variables merely specify *explanations*, or some other weak cousin of causation. This might be because only the physical model promises to be maximally predictively accurate and therefore maximally strongly closed (Davidson 1963, 1967, 1970, 1995), or because the physical model is a truth-maker for the mental model (Crane 2008; Robb, Heil, and Gibb 2023, Section 5.3), or for more recherché metaphysical reasons (Jackson and Pettit 1988, 1990a, 1990b). In my view the arguments for these claims are unsound, but for present purposes I simply note that if they succeed they are arguments against interventionism in general and therefore should be addressed as independent claims about the nature of causation and causal explanation. Proceeding under the presumption of the truth of interventionism, I here set them to one side.[29]

Third, while this will also make no difference to the argument below, note that **Mental Causation**$_j$ does not require that in order for mental variables to causally

most interesting arguments of this sort are those made by Cartwright (1983, 1994) as developed in the case of chemistry by Hendry (2006, 2010b, 2010a, 2017). See Sklar (2003) for the general line of response that I think blocks these arguments. The other is to accept **Physical Sufficiency** but deny that it entails **Completeness**$_j$. One strategy here turns on the idea that causes must be 'proportional' to their effects (Yablo 1992; List and Menzies 2009; Menzies and List 2010; Raatikainen 2010; and for critical discussion Weslake 2017). Another strategy, the 'dual explanandum' or 'intralevelist' solution to the exclusion problem, turns on the idea that the effects of mental causes are individuated mentally rather than physically (the position dates at least to Putnam 1975; see also Marras 1998; Thomasson 1998; Gibbons 2006; Schlosser 2009; and for critical discussion Sober 1999b; Buckareff 2011, 2012).

[29] See Burge (2007) and Woodward (2008a, 244–249) for arguments against some of these lines of objection.

explain, they must be either weakly or strongly sufficient for their effects. It is unclear to me why the exclusion problem is often framed so that mental causes must be sufficient for their effects in a stronger sense than sufficiency in the circumstances. Bennett (2003, Section 5) thinks that anything less than sufficiency would endanger the 'full-fledged causal efficacy of the mental' (481), granting it merely 'a derivative efficacy' (482). I cannot see the motivation for claims of this form if sufficiency is supposed to be stronger than circumstantial sufficiency—especially given the metaphors that are often used to characterise the exclusion problem.[30] If an event has a complete physical cause, mental causes are often said to have 'no work left to do' (Kim 1998, 35, 37, 54, 110, 126 n. 6), 'no gaps left to fill' (Menzies 2003), no opportunity to 'inject themselves' into the causal order (Kim 1998, 41; 2005, 16); if there is no lowest level of causation, we are supposed to worry that causal powers will 'drain away' (Block 2003; Kim 2003). But if there *was* work left to do, a gap to be filled, an injection to be provided, or a drain to be plugged, presumably the context would be almost sufficient, and the additional impetus plus context would be wholly sufficient. The work, filler, injection, or plug would not itself be wholly sufficient, but rather would be sufficient in the circumstances. Now perhaps these are all just poor metaphors for what is supposed to be at issue here; but metaphors aside, the claim in question would be that any actual causes that are not at least weakly sufficient must have merely derivative efficacy. Given that sufficiency in the circumstances is the sort of efficacy most causes have in most scientific theories, I say that derivative causes in this idiosyncratic sense would be causes enough for mental causation.[31]

3.4 Exclusion

It may seem that the formulation of **Exclusion** is now straightforward: it should simply be the strongest principle that is inconsistent with **Non-Reductionism**$_j$, **Completeness**$_j$, and **Mental Causation**$_j$. However, for the principle to have any *prima facie* plausibility, it needs to be weaker. To translate a point first made by Goldman (1969, 470–473; 1970, 159–161) into this context, **Completeness**$_j$ is perfectly compatible with **Mental Causation**$_j$ in a case where there is a path from the mental variable that is an actual cause of the event to the physical variable that is strongly sufficient for the event, or in a case where the mental variable is on a path from the physical variable to the event. A principle that says that if an event has a sufficient cause it has no other causes is clearly far too strong to be plausible, for it is inconsistent with the existence of causal chains.

[30] I do not suggest Bennett endorses the position I here criticise.

[31] For a more detailed argument for this claim, see Woodward (2008a, 245–249).

My final formulation of **Exclusion** is therefore:

- **Exclusion**$_j$. If there exists a causal model with variables drawn from a vocabulary F, which is strongly closed with respect to variables drawn from vocabulary G, and in which a variable F_1 is a strongly sufficient actual cause for an event, then there exists no causal model in which a variable G_1 drawn from vocabulary G is an actual cause of that event, unless there is a causal chain containing F_1 and G_1.

An attractive feature of this formulation is that it reveals a way in which someone who rejects **Non-Reductionism**$_j$ can evade the exclusion problem. Here I have in mind Lowe (2000, 2003), who has argued that all causal closure principles with strong empirical support are logically consistent with non-physicalist theories on which mental causes occupy a place in causal chains *between* sufficient physical causes and their effects. While I think we have overwhelmingly strong reasons to reject theories of this sort, my formulations of **Completeness**$_j$, **Exclusion**$_j$, and **Mental Causation**$_j$ support Lowe's claim.

The non-reductive physicalist, on the other hand, is in no position to make a similar move. They would thereby be committed to a position on which mental causes occupy a place in causal chains between sufficient physical causes and their effects, but are metaphysically necessitated by *different* physical variables, which *are not themselves sufficient for those effects*. It is rare to find a position in logical space no philosopher is willing to occupy, but this must be one of them.[32] My final formulation of **Mental Causation**$_j$ is therefore the following, which closes this loophole and renders the propositions that form the exclusion problem logically inconsistent:

- **Mental Causation**$_j$. There exists a causal model with variables drawn from M in which a mental variable M_1 is an actual cause of an event, and there is no causal chain containing P_1 and M_1.

3.5 The Interventionist Exclusion Problem

Putting this all together, I conclude that the exclusion problem in the interventionist framework should be formulated as follows:

- **Non-Reductionism**$_j$. Mental variables are distinct from physical variables in the sense that they are drawn from distinct vocabularies M and P, and the values of the M-variables are metaphysically necessitated by the values of the P-variables.

[32] See Kim (1998, 37, 40, 44). As Kim notes, the non-reductive physicalist will invariably be committed to versions of physicalism and closure on which this option is ruled out.

- **Completeness**$_j$. For every event, there exists a causal model with variables drawn from P, which is strongly closed with respect to M, in which there is a strongly sufficient actual cause P_1 for that event.
- **Mental Causation**$_j$. There exists a causal model with variables drawn from M in which a mental variable M_1 is an actual cause of an event, and there is no causal chain containing P_1 and M_1.
- **Exclusion**$_j$. If there exists a causal model with variables drawn from a vocabulary F, which is strongly closed with respect to variables drawn from vocabulary G, and in which a variable F_1 is a strongly sufficient actual cause for an event, then there exists no causal model in which a variable G_1 drawn from vocabulary G is an actual cause of that event, unless there is a causal chain containing F_1 and G_1.

One of these claims must go. We are finally in a position to consider arguments for **Exclusion**$_j$.

4 Compatibilism Examined

In this section I evaluate two arguments for **Exclusion**$_j$. The first is familiar from discussion of the exclusion problem in the interventionist setting, but the second is not. This is because the discussion has almost universally assumed a particular conception of the relationship between physical and mental variables, according to which the mental cause M_1 that figures in **Mental Causation**$_j$ is metaphysically necessitated by the strongly sufficient actual cause P_1 that figures in **Completeness**$_j$. I will call this assumption *subvenience sufficiency*, the non-reductive physicalist position that accepts it *internalism*, and the non-reductive physicalist position that denies it *externalism*. As I will show, the distinction is important. I begin with a discussion of subvenience sufficiency itself, and then consider internalism and externalism in turn. I side with those who take the interventionist to have a good response to the argument for **Exclusion**$_j$ under the assumption of internalism. But I go on to argue that the response the interventionist has to the argument for **Exclusion**$_j$ under the assumption of externalism serves to expose a weakness in interventionism itself.

4.1 Subvenience Sufficiency

As formulated, the exclusion problem invites us to consider two causal models. Each model contains a variable E_1 that is a candidate effect for a mental cause. The first model, $\mathcal{M}_P$, the existence of which is entailed by **Completeness**$_j$, contains (in addition to E_1) only variables drawn from P, is strongly closed with respect to M, and contains a strongly sufficient actual cause P_1 for E_1. The second model, $\mathcal{M}_M$, the existence of which is entailed by **Mental Causation**$_j$, contains (in addition to

E_1) variables drawn from M, and contains an actual cause M_1 of E_1. As is also entailed by **Mental Causation**$_j$, there is no causal chain containing P_1 and M_1. I will make the simplifying assumptions that $\mathcal{M}_M$ contains (in addition to E_1) only variables drawn from M, and that both $\mathcal{M}_P$ and $\mathcal{M}_M$ only contain variables on paths terminating in variable E_1.

With respect to models of this sort, subvenience sufficiency can be defined as follows:

- **Subvenience Sufficiency.** Given two models $\mathcal{M}_F$ and $\mathcal{M}_G$, where each model only contains variables on paths terminating in variable E_1, $\mathcal{M}_F$ is *subvenience sufficient* with respect to $\mathcal{M}_G$ and E_1 *iff* the values of all other variables in $\mathcal{M}_G$ are metaphysically necessitated by the values of strongly sufficient causes of E_1 in $\mathcal{M}_F$.

It is important to see that $\mathcal{M}_P$ being subvenience sufficient with respect to $\mathcal{M}_M$ is a substantive assumption that is not itself entailed by **Non-Reductionism**$_j$ either alone or in conjunction with **Completeness**$_j$ and **Mental Causation**$_j$. In particular, while **Non-Reductionism**$_j$ merely requires that the values of the M-variables are metaphysically necessitated by the values of the P-variables, $\mathcal{M}_P$ being subvenience sufficient with respect to $\mathcal{M}_M$ imposes the much stronger constraint that the M-variables are metaphysically necessitated by the very P-variables that are strongly sufficient for their effects. The assumption is vividly illustrated by what Loewer (2015, 60) calls 'Kim's Favourite Diagram' (2003, 159), a way to represent the exclusion argument that is ubiquitous in Kim's work, in which one and the same physical event is represented as both the cause of a given effect, and as the subvenience basis for the mental event which Kim takes it to exclude (see Figure 4.2).

I will refer to non-reductive physicalism in conjunction with **Subvenience Sufficiency** as *internalism* and non-reductive physicalism without **Subvenience Sufficiency** as *externalism*. The fact that internalism has been so frequently assumed in the discussion of the exclusion problem would be unremarkable if it were not the case that most non-reductive physicalists are committed to rejecting it, and if it did not make a difference to the arguments available to the non-reductive

$$\begin{tikzcd}
\mathrm{M} & \mathrm{M}^* \\
\mathrm{P} \arrow[u, "\text{Supervenes}"] \arrow[r, "\text{causes}"] & \mathrm{P}^* \arrow[u, "\text{Supervenes}"']
\end{tikzcd}$$

Figure 4.2 Kim's Favourite Diagram.

physicalist for rejecting **Exclusion**$_J$. As Bennett (2003) notes, internalism is inconsistent with content externalism, with functionalism in general, and with conceptual role semantics in particular.[33] These are the most prominent of the theories of mental properties that motivate non-reductionism in the first place, so it is hardly open to the non-reductive physicalist to ignore their consequences. I begin, however, with internalism.

4.2 Internalism

Baumgartner (2009, 2010) has argued that if interventionism is true, then whenever variables stand in relationships of metaphysical necessitation, the necessitated variable cannot have any of the same effects as the necessitating variable. If that is right, then internalism is incoherent. For the internalist is committed, by virtue of the claim that $\mathcal{M}_P$ is subvenience sufficient with respect to $\mathcal{M}_M$, to the existence of variables that stand in exactly this relationship. In this section I argue that the correct formulation of interventionism blocks this argument.[34]

As was clear from the definitions introduced in Section 2, all of the fundamental interventionist causal concepts are defined in terms of interventions. Baumgartner's argument depends on the way in which interventions are defined by Woodward (2003, 98). Woodward first introduces the type-level notion of an *intervention variable*:

- **IV.** *I* is an *intervention variable* for *X* with respect to *Y iff*:
  - – **I**$_1$. *I* is a contributing cause of *X*.
  - – **I**$_2$. There is a model in which *I* has at least one value that is weakly sufficient for the value of *X*.
  - – **I**$_3$. Every causal chain from *I* to *Y* contains *X*.
  - – **I**$_4$. *I* is statistically independent of every contributing cause of *Y* on causal chains that do not contain *X*.

This is then used to define the token-level notion of an *intervention*:

- **IN.** $I = i$ is an *intervention* on *X* with respect to *Y iff* is an intervention variable for *X* with respect to and there is a model in which $I = i$ is a weakly sufficient cause of the value of *X*.

[33] In addition, Worley (1993, Section 5) argues that internalism is inconsistent with anomalous monism and the folk-psychological platitude that a single mental state may be a cause of different effects on different occasions.

[34] I am indebted here to correspondence with Michael Baumgartner.

Note that I have presented definitions that are weaker than Woodward's in one respect, since $\mathbf{I}_2$ is weaker than Woodward's condition. The distinction amounts to whether interventions must be *hard*, so that they override all other causal connections to the variable intervened on, or whether they may be *soft*, and merely make an additional causal impact to the variable intervened on. I opt for the weaker definitions because Baumgartner's argument works either way, and because Woodward himself accepts both formulations (2015b, 3584; 2015a, 321 fn. 15; 2017, 254 fn. 3).[35]

Baumgartner's argument is simple, with each premise following from the definitions of the relevant notions, or from claims the internalist is committed to accepting. For M_1 to be an actual cause of E_1 in $\mathcal{M}_M$, it must be a contributing cause of E_1 in $\mathcal{M}_M$ ($\mathbf{AC}_\mathrm{A}$). For it to be a contributing cause of E_1 in $\mathcal{M}_M$, there must be an intervention on M_1 with respect to E_1 (**CC**). For there to be an intervention on M_1 with respect to E_1, there must be an intervention variable I for M_1 with respect to E_1 (**IN**). For there to be an intervention variable I for M_1 with respect to E_1, every causal chain from I to E_1 must contain M_1 (**IV**, $\mathbf{I}_3$), and I must be statistically independent of every contributing cause of E_1on causal chains that do not contain M_1 (**IV**, $\mathbf{I}_4$). But internalism is committed to the claim that $\mathcal{M}_P$ is subvenience sufficient with respect to $\mathcal{M}_M$. This entails that there is no way to make a change to M_1 without also changing P_1, which in turn entails both that there is a causal chain from I to P_1 to E_1 that does not contain M_1, and that P_1 is not statistically independent of M_1. So there is no such intervention variable I, M_1 is not an actual or contributing cause of E_1, and there is no such model $\mathcal{M}_M$ as required by the internalist.[36]

In response to this argument, Woodward (2015a, 323) helpfully distinguishes three questions. First, are the definitions that lead to this result adequate interpretations of Woodward (2003)? Second, must any interventionist theory that deserves the name adopt definitions that lead to this result? Third, in order for variables to be causes, must they make a difference to their effects beyond the differences made by variables that metaphysically necessitate them? I set Woodward's first question aside.[37] On Woodward's second question, some authors have considered ways in which the basic interventionist framework can be expanded, so that variables related both causally and by metaphysical necessitation can appear in the same model. The debate then becomes whether the principles that should

[35] For discussion of hard and soft interventions, see Korb et al. (2004); Markowetz, Grossmann, and Spang (2005); Eberhardt and Scheines (2007); Eberhardt (2014), and in the psychological context Campbell (2007); Kaiserman (2020).

[36] As Baumgartner (2009, 171) notes, the argument does not depend on a premise concerning causal closure: it can be used to show that, on these definitions, no variables related by metaphysical necessity can share any effects. Gebharter (2017a) argues that this is also the case with respect to the argument in Gebharter (2017b), and Stern and Eva (2023) agree.

[37] Woodward (2015a, 324–325; 2017, 257) has argued that the answer is 'no'. As he says, it is the least interesting of the three.

govern models of this sort should permit or prohibit causation by necessitated variables (Baumgartner 2010; Woodward 2015a; Gebharter 2017b; Stern and Eva 2023). If we wish to restrict our focus to causal relationships, however, there exists a more conservative amendment of the interventionist framework that is sufficient to block Baumgartner's argument.[38]

The amendment is as follows:

- **IV***. *I* is an *intervention variable* for *X* with respect to *Y iff*:
  - $\mathbf{I}_1$. *I* is a contributing cause of *X*.
  - $\mathbf{I}_2$. There is a model in which *I* has at least one value that is weakly sufficient for the value of *X*.
  - $\mathbf{I}_3^*$. Every path from *I* to *Y* goes through *X* in every model containing *I*, *X* and *Y*.
  - $\mathbf{I}_4^*$. *I* is statistically independent of every contributing cause of *Y* on paths that do not contain *X* in every model containing *I*, *X* and *Y* .

The difference between **IV** and **IV*** concerns the third and fourth conditions. Condition $\mathbf{I}_3$ requires that there are no paths, in any model, from *I* to *Y* without *X* . Condition $\mathbf{I}_3^*$ relaxes this requirement, requiring that there are no paths from *I* to *Y* without *X in any model that contains those variables*. Likewise, condition $\mathbf{I}_4$ requires that *I* is statistically dependent of contributing causes of *Y*, in all models, that are on paths without *X* . $\mathbf{I}_4^*$ relaxes this requirement, requiring that *I* is statistically dependent of contributing causes of *Y*, on paths without *X, in any model that contains those variables*.

The difference between these definitions shows up as a consequence of **Independent Manipulability** (Section 3.1), according to which it is a condition on variables coexisting in a model that it be metaphysically possible for every proper subset to be set to every combination of their possible values by independent interventions. This entails:

- **Non-Necessitation**. A causal model cannot contain a variable with a possible value that is metaphysically necessitated by a possible combination of values of any proper subset of the other variables in the model.

An immediate consequence of this, if internalism is true, is that there are no causal models that contain both P_1 and M_1. This in turn blocks Baumgartner's argument. According to **IV***, the presence of a causal chain from *I* to P_1 to E_1 that does not contain M_1, and the fact that P_1 is not statistically independent of M_1, do not threaten the satisfaction of $\mathbf{I}_3^*$ and $\mathbf{I}_4^*$. More generally, the fact that any change to M_1

[38] The basic strategy I develop in the remainder of this section is also proposed by Eronen and Brooks (2014), who cite an earlier version of this paper. I do not endorse their arguments for it.

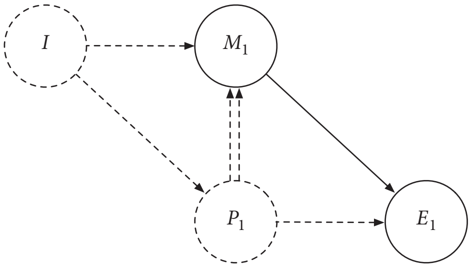

Figure 4.3 Diagram for $\mathcal{M}_M$, if internalism is true. Solid lines represent variables and direct causes that are in the model. Dashed lines represent variables and direct causes that are not in the model. Double arrows represent metaphysical necessitation. According to **IV**, *I* cannot be an intervention variable for M_1 with respect to E_1. According to **IV***, it can.

entails some corresponding change to P_1 is no obstacle to there being well-defined interventions on M_1 (see Figure 4.3).

The answer to Woodward's second question, therefore, is 'no'. There is a coherent formulation of an interventionist theory of causation that does not generate the consequence that a necessitated variable cannot have any of the same effects as the necessitating variable. Moreover, it is a formulation that is perfectly suited to the non-reductive physicalist, in the following sense. As Bennett (2008) argues, the non-reductive physicalist would ideally like a solution to the exclusion problem that can play two roles. On the one hand, it should show that causal considerations do not force her into reductive physicalism. On other hand, it should show that causal considerations are still a problem for the dualist.[39] Interventionism formulated in terms of **IV*** has both consequences, at least for the internalist. On the one hand, as I have just argued, the internalist can argue that their position is coherent, and compatible with mental causation. On the other hand, the internalist can point out that the dualist, in virtue of rejecting the metaphysical necessitation of the values of mental variables by the values of physical variables, cannot avail themselves of the same sorts of interventions on mental variables. Instead, they must commit to the existence of interventions on mental variables that do not entail any changes to physical variables, and that are statistically independent of physical variables. But in that case, accepting **Completeness**$_j$ would force the dualist to admit that while there may be such interventions, they could not result in any downstream effects. As a result, the dualist who accepts **Completeness**$_j$ is

[39] In the current framework, *reductive physicalism* can be defined as the position on which mental variables are not distinct from physical variables, and *dualism* can be defined as the position on which mental variables are distinct from physical variables, and the values of the mental variables are not metaphysically necessitated by the values of the physical variables.

committed to rejecting **Mental Causation**$_j$. Here is a different way to see the point. **Completeness**$_j$ says that $\mathcal{M}_p$ is strongly closed with respect to M. This means that there is no model containing all of the variables from $\mathcal{M}_p$, and any variables from M, in which any strongly sufficient causes lose that status. But there are two ways this can be true. The first is for there to be no such model containing all of the variables from $\mathcal{M}_p$ and any variables from M. This is what internalism is committed to, and it is consistent with the existence of a model in which M_1 is a cause. The second is for there to be such a model. This is what dualism is committed to, and it is not consistent with M_1 being a cause, in that model or any other.[40]

I turn now to Woodward's third question. What can be said to recommend interventionism formulated in terms of **IV*** over interventionism formulated in terms of **IV**, besides the fact that it facilitates a coherent non-reductionism?

One argument derives from the very motivation for the theory. If there is a single idea at the heart of interventionism, it is that the best way to understand the nature of causation is to theorise through the lens of an ideal experiment for detecting it (Woodward 2003, 14). So for example, when I introduced **Independent Manipulability** (Section 3.1), I said that it is motivated by the idea that for variables to coexist in a model, it must be possible to non-trivially test, for every pair, whether they are related by direct causation. The same motivation can be given for **IV***. The idea behind conditions $\mathbf{I}_3$ and $\mathbf{I}_4$ is that interventions should be independent of potential confounding causes. But just as the interventionist should say that variables are only candidates for being causally related if they can be independently manipulated, they should say that variables are only candidates for being potential confounding causes if they can be independently manipulated. $\mathbf{I}_3$ and $\mathbf{I}_4$ do not entail this constraint, but $\mathbf{I}_3^*$ and $\mathbf{I}_4^*$ do.

This is not a new form of argument. In his discussion of causal completeness in the context of probabilistic theories of causation, Sober (1999a, Section 2) considers theories according to which a positive causal factor must raise the probability of an effect in at least one background context, and lower it in none:

- **Positive Causal Factor**. C is a *positive causal factor* for E *iff* $P(E \mid C \,\&\, X_i) \geq P(E \mid \neg C \,\&\, X_i)$ for all background contexts X_i, with strict inequality for at least one X_i.

What counts as a background context? According to Sober, a necessary condition on a set of properties constituting a background context relative to a given cause and

[40] Here I disagree with Shapiro and Sober (2007), who suggest that a well-conceived argument for epiphenomenalism, under the assumption of interventionism, 'should aim to show that one class of properties does not affect a second class, not that the first has no effects at all' (241). This underestimates how strong the constraints are that completeness principles put on the sorts of properties that can be causes.

effect is that these probabilities are well defined. As he notes, this entails that when evaluating whether a necessitated property is a causal factor, necessitating properties cannot be part of any background context, since then $P(\neg C \,\&\, X_i)$ would be 0 and $P\left(E \mid \neg C \,\&\, X_i\right)$ not well defined. Sober's argument is identical to the argument I have just given for **IV***, transposed to the probabilistic case.[41]

The convergence of these arguments underscores that the issue concerning exclusion for difference-making theories of causation, under the assumption of internalism, concerns the contexts relative to which causes must make a difference. Must they make a difference controlling for *all* other causes, or must they make a difference controlling for all other *independent* causes? (Shapiro and Sober 2007, 241). I will briefly describe three other arguments that can be given for the second conception, before turning to externalism.

First, it is implicit in scientific practice that you do not need to control for necessitating variables in order to be justified in believing that necessitated variables are causes (Shapiro and Sober 2007; Shapiro 2010). As Sober puts it: 'This fact about scientific practice stands on its own' (1999a, 147). In this connection, it is also important to note that allowing necessitated properties to be causes does not mean that they trivially satisfy the requirements to be causes, simply in virtue of their being necessitated by causes (Segal and Sober 1991; Shapiro and Sober 2007, 256–259; Woodward 2015a, 2017). It is a substantive and difficult matter to determine whether a necessitated property meets the conditions for causation by the lights of interventionism under the assumption of **IV***.

Second, it is clear that in examples involving logical or conceptual necessitation between variables we do not need to hold fixed one variable in order to determine whether the other makes a difference (Woodward 2008a, 2015a). Indeed, since for any variable we can introduce others related to it in these ways, imposing this requirement would mean that no variables could possibly satisfy the requirements for being causally related. It can then be argued either that metaphysical necessitation is relevantly similar to those forms of dependence, or that **IV*** provides the correct theory in light of that fact.

Third, it can be argued that **IV** is, but **IV*** is not, subject to the argument that if there were no fundamental causal level, causation would drain away (Block 2003; Kim 2003).

I do not claim that these arguments are collectively decisive. But I do claim that in interventionism formulated with **IV***, the non-reductionist has a coherent and well-motivated theory of causation that entails the falsity of **Exclusion**$_j$. If the

[41] The same line of thought is arguably implicit in Eells (1991, 31). Similarly, Humphreys (1989, 74) requires that it be physically possible for the cause and its absence to occur relative to all background factors (for an application to the exclusion problem, see Henderson 1994). An analogous argument, in the context of a theory of causation along the lines of Mackie (1974), is given by Melnyk (2003, 137–138).

non-reductionist is an internalist, there is no obstacle to their endorsing **Mental Causation**$_j$.

4.3 Externalism

As it happens, very few non-reductionists can rest content at this point. For most of the conceptions of mental properties that motivate non-reductionism in the first place entail that internalism is false. So we need to consider arguments that target the externalist conception of mental properties. I will begin by discussing the externalist position generally, and then discuss some of the more concrete forms it may take when they become relevant.

The first point to note is that the externalist cannot make use of the same line of reasoning available to the internalist, who can appeal to **Non-Necessitation** in order to argue that there is no model containing both M_1 and P_1. Since the externalist by definition rejects the necessitation of M_1 by P_1, there is no obstacle to the existence of a model that contains both variables. Since the externalist remains a physicalist, they must thereby be committed to the existence of other physical variables that, together with P_1, necessitate M_1. For simplicity, I will use a single variable P_2 to represent these. So the externalist is committed to P_1 and P_2 together necessitating M_1, and neither P_1 nor P_2 alone necessitating M_1.

It now appears that epiphenomenalism looms. Consider model $\mathcal{M}_{PM1}$, containing variables P_1, M_1, and E_1. As I noted in Section 3.2, **Completeness**$_j$ entails that there is no path from M_1 to E_1 in $\mathcal{M}_{PM1}$. Moreover, this is not because there cannot be an intervention variable for M_1 with respect to E_1. There can be, but it must involve changing M_1 by changing P_2 (which is permissible according to **IV***). However, a difference of that sort cannot make any additional difference to E_1. At least in model $\mathcal{M}_{PM1}$, M_1 is epiphenomenal (see Figure 4.4).[42]

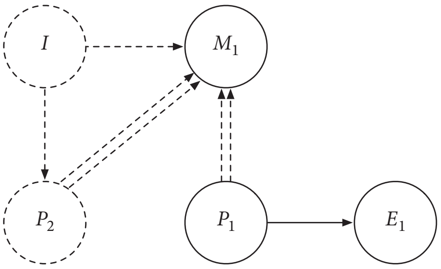

Figure 4.4 Diagram for $\mathcal{M}_{PM1}$.

[42] Arguments of this form are discussed by Block (1990), Worley (1993), and Rescorla (2012, 2014, Section 7).

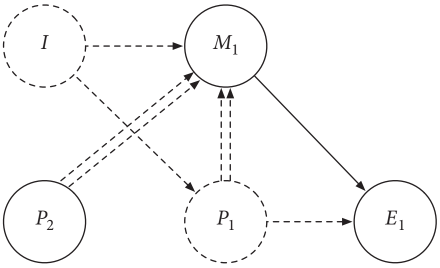

Figure 4.5 Diagram for $\mathcal{M}_{PM2}$.

Can we conclude that M_1 is epiphenomenal *simpliciter*? Not without additional argument. For according to the definitions provided in Section 2, any immediate inference from a variable not causing another in a model to its not causing it *simpliciter* is invalid. Recall that $X = x$ is an actual cause of $Y = y$ *simpliciter iff* there exists a model in which $X = x$ is an actual cause of $Y = y$. It follows that there is an asymmetry in what it takes to show that a variable value is or is not an actual cause of another. To show that a variable value *is* an actual cause, we simply need to identify a model in which it is. But to show that a variable value *is not* an actual cause, we need to show that there *does not exist* a model in which it is. What is needed to establish **Exclusion**$_j$ under the assumption of externalism is an argument that could establish the non-existence of a model in which M_1 is an actual cause of E_1, on the basis that it is not an actual cause in $\mathcal{M}_{PM1}$.

In fact, the externalist can do better than simply rejecting this inference. For they can exhibit a model in which M_1 is a cause of E_1. Consider model $\mathcal{M}_{PM2}$, containing variables P_2, M_1, and E_1. In this model, if we hold P_2 fixed at some particular value, then any intervention on M_1 must change P_1. So long as there exists a change of this sort that is associated with a change to E_1, then M_1 will be a direct cause of E_1, and for at least one state of the model will be an actual cause of E_1 (see Figure 4.5). Interventionism is therefore consistent not only with mental causation under the assumption of internalism, but with mental causation under the assumption of externalism.

For illustration, consider the case of content externalism. Here P_1 can be interpreted as representing neurophysiological properties, P_2 can be interpreted as representing content-fixing environmental properties, and M_1 can be interpreted as representing externally individuated mental properties, where the values of M_1 are metaphysically necessitated by the values of P_1 and P_2. $\mathcal{M}_{PM1}$ reveals the fact that if we hold fixed the neurophysiological properties, altering mental properties by altering the content-fixing environmental properties on which they partly depend would make no difference to behaviour. $\mathcal{M}_{PM2}$, on the other hand, reveals the fact that if we hold fixed the content-fixing environmental properties, altering mental properties by altering the neurophysiological properties on which they partly depend may make a difference to behaviour. According to interventionism, the existence of the former model does not entail that mental properties are not causes of behaviour, while the existence of the latter model entails that mental properties are causes of behaviour. If the non-reductionist is an externalist, there is no obstacle to their endorsing **Mental Causation**$_j$.

4.4 A Weakness in Interventionism

I have argued that interventionism allows both internalists and externalists to consistently accept **Non-Reductionism**$_j$, **Completeness**$_j$, and **Mental Causation**$_j$. In other words, an interventionist is entitled to reject **Exclusion**$_j$. However, in this section I suggest that attention to the externalist case reveals a weakness in interventionism.

The basic form of the problem is identified by Rescorla (2014, Section 11). As Rescorla notes, there are situations in which structurally identical models to $\mathcal{M}_{PM1}$ and $\mathcal{M}_{PM2}$ apply, and yet in which it is not the case that M_1 is a cause of E_1. Take, for example, a simple pocket calculator (Haugeland 1985, 121–123; Rescorla 2014, 180–181). The semantic properties instantiated by the calculator during the course of a calculation (let these be represented by M_1) are jointly determined by two factors: the physical properties it instantiates (let these be represented by P_1) and the interpretation to which they are subject (let this be represented by P_2). In this context, claims parallel to those concluding the previous section can now be introduced. Consider some particular output of the calculator (let this be represented by E_1). $\mathcal{M}_{PM1}$ reveals the fact that if we hold fixed the physical properties of the calculator, altering its semantic properties by altering the interpretation to which its physical properties are subject would make no difference to the output. $\mathcal{M}_{PM2}$, on the other hand, reveals the fact that if we hold fixed the interpretation to which its physical properties are subject, altering its semantic properties by altering its physical properties may make a difference to the output. But semantic properties don't cause the outputs of pocket calculators (Rescorla 2012). Something has gone wrong.

Moreover, the problem cannot be evaded by simply rejecting externalism. For there are many other situations in which structurally identical models to $\mathcal{M}_{PM1}$ and $\mathcal{M}_{PM2}$ apply, and in which M_1 is a cause of E_1. For example, consider a match struck in the presence of air, causing it to light. Let the presence of air be represented by M_1, the presence of oxygen be represented by P_1, the presence of all other constituents of air be represented by P_2, and the match lighting be represented by E_1. $\mathcal{M}_{PM1}$ reveals the fact that if we hold fixed the presence of oxygen, altering the presence of air by altering the presence of the other constituents would make no difference to the match lighting. $\mathcal{M}_{PM2}$, on the other hand, reveals the fact that if we hold fixed the presence of the other constituents, altering the presence of air by altering the presence of oxygen would make a difference to the match lighting.[43]

In sum, M_1 is a cause of E_1 in only some of the cases in which it appears that $\mathcal{M}_{PM2}$ applies, and the interventionist therefore owes us an account both of the difference between the cases, and why we should believe that mental properties fall on the right side of the line.[44]

5 Conclusion

It has been more than 25 years since the publication of Jaegwon Kim's *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation* (1998), the canonical investigation of causal exclusion principles. In summarising his discussion of counterfactual theories of causation, Kim wrote (71–72):

> […] what the counterfactual theorists need to do is to give an *account* of just what makes those mind-body counterfactuals we want for mental causation true, and show that on that account those counterfactuals we don't want, for example, epiphenomenalist counterfactuals, turn out to be false. Merely to point to the apparent truth, and acceptability, of certain mind-body counterfactuals as a vindication of mind-body causation is to misconstrue the philosophical task at hand … Such gestures only show that mind-body causation is part of what we normally take to be the real world; they go no further than a mere reaffirmation of our belief in the reality of mental causation. What we want—at least, what some of us are looking for—is a philosophical account of *how* it can be real in light of other principles and truths that seem to be forced upon us.

The 'principles and truths' Kim refers to here are those that he took as the premises in his arguments for causal exclusion principles. I have argued that an interventionist is entitled to reject those principles. But I have also argued that interventionists have not yet discharged the obligations that Kim here describes. In particular, they need to explain what is defective about the application of $\mathcal{M}_{PM2}$

[43] For discussion of this example in the context of mental causation, see Segal and Sober (1991), Tye (1991), Peacocke (1993a, 1993b), and Segal (2004, 2009). Note that examples of this sort place pressure on a condition Woodward (2008a, 2021a, 2021b, 2022) calls *realisation independence*, which requires that interventions must have the same effect no matter how they are realised. This condition seems to entail that in any case in which structurally identical models to $\mathcal{M}_{PM1}$ and $\mathcal{M}_{PM2}$ apply, M_1 is not a cause of E_1. See Hoffman-Kolss (2014, Section 5) for a different argument against realisation independence.

[44] A referee for this volume suggests that the notion of *conditional independence* recently discussed by Woodward (2021c, 2020, 2021a, 2021b, 2022) may help here. I do not think so, for two reasons. First, and in my view correctly, Woodward does not propose that conditional independence is a necessary condition on causation. Second, nothing I have said entails whether or not conditional independence is satisfied in either the case of mental properties or the case of the calculator, and I do not see any principled reason for saying it must always hold in the former and never in the latter.

to the case of the pocket calculator. Work of this sort must proceed along two paths: the development of principled constraints on when a causal model is appropriate for a given situation, as in, for example, Hitchcock (2001, 2004, 2012), Halpern and Hitchcock (2010), Halpern (2016), Woodward (2016), Blanchard and Schaffer (2017), McDonald (forthcoming, 2023), and Hoffmann-Kolss (this volume); and the application of these constraints to specific conceptions of the relationship between physical and mental properties, as in, for example, Rescorla (2014). Only when this cumulative case has been made, for the difference between pocket calculators and minds, can an interventionist claim to have a fully principled basis for rejecting **Exclusion.**[45]

References

Baumgartner, Michael. 2009. "Interventionist Causal Exclusion and Non-Reductive Physicalism." *International Studies in the Philosophy of Science* 23 (2): 161–178. http://doi.org/10.1080/02698590903006909.

Baumgartner, Michael. 2010. "Interventionism and Epiphenomenalism." *Canadian Journal of Philosophy* 40 (3): 359–383. https://doi.org/10.1080/00455091.2010.10716727.

Baumgartner, Michael. 2013. "Rendering Interventionism and Non-Reductive Physicalism Compatible." *Dialectica* 67 (1): 1–27. http://doi.org/10.1111/1746-8361.12008.

Baumgartner, Michael. 2018. "The Inherent Empirical Underdetermination of Mental Causation." *Australasian Journal of Philosophy* 96 (2): 335–350. https://doi.org/10.1080/00048402.2017.1328451.

Bennett, Karen. 2003. "Why the Exclusion Problem Seems Intractable, and How, Just Maybe, to Tract It." *Noûs* 37 (3): 471–497. http://doi.org/10.1111/1468-0068.00447.

Bennett, Karen. 2008. "Exclusion Again." In *Being Reduced: New Essays on Reduction, Explanation and Causation*, edited by Jesper Kallestrup and Jakob Hohwy, 280–305. Oxford: Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199211531.003.0015.

Blackburn, Simon. 1991. "Losing Your Mind: Physics, Identity, and Folk Burglar Prevention." In *The Future of Folk Psychology: Intentionality and Cognitive Science*, edited by John D. Greenwood, 196–225. Cambridge: Cambridge University Press.

Blanchard, Thomas, and Jonathan Schaffer. 2017. "Cause Without Default." In *Making a Difference: Essays on the Philosophy of Causation*, edited by Helen Beebee, Christopher Hitchcock, and Huw Price, 175–214. Oxford: Oxford University Press. http://doi.org/10.1093/oso/9780198746911.001.0001.

Block, Ned. 1990. "Can the Mind Change the World?" In *Meaning and Method: Essays in Honor of Hilary Putnam*, edited by George Boolos, 137–170. Cambridge: Cambridge University Press.

[45] Earlier versions of this paper were presented to a Mellon Mental Causation Workshop at Syracuse University in 2007, to the CUNY Cognitive Science Symposium in 2008, and to a workshop at Harvard University in 2019. I am grateful to the audiences on those occasions, and especially to Karen Bennett, Tony Dardis, David Rosenthal, Samuel Lee, and Ned Hall. I also thank Arif Ahmed, Ben Blumson, David Braddon-Mitchell, Michael Brent, John Campbell, Brandon Carey, Shamik Dasgupta, Sommer Hodson, Adam Kay, Doug Kutach, David Macarthur, Kevin McCain, Peter Menzies, Huw Price, and Nandi Theunissen for discussion or feedback, and Michael Baumgartner and Michael Rescorla for very helpful correspondence, on earlier versions. More recently, I thank Michael Tye for sending me his work on this topic, and Alastair Wilson, Katie Robertson, Henry Taylor, Mike Hicks, Ben Curry, and two referees for this volume for their advice on preparing the version that appears in this volume.

Block, Ned. 2003. “Do Causal Powers Drain Away?” *Philosophy and Phenomenological Research* 67 (1): 133–150. http://doi.org/10.1111/j.1933-1592.2003.tb00029.x.

Bontly, Thomas D. 2002. “The Supervenience Argument Generalizes.” *Philosophical Studies* 109 (1): 75–96. http://doi.org/10.1023/A:1015786809364.

Buckareff, Andrei A. 2011. “Intralevel Mental Causation.” *Frontiers of Philosophy in China* 6 (3): 402–25. http://doi.org/10.2307/44259314.

Buckareff, Andrei A. 2012. “An Action-Theoretic Problem for Intralevel Mental Causation.” *Philosophical Issues* 22: 89–105. http://doi.org/10.2307/41683062.

Burge, Tyler. 1993. “Mind-Body Causation and Explanatory Practice.” In *Mental Causation*, edited by John Heil and Alfred Mele, 97–120. Oxford: Oxford University Press.

Burge, Tyler. 2007. “Postscript: Mind-Body Causation and Explanatory Practice.” In *Foundations of Mind*, 2: 363–382. Philosophical Essays. Oxford: Oxford University Press.

Campbell, John. 2007. “An Interventionist Approach to Causation in Psychology.” In *Causal Learning: Psychology, Philosophy, Computation*, edited by Alison Gopnik and Laura Schulz, 58–66. New York: Oxford University Press.

Campbell, John. 2007. 2008. “Causation in Psychiatry.” In *Philosophical Issues in Psychiatry: Explanation, Phenomenology and Nosology*, edited by Kenneth S. Kendler and Josef Parnas, 196–215. Baltimore MD: Johns Hopkins University Press.

Cartwright, Nancy. 1983. *How the Laws of Physics Lie*. Oxford: Oxford University Press. http://doi.org/10.1093/0198247044.001.0001.

Cartwright, Nancy. 1994. “Fundamentalism Vs. The Patchwork of Laws.” *Proceedings of the Aristotelian Society* 94 (1): 279–292. https://doi.org/10.1093/aristotelian/94.1.279.

Crane, Tim. 2008. “Causation and Determinable Properties: On the Efficacy of Colour, Shape, and Size.” In *Being Reduced: New Essays on Reduction, Explanation and Causation*, edited by Jesper Kallestrup and Jakob Hohwy, 176–195. Oxford: Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199211531.003.0011.

Davidson, Donald. 1963. “Actions, Reasons, and Causes.” *Journal of Philosophy* 60 (23): 685–700. http://doi.org/10.2307/2023177.

Davidson, Donald. 1967. “Causal Relations.” *Journal of Philosophy* 64 (21): 691–703. https://doi.org/10.2307/2023853.

Davidson, Donald. 1970. “Mental Events.” In *Experience and Theory*, edited by Lawrence Foster and Joe William Swanson, 79–101. Amherst MA: University of Massachusetts Press. http://doi.org/10.1093/0199246270.003.0011.

Davidson, Donald. 1995. “Laws and Cause.” *Dialectica* 49 (2–4): 263–279. https://doi.org/10.1111/j.1746-8361.1995.tb00165.x.

Eberhardt, Frederick. 2014. “Direct Causes and the Trouble with Soft Interventions.” *Erkenntnis* 79 (4): 755–777. https://doi.org/10.1007/s10670-013-9552-2.

Eberhardt, Frederick, and Richard Scheines. 2007. “Interventions and Causal Inference.” *Philosophy of Science* 74 (5): 981–995. https://doi.org/10.1086/525638.

Eells, Ellery. 1991. *Probabilistic Causality*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511570667.

Eronen, Markus I., and Daniel S. Brooks. 2014. “Interventionism and Supervenience: A New Problem and Provisional Solution.” *International Studies in the Philosophy of Science* 28 (2): 185–202. https://doi.org/10.1080/02698595.2014.932529.

Fenton-Glynn, Luke. 2021. *Causation*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108588300.

Field, Hartry. 1978. “Mental Representation.” *Erkenntnis* 13 (1): 9–61. http://doi.org/10.1007/BF00160888.

Field, Hartry. 2003. “Causation in a Physical World.” In *The Oxford Handbook of Metaphysics*, edited by Michael J. Loux and Dean W. Zimmerman, 435–460. Oxford: Oxford University Press. http://doi.org/10.1093/oxfordhb/9780199284221.003.0015.

Gallow, J. Dmitri. 2022. “The Metaphysics of Causation.” In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Stanford: Stanford University. https://plato.stanford.edu/archives/fall2022/entries/causation-metaphysics/.

Gebharter, Alexander. 2017a. “Causal Exclusion Without Physical Completeness and No Overdetermination.” *Abstracta* 10: 3–14.

Gebharter, Alexander. 2017b. “Causal Exclusion and Causal Bayes Nets.” *Philosophy and Phenomenological Research* 95 (2): 353–375. https://onlinelibrary.wiley.com/doi/abs/10.1111/phpr.12247.

Gibbons, John. 2006. “Mental Causation Without Downward Causation.” *Philosophical Review* 115 (1): 79–103. http://doi.org/10.1215/00318108-115-1-79.

Goldman, Alvin I. 1969. “The Compatibility of Mechanism and Purpose.” *Philosophical Review* 78 (4): 468–482. https://doi.org/10.2307/2184199.

Goldman, Alvin I. 1970. *A Theory of Human Action*. Englewood Cliffs NJ: Prentice Hall.

Halpern, Joseph Y. 2016. “Appropriate Causal Models and the Stability of Causation.” *The Review of Symbolic Logic* 9 (1): 76–102. http://doi.org/10.1017/S1755020315000246.

Halpern, Joseph Y., and Christopher Hitchcock. 2010. “Actual Causation and the Art of Modeling.” In *Heuristics, Probability and Causality: A Tribute to Judea Pearl,* edited by Rina Dechter, Hector Geffner, and Joseph Y. Halpern, 383–406. London: College Publications.

Halpern, Joseph Y., and Judea Pearl. 2005. “Causes and Explanations: A Structural-Model Approach. Part i: Causes.” *British Journal for the Philosophy of Science* 56 (4): 843–887. http://doi.org/10.1093/bjps/axi147.

Haugeland, John. 1985. *Artificial Intelligence: The Very Idea*. Cambridge MA: MIT Press.

Hausman, Daniel M. 2005. “Causal Relata: Tokens, Types, or Variables?” *Erkenntnis* 63 (1): 33–54. http://doi.org/10.1007/s10670-005-0562-6.

Heckman, James J. 2005. “The Scientific Model of Causality.” *Sociological Methodology* 35 (1): 1–98. http://doi.org/10.1111/j.0081-1750.2006.00163.x.

Henderson, David K. 1994. “Accounting for Macro-Level Causation.” *Synthese* 101 (2): 129–156. https://doi.org/10.1007/BF01064014.

Hendry, Robin. 2006. “Is There Downward Causation in Chemistry?” In *Philosophy of Chemistry: Synthesis of a New Discipline*, edited by Davis Baird, Eric Scerri, and Lee McIntyre, 242: 173–189. Boston Studies in the Philosophy of Science. Dordrecht: Springer. https://doi.org/10.1007/1-4020-3261-7_9.

Hendry, Robin. 2010a. “Emergence Vs. Reduction in Chemistry.” In *Emergence in Mind*, edited by Cynthia Macdonald and Graham Macdonald, 205–21. Oxford: Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199583621.003.0014.

Hendry, Robin. 2010b. “Ontological Reduction and Molecular Structure.” *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 41 (2): 183–191. https://doi.org/10.1016/j.shpsb.2010.03.005.

Hendry, Robin. 2017. “Prospects for Strong Emergence in Chemistry.” In *Philosophical and Scientific Perspectives on Downward Causation*, edited by Michele Paolini Paoletti and Francesco Orilia, 146–163. New York: Routledge. https://doi.org/10.4324/9781315638577-9.

Hitchcock, Christopher. 2001. “The Intransitivity of Causation Revealed in Equations and Graphs.” *Journal of Philosophy* 98 (6): 273–299. https://doi.org/10.2307/2678432.

Hitchcock, Christopher. 2004. “Routes, Processes, and Chance-Lowering Causes.” In *Cause and Chance: Causation in an Indeterministic World*, edited by Phil Dowe and Paul Noordhof, 138–152. London: Routledge. https://doi.org/10.4324/9780203494660.

Hitchcock, Christopher. 2007. “Prevention, Preemption, and the Principle of Sufficient Reason.” *Philosophical Review* 116 (4): 495–532. http://doi.org/10.1215/00318108-2007-012.

Hitchcock, Christopher. 2009. “Causal Modelling.” In *The Oxford Handbook of Causation*, edited by Helen Beebee, Christopher Hitchcock, and Peter Menzies, 299–314. Oxford: Oxford University Press. http://doi.org/10.1093/oxfordhb/9780199279739.003.0015.

Hitchcock, Christopher. 2012. “Events and Times: A Case Study in Means-Ends Metaphysics.” *Philosophical Studies* 160 (1): 79–96. http://doi.org/10.1007/s11098-012-9909-4.

Hitchcock, Christopher. 2023. “Causal Models.” In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Stanford: Stanford University. https://plato.stanford.edu/archives/spr2023/entries/causal-models/.

Hoffman-Kolss, Vera. 2014. "Interventionism and Higher-Level Causation." *International Studies in the Philosophy of Science* 28 (1): 49–64. http://doi.org/10.1080/02698595.2014.915653.

Horgan, Terence. 1997. "Kim on Mental Causation and Causal Exclusion." *Noûs* 31: 165–184. https://doi.org/10.1111/0029-4624.31.s11.8.

Humphreys, Paul W. 1989. *The Chances of Explanation*. Princeton: Princeton University Press. https://doi.org/10.1515/9781400860760.

Ismael, Jenann. 2009. "Probability in Deterministic Physics." *Journal of Philosophy* 106 (2): 89–108. https://doi.org/10.5840/jphil2009106214.

Ismael, Jenann. 2011. "A Modest Proposal about Chance." *Journal of Philosophy* 108 (8): 416–442. https://doi.org/10.5840/jphil2011108822.

Jackson, Frank, and Philip Pettit. 1988. "Functionalism and Broad Content." *Mind* 97 (387): 381–400. http://doi.org/10.1093/mind/XCVII.387.381.

Jackson, Frank, and Philip Pettit. 1990a. "Causation in the Philosophy of Mind." *Philosophy and Phenomenological Research* 50: 195–214. http://doi.org/10.2307/2108039.

Jackson, Frank, and Philip Pettit. 1990b. "Program Explanation: A General Perspective." *Analysis* 50 (2): 107–117. http://doi.org/10.2307/3328853.

Kaiserman, Alex. 2020. "Interventionism and Mental Surgery." *Erkenntnis* 85 (4): 919–935. https://doi.org/10.1007/s10670-018-0059-8.

Kendler, Kenneth S., and John Campbell. 2009. "Interventionist Causal Models in Psychiatry: Repositioning the Mind-Body Problem." *Psychological Medicine* 39 (6): 881–887. http://doi.org/10.1017/S0033291708004467.

Kim, Jaegwon. 1973. "Causes and Counterfactuals." *Journal of Philosophy* 70 (17): 570–572. https://doi.org/10.2307/2025312.

Kim, Jaegwon. 1989. "Mechanism, Purpose, and Explanatory Exclusion." *Philosophical Perspectives* 3: 77–108. http://doi.org/10.2307/2214264.

Kim, Jaegwon. 1997. "Does the Problem of Mental Causation Generalize?" *Proceedings of the Aristotelian Society* 97 (3): 281–297. http://doi.org/10.1111/1467-9264.00017.

Kim, Jaegwon. 1998. *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. Cambridge MA: MIT Press. https://doi.org/10.7551/mitpress/4629.001.0001.

Kim, Jaegwon. 2003. "Blocking Causal Drainage and Other Maintenance Chores with Mental Causation." *Philosophy and Phenomenological Research* 67 (1): 151–176. http://doi.org/10.1111/j.1933-1592.2003.tb00030.x.

Kim, Jaegwon. 2005. *Physicalism, or Something Near Enough*. Princeton: Princeton University Press. https://doi.org/10.1515/9781400840847.

Korb, Kevin B., Lucas R. Hope, Ann E. Nicholson, and Karl Axnick. 2004. "Varieties of Causal Intervention." In *PRICAI 2004: Trends in Artificial Intelligence*, edited by Chengqi Zhang, Hans W. Guesgen, and Wai-Kiang Yeap, 322–331. Berlin: Springer. https://doi.org/10.1007/978-3-540-28633-2_35.

List, Christian, and Peter Menzies. 2009. "Nonreductive Physicalism and the Limits of the Exclusion Principle." *Journal of Philosophy* 106 (9): 475–502. http://doi.org/10.2307/20620197.

Loewer, Barry. 2008. "Why There Is Anything Except Physics." In *Being Reduced: New Essays on Reduction, Explanation and Causation*, edited by Jesper Kallestrup and Jakob Hohwy, 149–163. Oxford: Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199211531.001.0001.

Loewer, Barry. 2009. "Why Is There Anything Except Physics?" *Synthese* 170 (2): 217–233. http://doi.org/10.1007/s11229-009-9580-2.

Loewer, Barry. 2015. "Mental Causation: The Free Lunch." In *Qualia and Mental Causation in a Physical World: Themes from the Philosophy of Jaegwon Kim*, edited by David Sosa, Terence Horgan, and Marcelo Sabatés, 40–63. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139939539.004.

Lowe, E. J. 2000. "Causal Closure Principles and Emergentism." *Philosophy* 75 (294): 571–585. https://doi.org/10.1017/S003181910000067X.

Lowe, E. J. 2003. "Physical Causal Closure and the Invisibility of Mental Causation." In *Physicalism and Mental Causation: The Metaphysics of Mind and Action*, edited by Sven Walter and Heinz-Dieter Heckmann, 137–154. Exeter: Imprint Academic.

Mackie, John L. 1974. *The Cement of the Universe: A Study of Causation*. Oxford: Oxford University Press. http://doi.org/10.1093/0198246420.001.0001.

Malcolm, Norman. 1968. "The Conceivability of Mechanism." *Philosophical Review* 77 (1): 45–72. https://doi.org/10.2307/2183182.

Markowetz, Florian, Steffen Grossmann, and Rainer Spang. 2005. "Probabilistic Soft Interventions in Conditional Gaussian Networks." In *IN 10TH AI/STATS*, edited by Robert Cowell and Zoubin Ghahramani, 214–221. Savannah Hotel, Barbados: The Society for Artificial Intelligence and Statistics. https://www.gatsby.ucl.ac.uk/aistats/fullpapers/139.pdf.

Marras, Ausonio. 1998. "Kim's Principle of Explanatory Exclusion." *Australasian Journal of Philosophy* 76 (3): 439–451. https://doi.org/10.1080/00048409812348551.

McDermott, Michael. 1995. "Redundant Causation." *British Journal for the Philosophy of Science* 46 (4): 523–544. http://doi.org/10.1093/bjps/46.4.52.

McDermott, Michael. 2002. "Causation: Influence Versus Sufficiency." *Journal of Philosophy* 99 (2): 84–101. https://doi.org/10.5840/jphil200299219.

McDonald, Jenn. forthcoming. "Essential Structure for Causal Models." *Australasian Journal of Philosophy*, forthcoming.

McDonald, Jenn. 2023. "Causal Models and Contrastivism.", unpublished manuscript.

Mellor, D. H. 1995. *The Facts of Causation*. London: Routledge.

Melnyk, Andrew. 2003. *A Physicalist Manifesto: Thoroughly Modern Materialism*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511498817.

Menzies, Peter. 2003. "The Causal Efficacy of Mental States." In *Physicalism and Mental Causation: The Metaphysics of Mind and Action*, edited by Sven Walter and Heinz-Dieter Heckmann, 195–224. Exeter: Imprint Academic.

Menzies, Peter. 2010. "Reasons and Causes Revisited." In *Naturalism and Normativity*, edited by Mario De Caro and David Macarthur, 142–170. New York: Columbia University Press.

Menzies, Peter, and Christian List. 2010. "The Causal Autonomy of the Special Sciences." In *Emergence in Mind*, edited by Cynthia Macdonald and Graham Macdonald, 108–128. Oxford: Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199583621.003.0008.

Noordhof, Paul. 1997. "Making the Change: The Functionalist's Way." *British Journal for the Philosophy of Science* 48 (2): 233–250. http://doi.org/10.1093/bjps/48.2.233.

Papineau, David. 2001. "The Rise of Physicalism." In *Physicalism and Its Discontents*, edited by Carl Gillett and Barry Loewer, 3–36. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511570797.002.

Papineau, David. 2002. *Thinking about Consciousness*. Oxford: Oxford University Press. http://doi.org/10.1093/0199243824.001.0001.

Patterson, Sarah. 2005. "Epiphenomenalism and Occasionalism: Problems of Mental Causation, Old and New." *History of Philosophy Quarterly* 22 (3): 239–257.

Peacocke, Christopher. 1993a. "Externalist Explanation." *Proceedings of the Aristotelian Society* 93 (3): 203–230. https://doi.org/10.1093/aristotelian/93.1.203.

Peacocke, Christopher. 1993b. "Review of the Imagery Debate." *Philosophy of Science* 60 (4): 675–677. https://doi.org/10.1086/289774.

Pearl, Judea. 2000. *Causality*. Cambridge: Cambridge University Press.

Pereboom, Derk, and Hilary Kornblith. 1991. "The Metaphysics of Irreducibility." *Philosophical Studies* 63 (2): 125–145. http://doi.org/10.1007/BF00381684.

Polger, Thomas W., Lawrence A. Shapiro, and Reuben Stern. 2018. "In Defense of Interventionist Solutions to Exclusion." *Studies in History and Philosophy of Science Part A* 68 (April): 51–57. https://doi.org/10.1016/j.shpsa.2018.01.012.

Price, Huw, and Richard Corry. 2007. *Causation, Physics and the Constitution of Reality: Russell's Republic Revisited*. Edited by Huw Price and Richard Corry. Oxford: Oxford University Press.

Putnam, Hilary. 1975. "Philosophy and Our Mental Life." In *Mind, Language and Reality*, 2: 291–303. Philosophical Papers. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511625251.016.

Raatikainen, Panu. 2010. "Causation, Exclusion, and the Special Sciences." *Erkenntnis* 73 (3): 349–363. https://doi.org/10.1007/s10670-010-9236-0.

Rescorla, Michael. 2012. "Are Computational Transitions Sensitive to Semantics?" *Australasian Journal of Philosophy* 90 (4): 703–721. https://doi.org/10.1080/00048402.2011.615333.

Rescorla, Michael. 2014. "The Causal Relevance of Content to Computation." *Philosophy and Phenomenological Research* 88 (1): 173–208. http://doi.org/10.1111/j.1933-1592.2012.00619.x.

Rescorla, Michael. 2018. "An Interventionist Approach to Psychological Explanation." *Synthese* 195 (5): 1909–1940. https://doi.org/10.1007/s11229-017-1553-2.

Robb, David, John Heil, and Sophie Gibb. 2023. "Mental Causation." In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Stanford: Stanford University. https://plato.stanford.edu/archives/spr2023/entries/mental-causation/.

Russell, Bertrand. 1913. "On the Notion of Cause." *Proceedings of the Aristotelian Society* 13 (1): 1–26. https://doi.org/10.1093/aristotelian/13.1.1.

Schaffer, Jonathan. 2016. "The Metaphysics of Causation." In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Stanford: Stanford University. https://plato.stanford.edu/archives/fall2016/entries/causation-metaphysics/.

Schlosser, Markus E. 2009. "Non-Reductive Physicalism, Mental Causation, and the Nature of Action." In *Reduction: Between the Mind and the Brain*, edited by Alexander Hieke and Hannes Leitgeb, 73–90. Frankfurt: Ontos Verlag. https://doi.org/10.1515/9783110328851.73.

Segal, Gabriel. 2004. "Reference, Causal Powers, Externalist Intuitions, and Unicorns." In *The Externalist Challenge*, edited by Richard Schantz, 329–344. Berlin: Walter de Gruyter. https://doi.org/10.1515/9783110915273.329.

Segal, Gabriel. 2009. "The Causal Inefficacy of Content." *Mind and Language* 24 (1): 80–102. http://doi.org/10.1111/j.1468-0017.2008.01354.x.

Segal, Gabriel, and Elliott Sober. 1991. "The Causal Efficacy of Content." *Philosophical Studies* 63 (1): 1–30. http://doi.org/10.1007/BF00375995.

Shapiro, Lawrence A. 2010. "Lessons from Causal Exclusion." *Philosophy and Phenomenological Research* 81 (3): 594–604. http://doi.org/10.1111/j.1933-1592.2010.00382.x.

Shapiro, Lawrence A., and Elliott Sober. 2007. "Epiphenomenalism: The Dos and the Don'ts." In *Thinking about Causes: From Greek Philosophy to Modern Physics*, edited by Peter Machamer and Gereon Wolters, 235–264. Pittsburgh-Konstanz Series in the Philosophy and History of Science. Pittsburgh: University of Pittsburgh Press.

Sklar, Lawrence. 2003. "Dappled Theories in a Uniform World." *Philosophy of Science* 70 (2): 424–441. http://doi.org/10.1086/375476.

Sober, Elliott. 1999a. "Physicalism from a Probabilistic Point of View." *Philosophical Studies* 95 (1–2): 135–174. http://doi.org/10.1023/A:1004519608950.

Sober, Elliott. 1999b. "The Multiple Realizability Argument Against Reductionism." *Philosophy of Science* 66 (4): 542–564. https://doi.org/10.1086/392754.

Statham, Georgie. 2018. "Woodward and Variable Relativity." *Philosophical Studies* 175 (4): 885–902. https://doi.org/10.1007/s11098-017-0897-2.

Stern, Reuben, and Benjamin Eva. 2023. "Antireductionist Interventionism." *British Journal for the Philosophy of Science* 74 (1): 241–267. https://doi.org/10.1086/714792.

Stich, Stephen. 1983. *From Folk Psychology to Cognitive Science: The Case Against Belief.* Cambridge MA: MIT Press.

Strevens, Michael. 2007. "Review of Woodward, Making Things Happen." *Philosophy and Phenomenological Research* 74 (1): 233–249. http://doi.org/10.1111/j.1933-1592.2007.00012.x.

Thomasson, Amie. 1998. "A Nonreductivist Solution to Mental Causation." *Philosophical Studies* 89 (2–3): 181–195. http://doi.org/10.1023/A:1004280812099.

Tye, Michael. 1991. *The Imagery Debate*. Cambridge MA: MIT Press.

Weslake, Brad. unpublished. "A Partial Theory of Actual Causation."

Weslake, Brad. 2010. “Explanatory Depth.” *Philosophy of Science* 77 (2): 273–294. http://doi.org/10.1086/651316.

Weslake, Brad. 2017. “Difference-Making, Closure and Exclusion.” In *Making a Difference: Essays on the Philosophy of Causation*, edited by Helen Beebee, Christopher Hitchcock, and Huw Price, 215–231. Oxford: Oxford University Press. http://doi.org/10.1093/oso/9780198746911.003.0011.

Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation.* New York: Oxford University Press. http://doi.org/10.1093/0195155270.001.0001.

Woodward, James. 2008a. “Mental Causation and Neural Mechanisms.” In *Being Reduced: New Essays on Reduction, Explanation and Causation*, edited by Jesper Kallestrup and Jakob Hohwy, 218–262. Oxford: Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199211531.003.0013.

Woodward, James. 2008b. “Response to Strevens.” *Philosophy and Phenomenological Research* 77 (1): 193–212. http://doi.org/10.1111/j.1933-1592.2008.00181.x.

Woodward, James. 2015a. “Interventionism and Causal Exclusion.” *Philosophy and Phenomenological Research* 91 (2): 303–347. http://doi.org/10.1111/phpr.12095.

Woodward, James. 2015b. “Methodology, Ontology, and Interventionism.” *Synthese* 192 (11): 3577–3599. https://doi.org/10.1007/s11229-014-0479-1.

Woodward, James. 2016. “The Problem of Variable Choice.” *Synthese* 193 (4): 1047–1072. https://doi.org/10.1007/s11229-015-0810-5.

Woodward, James. 2017. “Intervening in the Exclusion Argument.” In *Making a Difference: Essays on the Philosophy of Causation*, edited by Helen Beebee, Christopher Hitchcock, and Huw Price, 251–268. Oxford: Oxford University Press. http://doi.org/10.1093/oso/9780198746911.003.0013.

Woodward, James. 2020. “Causal Complexity, Conditional Independence, and Downward Causation.” *Philosophy of Science* 87 (5): 857–867. https://doi.org/10.1086/710631.

Woodward, James. 2021a. “Downward Causation and Levels.” In *Levels of Organization in the Biological Sciences*, edited by Daniel S. Brooks, James DiFrisco, and William C. Wimsatt, 175–194. Cambridge MA: MIT Press. https://doi.org/10.7551/mitpress/12389.003.0013.

Woodward, James. 2021b. “Downward Causation Defended.” In *Top-Down Causation and Emergence*, edited by Jan Voosholz and Markus Gabriel, 217–52. Cham: Springer. https://doi.org/10.1007/978-3-030-71899-2_9.

Woodward, James. 2021c. “Explanatory Autonomy: The Role of Proportionality, Stability, and Conditional Irrelevance.” *Synthese* 198 (1): 237–65. https://doi.org/10.1007/s11229-018-01998-6.

Woodward, James. 2022. “Levels, Kinds and Multiple Realizability: The Importance of What Does Not Matter.” In *Levels of Reality in Science and Philosophy*, edited by Stavros Ioannidis, Gal Vishne, Meir Hemmo, and Orly Shenker, 261–292. Cham: Springer. https://doi.org/10.1007/978-3-030-99425-9_14.

Worley, Sara. 1993. “Mental Causation and Explanatory Exclusion.” *Erkenntnis* 39 (3): 333–358. https://doi.org/10.1007/BF01128507.

Yablo, Stephen. 1992. “Mental Causation.” *Philosophical Review* 101 (2): 245–280. https://doi.org/10.2307/2185535.

Yablo, Stephen. 1997. “Wide Causation.” *Noûs* 31: 251–281. https://doi.org/10.1111/0029-4624.31.s11.12.