Exclusion Excluded*

Brad Weslake†

June 7, 2011

Contents

*Earlier versions of this paper were presented to a Mellon Mental Causation Workshop at Syracuse University, Friday 30 November–Sunday 2 December 2007, and to the CUNY Cognitive Science Symposium, 14 November 2008. I am grateful to both audiences, and especially to Karen Bennett, Tony Dardis and David Rosenthal, for feedback on those occasions. 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 on earlier versions. I owe special thanks to Michael Baumgartner and Michael Rescorla for extremely helpful correspondence on the penultimate version.

†Department of Philosophy
University of Rochester
Box 270078
Rochester, NY 14627-0078
bradley.weslake@rochester.edu
http://www.rochester.edu/college/faculty/bweslake/

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:

> *Nonreductionism* Mental properties are distinct from, though metaphysically necessitated by, physical properties.
>
> *Completeness* Every event has a complete causal explanation in terms of physical 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, unless it is causally overdetermined.
>
> *Mental Causation* There exist causal explanations of events in terms of mental properties that are not cases of causal overdetermination[1].

In this paper I examine the prospects for a principled rejection of *Exclusion*. Following Horgan (1997) and Bennett (2003, p. 473), I will refer to this position as *causal compatibilism*. Causal compatibilism is a popular position[2]. However, compatibilism has typically been defended in the absence of an independently justified general framework for thinking about causation and causal explanation, and to that extent has yet to be made good. In particular, compatibilists have yet to provide a principled reason for thinking of overdetermination in a way that justifies the rejection of *Exclusion*[3].

The framework for thinking about causation and explanation I employ is a justly influential theory developed by James Woodward (2003), which I will refer to as *interventionism*. I argue that interventionism entails that *Exclusion* is false. The structure

[1] Nothing here turns on my formulation of the issue in terms of explanation, properties and events. I defer discussion of completeness and overdetermination to §3. 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 does not make any difference to the argument below. I assume throughout that "explanation" is a factive term, and that in a causal explanation all explanans properties are causally relevant to the explanandum.

[2] Bennett (2003) cites Goldman (1969), Blackburn (1991), Pereboom and Kornblith (1991), Yablo (1992), Burge (1993), Mellor (1995, pp. 103–104), Horgan (1997), Noordhof (1997) and Yablo (1997), to which we can add van Gulick (1992), Baker (1993), Menzies (2003), Ross and Spurrett (2005), Shapiro and Sober (2007) (see also Shapiro 2010) and Woodward (2008*a*).

[3] Here I agree with Bennett (2003, §2). Bennett herself is a clear exception, and I hope that her influence on my approach to the problem is obvious.

of the paper is as follows. In §2 I introduce interventionism. In §3 I use the interventionist framework to provide definitions of overdetermination and completeness, show how this improves on an influential account of overdetermination defended by Bennett (2003), and use these definitions to reformulate the exclusion problem. In §4 I identify two conditions which are individually sufficient for the falsity of *Exclusion*, but argue that they ought to be rejected by most non-reductionists. However I then argue that *Exclusion* should also be rejected when the first condition is false. So the non-reductionist has a general argument against *Exclusion*, whether or not they accept this condition. I conclude in §5.

2 Interventionism Introduced

Central to the interventionist framework is the notion of a causal model. 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 the variables on the right hand side. The possible values of variables in a causal model must represent particulars capable of being set to different values by interventions, but the framework is otherwise consistent with a range of different metaphysical views concerning the nature of the causal relata[4]. More formally, a causal model is an ordered pair $\langle \mathscr{V}, \mathscr{E} \rangle$, where $\mathscr{V}$ is a set of variables and $\mathscr{E}$ a set of equations, and every variable appears on the left hand side of exactly one equation. I will refer to a possible assignment of values to all variables in a model as a *state* of the model, and will talk freely of actual and possible variable values, changes to variable values, states and changes of state of models. This sort of talk should be interpreted throughout as reflecting corresponding actual or possible changes in 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[5].

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

[4]For simplicity, I will at times nevertheless stay with talk of properties and events.

[5]As I will explain below, this does not entail that for all causal relations, if one veridical model entails that the relation obtains, then all veridical models entail that the relation obtains.

on the intervention[6]. For a given causal model $\mathscr{M}$, a variable X is a (type-level) cause of a variable Y *iff* there is some state of $\mathscr{M}$ for which an intervention on X would change the value of Y. The particular value of X in turn figures in the (token-level) causal explanation for the particular value of Y in $\mathscr{M}$ *iff* X figures in the answer to at least one interventionist *what-if-things-had-been-different-question* (*w-question*) with respect to Y[7]. X is a (type-level) cause of Y *simpliciter iff* there is a model in which it is so represented. Likewise, the value of X is a (token-level) cause of Y *simpliciter iff* there is a model in which it is so represented.

More precise formulations of these notions will be elaborated below in the context of examples that will help to make them clear, but this brief sketch is already 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. Note that this is not a version of causal anti-realism. Causal claims are not made true or false by causal models, they are made true by the counterfactuals regarding experimental interventions that are represented by those models[8]. 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.

[6]Interventions must also be statistically independent of the values of other variables in the model. I provide a formal definition in §4.2.

[7]The precise form of the *w-questions* relevant to causation and explanation is provided in §3.1.

[8]This may seem obvious, but the following mistake is routinely made: "A model is in the mind. As a consequence, causality is in the mind" (Heckman 2005, p. 2).

3 Exclusion Reformulated

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

> *Nonreductionism*$_{\text{i}}$ The values of mental variables are distinct from, though metaphysically necessitated by, the values of physical variables.
>
> *Completeness*$_{\text{i}}$ For every event, there exists a causal model containing only physical variables which specifies a complete explanation of that event.
>
> *Exclusion*$_{\text{i}}$ 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, unless it is a model in which the event is causally overdetermined.
>
> *Mental Causation*$_{\text{i}}$ There exists a causal model in which a mental variable explains an event in which that event is not causally overdetermined.

In order to evaluate these claims we need to understand how interventionism defines the crucial notions of overdetermination, completeness and variable distinctness. In the remainder of this section I explain these notions in turn, and then employ them to more precisely reformulate the exclusion problem in the interventionist framework.

3.1 Overdetermination

Exclusion$_{\text{i}}$ requires clarification of the notion of an event being causally overdetermined in a model. I will work up to a definition of this notion by way of considering a similar proposal made by Karen Bennett (2003).

Bennett (§4) proposes a necessary condition (NC) on overdetermination, which specifies that e is overdetermined by c_1 and c_2 only if the following counterfactuals are non-vacuously true:

$$(c_1 \& \neg c_2) \boxdot\!\!\rightarrow e \qquad (O_1)$$

$$(c_2 \& \neg c_1) \boxdot\!\!\rightarrow e \qquad (O_2)$$

This is a natural and intuitive criterion for overdetermination. As Bennett notes, the *prima facie* appeal of the two parts of this condition is "simply that they capture the reasoning we engage in when we want to distinguish cases of genuine overdetermination from cases of joint causation" (p. 477). Be that as it may, Thomson-Jones (2007) has raised a convincing counterexample to NC that he calls the *staggered firing*

squad. Suppose Billy and Suzy both fire at Victim, though Suzy fires first from a long distance away (c_1) and Billy fires second from a short distance away (c_2). The bullets arrive simultaneously at the same place with the same momentum and kill Victim (e). This is as paradigmatic a case of overdetermination as there is. Suppose however that it is also the case that Billy has an unmanifested squeamish (or subservient) disposition, so that had Suzy not fired, Billy would have fired inaccurately and Victim would not have died. But then (O_2) is false—if Billy had fired and Suzy had not, Victim would not have died. However it is plausible that we still have a case of overdetermination—the presence of this unmanifested disposition can hardly make a difference to whether there is overdetermination. So NC is false[9].

Here is how to represent the situation in a causal model. Call our variables B for Billy's firing, S for Suzy's firing, and V for Victim's death. Let B be 0 if Billy does not fire, 1 if he fires inaccurately, and 2 if he fires accurately. Let S be 0 if Suzy does not fire and 1 if Suzy does fire. Let V be 1 if Victim dies and 0 if Victim does not die. The equations specifying the counterfactual relationships between variables in the model are[10]:

$$V := S \vee (B = 2) \tag{EQ1}$$

$$B := S + 1 \tag{EQ2}$$

$$S := 1 \tag{EQ3}$$

In the actual world, $B = 2$, $S = 1$ and $V = 1$. According to the definition of *actual cause* which I will assume, the procedure for identifying actual causes is as follows[11].

[9]Note that the case is explicitly designed to not require a backtracking reading of any counterfactuals (*cf.* Bennett 2003, pp. 477–478). This example also shows that the "counterfactual signature of redundant causation" proposed by Hitchcock (2011, §3) is mistaken.

[10]Note that these equations are not to be interpreted symmetrically. ":=" represents an assignment of values to variables on the left hand side in the manner specified on the right hand side; "=" is a function returning 1 if the two sides are equal and 0 if otherwise; "$\vee$" is a function returning 1 if either side is 1 and 0 otherwise.

[11]Here I mostly follow Woodward (2003, §2.7). Similar definitions of actual cause are endorsed by Hitchcock (2001), and Halpern and Pearl (2005*a*; 2005*b*), to whom the idea can be credited. A similar strategy, though pursued without appeal to causal models, is also adopted by Yablo (2002; 2004). As Oisín Deery reminded me, Woodward (p. 84, fn. 46) is ambivalent on the question of whether a more complicated definition is required in order to handle cases of trumping. Hitchcock (2007) also amends the definition, in response to counterexamples in Hiddleston (2005) and Björnsson (2007) (a similar amendment is suggested by Menzies 2004). There are other alleged counterexamples in Hall (2007) (see also Glymour et al. 2009), to which Hitchcock (2009) replies. For my own account of actual causation, see Weslake (ms). Fortunately, the difference between these various theories does not

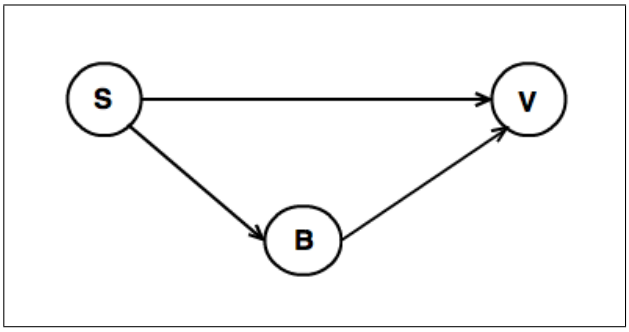

Figure 1: *Directed Graph for Suzy, Billy and Victim*

First we need the (type-level) notion of a *direct cause* (Woodward 2003, p. 55)[12]:

> *(DC)* X is a direct cause of Y in model $\mathscr{M}$ *iff* there is a possible intervention on X that would change Y (or the probability distribution of Y)[13] when all other variables in $\mathscr{M}$ besides X and Y are held fixed at some combination of values by interventions.

For example, if we hold B fixed at 0, changing S from 0 to 1 would change V from 0 to 1. So S is a direct cause of V. By similar reasoning, we arrive at the full set of direct causes in our example: S is a direct cause of B and V; and B is a direct cause of V. Second we need the (type-level) notion of a *directed path* (*ibid*, p. 42):

> *(P)* A *directed graph* for $\mathscr{M}$ is an ordered pair $\langle V, E\rangle$ where V is a set of vertices that correspond to the set of variables in $\mathscr{M}$ and E a set of directed edges connecting these vertices. A directed edge from vertex X to vertex Y represents that X directly causes Y in $\mathscr{M}$. A sequence of variables $\{V_1 \ldots V_n\}$ is a *directed path* from V_1 to V_n in $\mathscr{M}$ *iff* for all $i (1 \leqslant i < n)$ there is a directed edge from V_i to V_{i+1} in the directed graph for $\mathscr{M}$.

From here on, *path* should be read as equivalent to *directed path*. For example, there are two paths from S to V, one that does not include any intermediate variables and

make a difference to the arguments below, since they all deliver the correct result in cases of symmetric overdetermination.

[12] 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.

[13] For the remainder of the paper, all talk of actual or counterfactual changes to variables should be interpreted with this qualification to handle the probabilistic case. Matters are simplified by leaving it implicit.

one that includes variable B. This is easily seen by constructing a diagram with the same structure as the associated directed graph, as for example in Figure 1 where circles correspond to vertices (variables) and arrows correspond to directed edges (direct causes). As an aside, another definition that will be important below is most helpfully introduced at this point (*ibid*, p. 59):

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

Third we need to define the (token-level) notion of a *redundancy range* (*ibid*, p. 83)[14]:

> *(RR)* For a path P from X to Y in model $\mathscr{M}$, define $V_1 \ldots V_n$ as all variables in $\mathscr{M}$ that are not on P. Values $v_1 \ldots v_n$ are on the redundancy range for V_i with respect to P *iff* no intervention setting $V_1 \ldots V_n$ to $v_1 \ldots v_n$ while holding the actual value of X fixed would result in a change to the actual values of any variables on P.

Now we can define the following necessary and jointly sufficient conditions on a (token-level) *actual cause* $X = x$ of $Y = y$ relative to a model $\mathscr{M}$ (*ibid*, p. 84):

> *(AC*$^*_1$*)* The actual value of $X = x$ and the actual value of $Y = y$.
>
> *(AC*$^*_2$*)* For each path P_i from X to Y in $\mathscr{M}$ and for each possible combination of values for the direct causes Z_i of Y in $\mathscr{M}$ that are not on this path and that are in the redundancy range of Z_i with respect to P_i, determine whether there is an intervention on X that will change the value of Y. If there is at least one such intervention, *(AC*$^*_2$*)* is satisfied.

The definition appears at first glance a complicated one, but the intuitive idea is simple: an actual cause along a path is determined by whether there is an intervention that would change the effect along that path in background conditions which are irrelevant in the circumstances to the path. We are now also able to define the interventionist *w-questions* relevant to explanation: the particular value of X figures in the (token-level) causal explanation for the particular value of Y in $\mathscr{M}$ *iff* X is an actual

[14] My formulation is closer to Hitchcock (2001, p. 290), as Woodward's version generates problems with late preemption. Halpern and Pearl (2005*a*, §A.2) show that this version is too strict, but the details will not matter for what follows (see Weslake ms for discussion).

cause of Y in $\mathscr{M}$; the *w-questions* relevant to explanation concern the hypothetical results of the interventions specified by *(AC*$_2$)*.

Before returning to our example, note several consequences of these definitions that will be important later. First, notice that if $X = x$ is an actual cause of $Y = y$ in $\mathscr{M}$ then X is a contributing cause of Y in $\mathscr{M}$. Second, notice that each of these definitions is relativised to a causal model. The corresponding de-relativised definitions are as follows[15]: 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 §5.

Now to return to the case of Billy and Suzy, we need to apply *(AC*)* to each potential actual cause. Starting with Suzy, it is clear that all possible values of B are in the redundancy range of S, since no intervention on B while holding S fixed would result in a change to V[16]. So we are permitted to test for the efficacy of S by setting $B = 0$ or $B = 1$. But if $B = 0$ or $B = 1$, setting $S = 0$ would result in $V = 0$. So $S = 1$ is an actual cause of $V = 1$. Turning to Billy, we see that the situation is symmetrical. It is clear that all possible values of S are in the redundancy range of B, since no intervention on S while holding B fixed would result in a change to V. So we are permitted to test for the efficacy of B by setting $S = 0$. But if $S = 0$, setting $B = 0$ would result in $V = 0$. So $B = 2$ is an actual cause of $V = 1$.

Given this definition of *actual cause*, the natural way to define overdetermination should now be obvious. Intuitively, an actual cause is sufficient for an effect in the sense required for it to overdetermine that effect just in case any other actual or potential causes of the effect did not make or could not have made any further difference to the effect than the difference made by the actual cause in question. That is, just in case the effect would have occurred regardless of the other values any other actual or potential causes of the effect could have taken. Let us say then that an actual cause $X = x$ along path P is a (token-level) *weakly sufficient actual cause* for an effect $Y = y$ in $\mathscr{M}$ *iff* all possible values $v_1 \ldots v_n$ for all variables $V_1 \ldots V_n$ not on P in $\mathscr{M}$ are on the redundancy range for $X = x$ and $Y = y$ with respect to P[17]. *Overdetermina-*

[15] Here I follow Woodward (2008*b*), though he does not provide a de-relativised definition of actual cause.

[16] On the assumption, which is built into my specification of the model, that if Suzy fires and Billy fires inaccurately, Victim dies.

[17] Pearl (2000, §10.2) calls this condition *sustenance* and Halpern and Pearl (2005*a*, p. 855) say that when it holds $X = x$ *strongly causes* $Y = y$. I use the term "weak sufficiency" since a stronger notion will be required for the formulation of *Completeness* in §3.2. The generalisation to jointly weakly sufficient causes is obvious, though I will assume for simplicity for the remainder of the paper that it is always

tion (token-level) is in turn defined as occurring when there are at least two weakly sufficient actual causes of an effect in a model[18].

Note that according to this criterion S and B overdetermine V[19]. We have already seen how it is that they are adjudged actual causes, so all that remains is to show that they are both weakly sufficient. Recall our observation that all possible values of each variable are on the redundancy range of the other. By our definition of weak sufficiency then, each is weakly sufficient. Since S and B are both weakly sufficient actual causes of V, they overdetermine V. The interventionist definition of overdetermination succeeds where Bennett's NC does not[20].

3.2 Completeness

Exclusion$_{\text{i}}$ also requires clarification of the notion of a complete explanation for an event. To begin, note that *Completeness*$_{\text{i}}$ should not be interpreted as being equivalent to the claim that every event can be 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[21]; and 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 interventionist causal models[22]. Now in making these points, I do not mean to weaken the support that causal completeness assump-

single variables that provide sufficient causes.

[18] In fact a further condition is required so that the definition does not entail that chains of sufficient causes overdetermine their effects: two causes only overdetermine an effect if neither is on every path from the other to the effect. This detail will not play any role in what follows.

[19] Here and in what follows, I frequently omit explicit reference to the actual values of variables when making claims concerning the token-level relations that hold between them, when context makes those values clear.

[20] The key to the solution is that the interventionist model provides us with a rationale for considering a counterfactual where we explicitly have Billy shoot accurately without Suzy shooting even though Billy would not have shot accurately had Suzy not shot. The rationale for considering this counterfactual is provided by the very notion of an intervention, which allows us to hypothetically manipulate a variable independently of how that variable would ordinarily be caused (Woodward 2008*a*, pp. 240–241 also notes the importance of this feature of interventionism in this context).

[21] See Russell (1912–1913), Field (2003) and the essays in Price and Corry (2007).

[22] One reason is that causal models do not allow the representation of continuous processes (Strevens 2007, pp. 242–244). Strevens puts the point by saying that interventionist causal models "represent less of causal reality than is actually out there" (p. 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 (*cf.* Woodward 2008*b*, pp. 210–211).

tions 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*$_{\mathrm{i}}$ does not say, let us examine what it does say. The exclusion problem is often framed in terms of causal sufficiency rather than completeness, since overdetermination is typically defined in terms of sufficiency, so any defensible notion of completeness must bear some close relationship to a notion of causal sufficiency. There are at least five different notions of sufficiency that can be discriminated with the interventionist framework, only one of which, I will argue, is suitable for figuring in formulations of *Completeness*. These notions are, in order of strength:

> *Sufficiency in the Circumstances in a Model* A cause is sufficient in the circumstances for an effect in a model *iff* it is a non-overdetermining actual cause of the effect in that model[23].
>
> *Weak Sufficiency in a Model* A cause is weakly sufficient for an effect in a model *iff* it is a *weakly sufficient actual cause* of that effect in the sense defined above. (That is, *iff* it is an actual cause of that effect along path P and all possible values $v_1 \ldots v_n$ for all variables $V_1 \ldots V_n$ not on P are on the redundancy range for the actual values of the cause and effect with respect to P).
>
> *Weak Sufficiency in an Effectively Closed Model* Call a model $\mathscr{M}_{\mathrm{F}}$ framed in variables characterised by vocabulary F *effectively closed* with respect to variables characterised by vocabulary G with respect to an effect *iff* for every model $\mathscr{M}_{\mathrm{FGi}}$ constructed by adding variables from G to $\mathscr{M}_{\mathrm{F}}$, no weakly sufficient causes of the effect for any state of $\mathscr{M}_{\mathrm{F}}$ are not also weakly sufficient causes of the effect for the corresponding state of $\mathscr{M}_{\mathrm{FGi}}$[24]. A cause is weakly sufficient for an effect in an effectively closed model $\mathscr{M}_{\mathrm{F}}$ with respect to G *iff* it is weakly sufficient for the effect in $\mathscr{M}_{\mathrm{F}}$.

[23] This is a somewhat permissive definition of circumstantial sufficiency. A more strict definition would make it equivalent to Woodward's *(AC)* (2003, p. 77), which is equivalent to the definition of causation defined in terms of "Act" in Hitchcock (2001, pp. 286–287). The difference will not matter for what follows.

[24] To keep our emphasis on metaphysical questions, let us suppose that different vocabularies correspond to different types of particulars. It is a standard presupposition here that there exists some relatively clear distinction between physical properties and mental properties, and therefore the particulars involving them. I make no stand on how these property types are to be demarcated.

Strong Sufficiency in an Effectively 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. A cause is strongly sufficient for an effect in an effectively closed model $\mathscr{M}_F$ with respect to G *iff* it is strongly sufficient for the effect in $\mathscr{M}_F$.

Nomological Sufficiency An event A is nomologically sufficient for an event B *iff* the occurrence of A and the laws of nature 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)[25].

It is important to note a difference between the first two definitions and the last three. The first two are defined in a wholly model-internal manner, while the last three involve model-external facts. Indeed, the final notion of *Nomological Sufficiency* makes no reference to causal models at all. 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.

The first two model-internal notions of sufficiency on the other hand are inadequate because they do not even pose a *prima facie* exclusion problem. Consider the stronger notion, *Weak Sufficiency in a Model*. Suppose that we have a model $\mathscr{M}_P$ framed in variables characterised by physical vocabulary P which is not effectively closed with respect to variables characterised by mental vocabulary M. Suppose further that $\mathscr{M}_P$ specifies a cause that is weakly sufficient for some action. Since $\mathscr{M}_P$ is not effectively closed, there is simply no problem in supposing that there exists some model $\mathscr{M}_{PM}$ constructed by adding variables characterised by M to $\mathscr{M}_P$, in which the M-variables specify an actual cause for the action and the P-variables do not specify a cause that is weakly sufficient for the action. What this possibility reveals is that the first two 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 an Effectively Closed Model*.

In fact it needs to be stronger. *Weak Sufficiency in an Effectively Closed Model* is compatible with the actual values of $\mathscr{M}_P$ specifying weakly sufficient actual causes

[25]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 lightcone in spacetime, if any events are going to turn out to be nomologically sufficient for any others (Field 2003).

that remain weakly sufficient in $\mathscr{M}_{\mathrm{PM}}$, and yet it being the case that no *alternative* values of the P-variables would have specified a weakly sufficient cause in either $\mathscr{M}_{\mathrm{P}}$ or $\mathscr{M}_{\mathrm{PM}}$. 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. This does not properly 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 an Effectively Closed Model*.

Sufficiency weak or strong in an effectively closed model is a relative matter, in the sense that a cause could be sufficient in an effectively closed model with respect to one set of variables, but not with respect to a different set of variables. While it would not make a difference to the argument below if we strengthened our understanding of completeness so that it involved the idea of a model effectively 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. 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 effectively closed with respect to the M-variables. If I had all of the physical information about you my knowledge of counterfactuals 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[26]. 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 effectively closed with respect to the variables of the biological causal model, the biological causal model is effectively closed with respect to the variables of the neuroscientific causal model, and so on up the hierarchy of the sciences and never *vice versa*[27]. And so if the exclusion problem arises for mental variables it also arises for any variables not appearing in some maximally

[26] 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).

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

effectively closed causal model[28].

Finally, while this will also make no difference to the argument below, note that my definitions of *Exclusion*$_{\text{i}}$ and *Mental Causation*$_{\text{i}}$ do not require that in order for mental variables to causally 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, §5) thinks that anything less than sufficiency would endanger the "full-fledged causal efficacy of the mental" (p. 481), granting it merely "a derivative efficacy" (p. 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[29]. If an event has a complete physical cause, mental causes are often said to have "no work left to do" (Kim 1998, p. 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, p. 41; 2005, p. 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 *were* 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[30].

3.3 Variable Distinctness

Finally, *Exclusion*$_{\text{i}}$ requires clarification of the notions of distinct variables and variable values.

Firstly, since variables are best understood as features of causal models rather than as features of the world, framing the problem in terms of variables may appear to have trivialised the issue. In order for *Nonreductionism*$_{\text{i}}$ to occupy the proper role in the exclusion problem it needs to express a claim about the world, not about our

[28] Here I side with Bontly (2002) against Kim (1997; 1998).

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

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

ways of representing the world. I will suppose then that variable values are distinct only if they represent distinct particulars, and that distinct vocabularies correspond to distinct types of particulars[31]:

> *Value Distinctness* Variable values are distinct only if they represent distinct particulars.

Secondly, 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 §3.1. Whether X is a direct cause of Y in $\mathscr{M}$ depends, by the definition of *(DC)*, on whether there exists an intervention on X that will change Y when all other variables in $\mathscr{M}$ are held fixed at some combination of values by interventions[32]. This implies an independence condition on variables coexisting in a model: if X is a direct cause of Y in $\mathscr{M}$ 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 $\mathscr{M}$ 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* whether *(DC)* holds. According to Woodward (2003, §3.5; 2011), the relevant sense of possibility here is *at a minimum* metaphysical possibility. The corresponding independence condition on variables coexisting in a model $\mathscr{M}$ is this:

> *Independent Manipulability* It is metaphysically possible that every proper subset of the variables in $\mathscr{M}$ be set to every combination of their possible values by independent interventions[33].

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[34]. When I refer to the interventionist theory of

[31] See fn. 24.

[32] In interpreting the condition in this way I agree with Baumgartner (2009). Woodward (2011) confirms that this was his intended interpretation.

[33] Woodward (2011) calls this condition (without the proper subset clause) *Independent Fixability*.

[34] For more detailed discussion of the reasons for imposing constraints of this sort, and proposals for further necessary conditions on variables, see Hitchcock (2001; 2004) and Hitchcock and Halpern (2010).

causation in what follows, I will take it to include all of the definitions provided in §3.1, as well as *Value Distinctness* and *Independent Manipulability*.

3.4 The Non-Embedding Constraint

The exclusion problem in the interventionist setting now has this shape:

> *Nonreductionism*$_{\mathrm{j}}$ 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.
>
> *Completeness*$_{\mathrm{j}}$ For every event, there exists an effectively closed causal model $\mathscr{M}_{\mathrm{Pi}}$ with respect to M which specifies a strongly sufficient actual cause for that event.
>
> *Exclusion*$_{\mathrm{j}}$ If there exists an effectively closed causal model $\mathscr{M}_{\mathrm{F}}$ with respect to variables characterised by vocabulary G which specifies a strongly sufficient actual cause for an event, there exists no other causal model $\mathscr{M}_{\mathrm{G}}$ specifying an actual cause for that event in terms of variables characterised by vocabulary G, unless it is a model in which the event is causally overdetermined.
>
> *Mental Causation*$_{\mathrm{j}}$ There exists a causal model $\mathscr{M}_{\mathrm{MI}}$ containing a mental variable specifying an actual cause of an event in which that variable does not overdetermine the event.

A deceptively simple solution to the exclusion problem now suggests itself. The solution is simply that *Exclusion*$_{\mathrm{j}}$ is not a principle that is entailed by the interventionist theory of causation, as I have presented it. In the remainder of this section I will challenge this solution in order to develop a more adequate formulation of the problem.

The idea behind the simple solution is this. Let us suppose for the sake of argument that $\mathscr{M}_{\mathrm{MI}}$ is a model specified by folk psychology and $\mathscr{M}_{\mathrm{PI}}$ an effectively closed model with respect to M specifying strongly sufficient causes in terms of physical variables[35]. And suppose that there is no overlap in the variables appearing in $\mathscr{M}_{\mathrm{PI}}$ and the variables appearing in $\mathscr{M}_{\mathrm{MI}}$, apart from a set of variables drawn from

[35]There are of course issues concerning whether folk psychological practice can reasonably be described as involving a commitment to an interventionist causal model of the kind specified by *Mental Causation*$_{\mathrm{j}}$, and if so whether the model is veridical (see Godfrey-Smith 2005 for a general discussion of the idea of folk psychological competence as involving facility with a model). The argument here only requires the claim that there exists some such veridical model.

vocabulary A representing a range of possible actions[36]. The important point is that none of these suppositions is inconsistent with supposing that both $\mathscr{M}_{\mathrm{P_I}}$ and $\mathscr{M}_{\mathrm{M_I}}$ specify non-overdetermining actual causes for some A-variable. Intuitively, the threat of overdetermination is the threat that physical and mental causes overdetermine actions. Overdetermination however, as defined in the interventionist framework, is a model-internal notion. So the fact that there are two models providing actual causes for some event, even if both provide strongly sufficient actual causes for that event, does not entail that the event is overdetermined. And there is no principle in the interventionist framework disallowing the possibility of two such models.

I see two ways in which this picture might be challenged. The first challenge, which I reject, is to deny on general metaphysical grounds that it is possible for there to be causal models which stand in this relationship to each other. It might be argued that only $\mathscr{M}_{\mathrm{P_I}}$ *really* specifies causes, and that $\mathscr{M}_{\mathrm{M_I}}$ merely specifies *explanations*, or some other weak cousin of causation. This might be because only $\mathscr{M}_{\mathrm{P_I}}$ promises to be maximally predictively accurate and therefore maximally effectively closed (Davidson, 1963; 1967; 1970; 1995), or because $\mathscr{M}_{\mathrm{P_I}}$ is a truthmaker for $\mathscr{M}_{\mathrm{M_I}}$ (Robb and Heil 2008, §5.3; Crane 2008)[37], or for more recherché metaphysical reasons (Jackson and Pettit, 1988; 1990*a*; 1990*b*). I think that the arguments for these claims are unsound, but for present purposes want simply to 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 leave them to one side[38].

The second challenge, which I accept, is to concede that there might exist models standing in this relationship to each other, but to argue that a further constraint must be satisfied in order for $\mathscr{M}_{\mathrm{M_I}}$ to specify genuine non-overdetermining causes. Call the constraint the *non-embedding* constraint. The non-embedding constraint is that there must not be a causal model $\mathscr{M}_{\mathrm{PM_I}}$ containing variables drawn from both $\mathscr{M}_{\mathrm{P_I}}$ and $\mathscr{M}_{\mathrm{M_I}}$ in which the P-variables and M-variables are overdetermining causes of the A-variable in $\mathscr{M}_{\mathrm{PM_I}}$. Surely, the challenge suggests, the existence of such a model

[36] This supposition again shows that $\mathscr{M}_{\mathrm{P_I}}$ ought not to be understood as provided by some fundamental physical theory, since these say nothing about actions as described in the vocabulary of folk psychology. Some have wished to press this point into a defence of mental causation (see for example Gibbons 2006), but I find the strategy unpromising and assume that there exists some such model $\mathscr{M}_{\mathrm{P_I}}$. If you are suspicious of actions, substitute behaviours.

[37] Something like this worry seems to haunt Kim's discussion in many places, and I have encountered it repeatedly in discussion, but arguments for the claim do not often appear explicitly in the literature.

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

would undermine the claims of $\mathscr{M}_{M_I}$ to have identified a non-overdetermining mental cause. Now there is clearly something correct in this challenge, since it would be a hollow victory for the interventionist framework if for all cases of mental causation there existed a model in which the M and P variables overdetermine the effect[39]. In fact, there are two ways to read the challenge. On the first way, the challenge is to the notion of overdetermination defined above. The thought is that if there can be two causal models, one in which an effect is not overdetermined by two variables and one in which it is, then it is overdetermined *simpliciter*[40]. On the second way, the challenge is not to the notion of overdetermination defined above but rather to what is required in order to answer the exclusion problem. The thought is that if we are using a model-internal notion of overdetermination, then in order to show that there is no overdetermination in the case of mental causation, we need to show that it is not the case that for each case of mental causation we can construct a model in which the effect is overdetermined by its mental cause. Now this is a choice that makes a difference to how we think about overdetermination in general, but it does not make any difference to the arguments in this paper, so I will simply take the second option, which I prefer, without argument. The upshot of accepting the non-embedding constraint in this form is reflected in the following revised formulations of *Exclusion* and *Mental Causation*, stated with the other assumptions required to generate the exclusion problem:

> *Nonreductionism*$_j$ 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.
>
> *Completeness*$_j$ For every event, there exists an effectively closed causal model $\mathscr{M}_{P_i}$ with respect to M which specifies a strongly sufficient actual cause for that event.
>
> *Exclusion*$_k$ If there exists an effectively closed causal model $\mathscr{M}_F$ with respect to variables characterised by vocabulary G which specifies a strongly sufficient

[39] Of course, there will be cases where such overdetermination does genuinely occur, as for example when we consider a model which incorporates Suzy's mental state and Billy's physical state in relation to Victim's death. We are concerned here with the question of whether such a model can *always* be constructed.

[40] Indeed, it might be urged more generally that it had better not be possible for there to be one model in which a variable satisfies some causal concept and another in which it does not. I agree for some causal concepts and disagree for others. However, this amounts to a question regarding the adequacy of the interventionist framework in general and so I will set it aside (though see §5). For the question of whether interventionism implies a problematic form of model relativity, see Strevens (2007; 2008), Woodward (2008*b*) and McCain (2010).

actual cause for an event, there exists no other causal model $\mathscr{M}_{G_I}$ specifying an actual cause for that event in terms of variable G_I, unless there exists a model $\mathscr{M}_{FG_I}$ in which the event is overdetermined by the F-variables and G_I.

Mental Causation$_{k_I}$ There exists a causal model $\mathscr{M}_{M_I}$ containing a mental variable M_x specifying an actual cause of an event represented by variable A_x.

Mental Causation$_{k_2}$ There does not exist a causal model $\mathscr{M}_{PM_I}$ in which A_x is overdetermined by the P-variables and M_x.

4 Compatibilism Examined

Having completed our reformulation of the exclusion problem, we are now in a position to evaluate it. In §4.1 I identify a condition concerning the connection between the F variables and G_i that is sufficient for the falsity of *Exclusion*$_k$. In §4.2 I consider an objection to this claim, and argue that responding to the objection requires an amendment to the definition of intervention provided by Woodward (2003). In §4.3 I show that if, as most non-reductionists believe, the condition is false, then there is a *prima facie* sound argument for the conclusion that mental properties are epiphenomenal with respect to physical properties. In §4.4 I consider and reject a response to this argument suggested by Bennett (2003). Along the way I identify a second condition sufficient for the rejection of *Exclusion*, which I argue most non-reductionists are also not in a position to endorse. Finally, in §4.5 I diagnose the mistaken assumption in the argument for epiphenomenalism and thereby provide a general argument against *Exclusion*.

4.1 Subvenience Sufficiency

It will be helpful to develop the argument in this subsection in a number of steps. As a first step, consider one way in which we might try to construct $\mathscr{M}_{PM_I}$. Suppose that $\mathscr{M}_{P_I}$ contains all P-variables and is effectively closed with respect to variables in all other vocabularies. That is, suppose that there is a single causal model that specifies strongly sufficient physical causes for all events and would lose no sufficient causes when incorporating variables from any other vocabularies. And as above, suppose that there is no overlap in the variables appearing in $\mathscr{M}_{P_I}$ and the variables appearing in $\mathscr{M}_{M_I}$, apart from a set of variables drawn from vocabulary A representing possible actions. Now suppose for *reductio* that we incorporate all variables from $\mathscr{M}_{M_I}$ and $\mathscr{M}_{P_I}$ into the same model. Recall that according to *Independent Manipulability*, in order for variables P_x and M_x to appear in a single causal model, it must be possible

to intervene on M_x while holding P_x and all other variables except one fixed by interventions, and *vice versa*[41]. Suppose further that the single variable we are not holding fixed does not form part of the subvenience basis for M_x[42]. Since the relevant notion of interventional possibility is at a minimum metaphysical possibility, such an intervention will only be possible if the relationship between the physical variables and mental variables is *weaker* than metaphysical necessity. However according to *Nonreductionism*$_j$ the values of the physical variables metaphysically necessitate the values of the mental variables. So there are no interventions on M_x that are not also interventions on P_x[43]. So assuming the form of dependence definitive of the non-reductive physicalist, $\mathscr{M}_{PM_I}$ cannot be constructed by combining all variables from $\mathscr{M}_{M_I}$ and $\mathscr{M}_{P_I}$ into a single model. The upshot is the following necessary condition on variables coexisting in a model:

> *Non-Supervenience* A causal model cannot contain a variable with a possible value that metaphysically supervenes on a possible combination values of any proper subset of the other variables in the model.

Interventionism requires that the variables in a model satisfy *Independent Manipulability*, which entails *Non-Supervenience*. *Non-Supervenience* in turn prevents the variables from $\mathscr{M}_{P_I}$ and $\mathscr{M}_{M_I}$ from being incorporated into a single model[44].

Of course, *Completeness*$_j$ does not require that there be a single causal model that provides strongly sufficient causes for every event. Instead, it merely requires that for each event there exists at least one physical causal model specifying a strongly sufficient cause. So as a second step, consider what is required to be true in order for the preceding pattern of argument to apply to every such causal model $\mathscr{M}_{Pi}$ for some event that has a mental cause M_x. In order for *Non-Supervenience* to prevent M_x

[41] No doubt there are many physical and mental variables that can be manipulated independently of each other in this way. I can change my mind about whether to shoot or pass the ball independently of whether my arm is moving this way or that. Our question is whether *all* physical and mental variables are independently manipulable in this way.

[42] I suppose it goes without saying, but that there exists such a variable is an *extremely* weak assumption. First, presumably nearly all other M-variables are in this class, and any reasonable physicalist theory of mind will also place most other P-variables in the class. Second, we may suppose that the variable in question is one of the A-variables that is a candidate effect of M_x. *This* variable had better not be part of the subvenience basis for M_x.

[43] Note that there may well be interventions on P_x that are not interventions on M_x, since it is implausible that every fine grained change at the subvening physical level makes for some difference at the supervening mental level.

[44] For a more detailed argument for *Non-Supervenience* from interventionist assumptions, see Woodward (2011). In what follows I will omit reference to the proper subset clause (see fn. 42).

from being incorporated into any such model $\mathscr{M}_{Pi}$, it must be the case that every such $\mathscr{M}_{Pi}$ contains physical variables the values of which metaphysically necessitate the value of M_x. Now this is a *much* stronger assumption than *Nonreductionism*$_j$ alone. Where *Nonreductionism*$_j$ merely requires that the values of the P-variables metaphysically necessitate the values of the M-variables, the present assumption requires, in addition, that the values of all M-variables which cause some event metaphysically supervene on the values of every set of P-variables providing strongly sufficient causes for that event. This condition I will call *Subvenience Sufficiency*:

> *Subvenience Sufficiency* The values of all M-variables which actually cause some event metaphysically supervene on the values of every set of P-variables providing strongly sufficient causes for that event.

If *Subvenience Sufficiency* is true, then by *Non-Supervenience* there will be no model which contains both a set of physical variables which provide a strongly sufficient cause for A_x, and a mental variable M_x which provides a cause for A_x[45]. Since there is no model containing these variables together, they do not overdetermine A_x. That is, I have argued that interventionism and *Subvenience Sufficiency* entail *Mental Causation*$_{k2}$. This, together with *Mental Causation*$_{k1}$, entails that *Exclusion*$_k$ is false. If *Subvenience Sufficiency* is true for the relationship between physical properties and mental properties, then non-reductive physicalism and causal compatibilism are, on the interventionist framework, made for each other.

But is *Subvenience Sufficiency* true? To my knowledge Bennett (2003) is the only person to have noticed the importance of *Subvenience Sufficiency* for the case for compatibilism. Moreover, as she points out (*ibid*, §6), most non-reductive physicalists accept theories of mind on which it is false. Suppose our candidate strongly sufficient cause P_x represents some neurophysiological property[46]. If role functionalism is true, then it is false that any mental properties supervene on P_x alone, for mental properties also supervene on the causal role filled by P_x, which in turn supervenes in part on the contingent background conditions in which P_x is instantiated. Likewise if content externalism is true, for then the corresponding mental properties supervene not only on P_x but also on whatever external conditions are required to fix external content. The class of non-reductionists who endorse one or both of these theses is large, and they do not exhaust the ways in which *Subvenience Sufficiency* can fail[47]. So most

[45] In this sense the present option for rejecting *Exclusion* is analogous to maintaining that Bennett's (O_2) is vacuously true.

[46] To avoid prolixity, I will use P_x to refer to both the property and the associated variable.

[47] Bennett notes for example that it also fails if any version of conceptual role semantics is true.

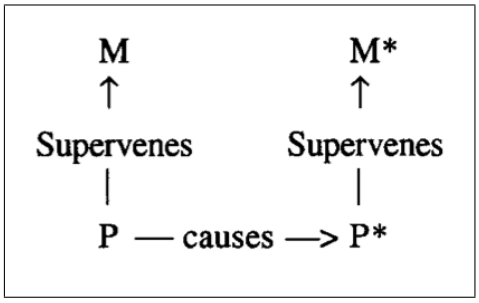

Figure 2: *Subvenience Sufficiency in Kim Diagrams*

non-reductionists are not in a position to reject *Exclusion*$_{\text{K}}$ by accepting *Subvenience Sufficiency*.

Let us pause to note some interim conclusions. First, it is important to notice that the argument to this point is enough to show that if interventionism is true, *Exclusion* is false in general. You cannot argue from there being two sufficient causes of an event to that event being overdetermined, since this does not entail that there exists a model in which that event is overdetermined.

Second, the argument to this point is also enough to show that *Exclusion* is false in one of the central contexts in which Jaegwon Kim, the most prominent defender of the principle, has so often defended it. From Kim's work, we are all familiar with diagrams taking the form of Figure 2[48]. This diagram and others like it display in vivid form the *Subvenience Sufficiency* assumption, for here one and the same physical event is depicted as both causally sufficient for the effect in question and as metaphysically sufficient for the candidate mental cause. There is an irony in the fact that the most prominent defender of *Exclusion* so frequently makes an assumption on which, if interventionism is true, it is invalid.

Third, a comment on related treatments of interventionism and mental causation. Earlier examinations of the exclusion argument from the interventionist perspective have also emphasised the impossibility of intervening on supervening properties while holding fixed their subvenience bases[49]. However, the emphasis of this

[48]This version is from Kim (2003, p. 159). The situation depicted is a recurring focus of the arguments in both Kim (1998) and Kim (2005). For reasons that should be clear from the foregoing discussion, these diagrams cannot be interpreted as representing directed graphs. For one way of expanding the interventionist framework to allow for the representation of non-causal relationships in directed graphs, see Woodward (2011).

[49]See in particular Shapiro and Sober (2007), who write: "It is not relevant, or even coherent, to ask what will happen if one wiggles X while holding fixed the micro-supervenience base of X". See also Woodward (2008*a*) and Shapiro (2010).

work has been on showing that in order to *test* whether mental properties are causal, we should not hold fixed their subvenience bases. I have argued further that considerations of the same type explain why we should not think of mental properties and their subvenience bases as overdetermining their effects. Overdetermination is a model-internal notion, so variables that cannot appear together in a model should not be thought of as overdetermining their effects. However, the argument I have given to this point also shows that these earlier interventionist arguments are subject to the important restriction of subvenience sufficiency. Since most non-reductionists reject this restriction, these earlier treatments have in effect successfully replied to Kim but failed to provide reason to reject *Exclusion* in general.

4.2 An Interventionist Argument for Epiphenomenalism (I)

The argument of the preceding subsection showed that if *Subvenience Suffiency* is true, then if *Mental Causation*$_{KI}$ is true *Exclusion*$_{K}$ is false. However, Michael Baumgartner (2009; 2010) has recently argued that interventionism and *Subvenience Sufficiency* in fact entail that *Mental Causation*$_{KI}$ is false. If so, then *Subvenience Suffiency* would not suffice for rejecting *Exclusion*$_{K}$ after all. In this subsection I argue that the correct definition of intervention blocks this argument[50].

The official definition of intervention provided by Woodward is as follows (*ibid*, p. 98). First we need the (type-level) notion of an *intervention variable*. I is an intervention variable for X with respect to Y *iff*:

> *(IV)*
>
> *(I_1)* I is a contributing cause of X.
>
> *(I_2)* Some values of I are weakly sufficient for the value of X.
>
> *(I_3)* Every path from I to Y goes through X.
>
> *(I_4)* I is statistically independent of every contributing cause of Y on a non-X path.

Then we define the (token-level) notion of an *intervention*:

> *(IN)* $I = z_i$ is an intervention on X with respect to Y *iff* I is an intervention variable for X with respect to Y and $I = z_i$ is a weakly sufficient cause of the value taken by X.

[50] My argument in this subsection is indebted to correspondence with Michael Baumgartner.

Baumgartner's argument is simple. If *Mental Causation*$_{K1}$ is true then there exists a causal model $\mathscr{M}_{M1}$ containing a mental variable M_x that is a contributing cause of A_x. This in turn requires that there exist an intervention variable for M_x. If *Completeness*$_J$ is true then there exists a contributing cause P_x for A_x, along a path that does not go through M_x. *(I_4)* therefore requires that an intervention variable for M_x be statistically independent of P_x. But if *Subvenience Sufficiency* is true then the values of M_x are metaphysically necessitated by the values of P_x and therefore not statistically independent of them. Therefore there is no such intervention variable. Therefore there is no such model $\mathscr{M}_{M1}$.

I claim that the correct conclusion to draw from this argument is not that *Subvenience Suffiency* entails epiphenomenalism, but that the official definition of intervention should be amended[51]. As Shapiro and Sober (2007) and Woodward (2008*a*) (see also Shapiro 2010 and Woodward 2011) have shown, the justification for controlling for confounding causes that motivates the definition *(IV)* does not carry across to variables that are related by metaphysical necessitation. Consider a simple, and much discussed, example from Yablo (1992, p. 257). Sophie the pigeon pecks at all and only red things. So Sophie pecks at scarlet things. Consider two causal models. In the first model $\mathscr{M}_{S1}$, we have a variable with values for redness and all other colours at the same grain (R), and a variable with values for pecking and not pecking (K). In the second model $\mathscr{M}_{S2}$, we have a variable with values for scarlet and all other colour shades at the same grain (S), and K. Baumgartner's argument would have us reason as follows. Since S is a contributing cause of K in $\mathscr{M}_{S2}$, an intervention variable for R should be statistically independent of S. But the values of R are metaphysically necessitated by the values of S and therefore not statistically independent of them. Therefore there is no such intervention variable. Therefore there is no such model $\mathscr{M}_{S1}$. This is clearly the wrong result. When we wish to test for the causal relevance of redness, it should not be required that we hold fixed conditions which metaphysically suffice for redness. More generally, a test for the causal relevance of some property should not be required to hold fixed conditions that are metaphysically sufficient for the property to be instantiated. We require an amendment to *(IV)* that does not license this sort of reasoning.

I propose to amend *(IV)* as follows. I is an intervention variable for X with respect to Y in model $\mathscr{M}$ *iff*:

(IV)*

[51] I say amended rather than clarified because Woodward (2008*b*, §4) claims that his definition was not intended to involve the reference to models that my proposed definition will employ. For a different amendment, which I believe is entailed by my proposal, see Woodward (2011).

(I_1)* I is a contributing cause of X in model $\mathscr{M}_i$ constructed by adding I to $\mathscr{M}$.

(I_2)* Some values of I are weakly sufficient for the value of X in $\mathscr{M}_i$.

(I_3)* Every path from I to Y goes through X in every model containing I, X and Y.

(I_4)* I is statistically independent of every contributing cause Z of Y except those that are on an X path in every model containing I, X,Y, and Z.

And:

(IN)* $I = z_i$ is an intervention on X with respect to Y in model $\mathscr{M}$ *iff* I is an intervention variable for X with respect to Y in $\mathscr{M}$ and $I = z_i$ is a weakly sufficient cause of the value taken by X in model $\mathscr{M}_i$ constructed by adding I to $\mathscr{M}$.

These are model-relative definitions, which we de-relativise in the familiar way: I is an intervention variable for X with respect to Y *simpliciter iff* there is a model in which it is so represented; and $I = z_i$ is an intervention on X with respect to Y *simpliciter iff* there is a model in which it is so represented.

Note that this blocks the mistaken line of reasoning in our simple example. Suppose our intervention variable I for R represents ways of presenting Sophie with differently coloured objects. Condition *(I_3*)* requires that every path from I to K go through R for every model containing I, R and K. By *Non-Supervenience* there is no model containing I, R, S and K together, since the values of R supervene on the values of K. So the existence of model $\mathscr{M}_{S_2}$ does not prevent the satisfaction of *(I_3*)*. And since *(I_4*)* only requires statistical independence from contributing causes in models containing I, R and K, and S is not a contributing cause in any such model, the failure of statistical independence between R and S is compatible with the satisfaction of *(I_4*)*.

For parallel reasons, Baumgartner's interventionist argument for epiphenomenalism fails when formulated with *(IV*)*. In what follows, I will take interventionism to include definition *(IV*)*. If *Subvenience Sufficiency* is true, then interventionism provides no *a priori* reason to reject *Mental Causation*$_{KI}$.

4.3 An Interventionist Argument for Epiphenomenalism (II)

Since most non-reductionists reject *Subvenience Sufficiency*, we need to explore the consequences of doing so. In this subsection I present an argument for the conclusion

that if *Subvenience Sufficiency* is false then *Mental Causation*$_{kI}$ is false. The remainder of the section will be dedicated to developing the correct response to this argument.

Here is the argument. If *Subvenience Sufficiency* is false, then there exists a model in which the strongly sufficient actual causes of A_x identified by some model $\mathscr{M}_{Pi}$ do not metaphysically necessitate the actual causes of A_x identified by $\mathscr{M}_{MI}$. So let us suppose first that P_x is a strongly sufficient cause of A_x in $\mathscr{M}_{Pi}$, and second that P_x alone is not metaphysically sufficient for M_x, say because M_x metaphysically supervenes on P_x and P_y together. Now *Non-Supervenience* entails that there is no model that contains M_x, P_x and P_y. But it does not entail that there is no model $\mathscr{M}_{PxM}$ that contains M_x and P_x but not P_y. Since P_y is omitted from the model, we can hold P_x fixed and intervene on M_x by intervening on P_y. And in the other direction, we can hold M_x fixed and intervene on P_x just in case the possible values of P_x include other subsets of subvenience bases of M_x[52]. Moreover, given that $\mathscr{M}_{Pi}$ is effectively closed with respect to the M-variables and that P_x is strongly sufficient for A_x, it follows that there exists no intervention on M_x that would make a difference to A_x, with all other variables in $\mathscr{M}_{PxM}$ held fixed at any of their possible values. So M_x is not a direct cause of A_x in $\mathscr{M}_{PxM}$. Moreover on the assumption that P_x is not on any path from M_x to A_x[53], then M_x is not a contributing cause of A_x and hence does not provide an actual cause of A_x for any possible state of $\mathscr{M}_{PxM}$. Therefore M_x is not a contributing cause of A_x *simpliciter*.

This line of reasoning appears to show that if interventionism is true and *Subvenience Sufficiency* is false, then *Mental Causation*$_{kI}$ is false. This in turn entails that *Exclusion*$_k$ is true. If *Subvenience Sufficiency* is false for the relationship between physical properties and mental properties, then non-reductionism, on the interventionist framework, appears to entail *epiphenomenalism*. In what follows, I will call this the interventionist argument for epiphenomenalism. The central problem with the argument is simple to state, though not easy to identify. Before identifying the problem, however, I will consider an alternative response suggested by Bennett (2003).

[52] On the way most non-reductive physicalists think about typical cases of multiple realisation, many or perhaps most of these interventions will make no difference to A_x. However, all that is required for this test to rule that P_x is a cause of A_x is that at least one realisation of M_x makes a difference to A_x. This possibility has been orthodoxy since Fodor (1974).

[53] This is another extremely weak assumption. If it is not granted, the same argument would arise again for whatever physical variable provides a strongly sufficient cause of P_x, and so on, until we reach a variable representing a cause occurring before M_x that therefore could not possibly be on any path from M_x (*cf.* Bennett 2003, p. 494 fn. 18).

4.4 Isolation Failure

Bennett (2003, §8) argues that the case for compatibilism can be rescued even if *Subvenience Sufficiency* is false[54]. One way to read her argument is as follows. Employing her necessary condition for overdetermination (NC) (see §3.1), the compatibilist has argued so far that when *Subvenience Sufficiency* is true (O_2) is vacuously true, since it is metaphysically impossible for the actual value of P_x to obtain without the actual value of M_x obtaining. Having noted that *Subvenience Sufficiency* is false by the lights of most non-reductionists, she then argues that when this is the case (O_2) will instead be *false*. It will be false, she says, because the closest relevant possible world in which M_x takes a different value while P_x takes the same value is one in which the background conditions in which P_x is causally sufficient for A_x do not obtain, and therefore a world in which P_x is not sufficient for A_x. Therefore A_x might not occur, and (O_2) is false[55]. As I will explain in this subsection, I believe that Bennett has identified an important additional condition in which *Exclusion* is false. However, I will also argue that she has not done enough to show that the non-reductionist is entitled to believe that the condition obtains.

In order to evaluate Bennett's argument from our perspective, we need to translate it into the interventionist framework. The crucial point here is that according to interventionism, the counterfactuals that must be true for causal claims to be true do not concern claims about which variable settings would have been most likely to obtain were they different, but rather concern the results of hypothetical interventions. This is illustrated nicely by the Billy and Suzy case discussed in §3.1, where it did not undermine Billy's causal relevance to Victim that if Suzy had not fired Billy would have fired inaccurately, since the crucial question concerns the result of an intervention that has Billy fire accurately without Suzy firing[56]. To see how this point applies to Bennett's argument, recall that an intervention requires a change to the value of a variable in a model such 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 (definition *(IV*)* expresses this formally). If Bennett is right, then the nearest possible world in which there is a change of M_x without a change of P_x is a

[54] My argument in this subsection and the next is heavily indebted to correspondence with Michael Rescorla. See Rescorla (2011) for a similar argument in the context of the causal relevance of content to computation.

[55] The same argument is made, for the case of conceptual role semantics, by Block (1990*a*, §4).

[56] See fn. 20. By making this point I do not deny that there might be a notion of similarity definable on which the interventionist theory can be formulated in terms of that notion. For a comparison of interventionism with the Lewis (1973) theory of causation, in which similarity plays a crucial role, see Woodward (2003, §3.6).

world in which there is a change of background conditions Z_i to values for which P_x is not weakly sufficient for A_x, and in which there is a change of A_x. But this is just to say that this world involves a change to M_x that has an effect on A_x along a path from Z_i to A_x that does not include M_x. If so, then the nearest possible world in which there is a change of M_x without a change of P_x is not a world in which M_x has been changed by an intervention. Now if this is all there is to Bennett's point, then we have no reason to reject the interventionist argument for epiphenomenalism, since that argument did not claim that worlds in which M_x is changed by interventions with P_x held fixed are *closer* than all other worlds in which M_x changes without P_x changing. Rather, it merely claimed that *there are* worlds in which M_x is changed by an intervention with P_x held fixed, and that these are worlds in which A_x is not changed.

But there is another way to read Bennett's argument, on which she is suggesting that there might not *be* any such worlds—that it might be impossible to intervene on M_x with P_x held fixed, even if M_x isn't metaphysically necessitated by P_x alone[57]. And indeed, according to interventionism this would be the case if every possible way of changing P_y would have an effect on A_x along a path that does not include M_x, for then the causal requirements on interventions could not possibly be satisfied. So *Non-Supervenience* is not the only way in which *Independent Manipulability* can fail, and *Subvenience Sufficiency* not the only condition in which *Exclusion* is false. We also have the following condition:

> *Isolation Failure* For every variable M_x which actually causes some event, every set of P-variables providing strongly sufficient causes for that event contains some variable P_z such that every metaphysically possible change to M_x has an independent effect on P_z or *vice versa*, with all other variables held fixed at some combination of possible values.

We can then re-trace our earlier line of reasoning. If *Isolation Failure* is true, then by *Independent Manipulability* there will be no model which contains both a set of physical variables which provide a strongly sufficient cause for A_x, and a mental variable M_x which provides a cause for A_x. Since there is no model containing these variables together, they do not overdetermine A_x. So interventionism and *Isolation Failure* entail *Mental Causation*$_{K2}$, which with *Mental Causation*$_{K1}$ entails that *Exclusion*$_K$ is false. These conclusions are summarised in Figure 3.

[57] Bennett certainly has both arguments in mind, but does not commit wholly to either, for example writing that her claim is that "if the mental cause had not happened, that just *constitutively involves* various changes in the world that change, or at least may well change, what p causes" (p. 488, her emphasis).

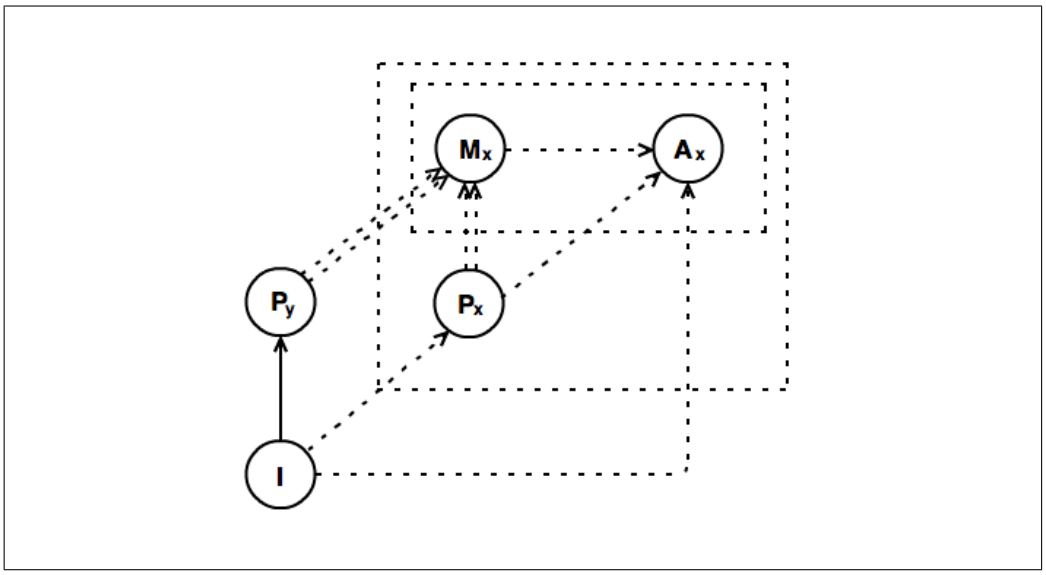

Figure 3: *Possible Directed Graphs for Nonreductionism*
Dotted arrows represent possible paths, double dotted arrows represent possible relations of metaphysical necessitation, dotted rectangles contain variables possibly appearing together in a model. When P_x is strongly sufficient for A_x, I have argued for the following results. If $P_x \rightrightarrows M_x$ then *Subvenience Sufficiency* is true, only the model represented by the smaller rectangle exists, and possibly $M_x \rightarrow A_x$. If $\{P_x, P_y\} \rightrightarrows M_x$ and either necessarily $I \rightarrow P_x$ or necessarily $I \rightarrow A_x$ (possibly *via* P_y, a possibility not explicitly represented here), then *Subvenience Sufficiency* is false, *Isolation Failure* is true, only the model represented by the smaller rectangle exists, and possibly $M_x \rightarrow A_x$. The interventionist argument for epiphenomenalism suggests that if $\{P_x, P_y\} \rightrightarrows M_x$ and $I \nrightarrow P_x$ and $I \nrightarrow A_x$, then *Subvenience Sufficiency* and *Isolation Failure* are false, models represented by both rectangles exist, and $M_x \nrightarrow A_x$.

Here is an example of a candidate model not ruled out by *Non-Supervenience* that is nevertheless ruled out by *Independent Manipulability*. Suppose that there are two hoses supplying water to a tank, and consider two models relating the flow of water through the hoses to the volume of water in the tank. In the first model $\mathscr{M}_{H1}$, we have variables for the throughput of hose one (H_1), the throughput of hose two (H_2), and the tank volume (T). In the second model $\mathscr{M}_{H2}$, we have variables for the throughput of hose one and hose two combined (H), and T. Suppose our intervention variables are I_1 for possible settings of a tap upstream that changes hose one throughput, I_2 for possible settings of a tap upstream that changes hose two throughput, and I for possible combinations of settings of these taps that result in different overall throughput. H_1 and H_2 are contributing causes of T in $\mathscr{M}_{H1}$, while H is a contributing cause of T in $\mathscr{M}_{H2}$. So all three are contributing causes *simpliciter*. By *Non-Supervenience* there is no model containing H_1, H_2 and H together, since the values of H supervene on the values of H_1 and H_2. However *Non-Supervenience* does not rule out a model $\mathscr{M}_{H3}$ containing H and H_1 without H_2, or a model $\mathscr{M}_{H4}$ containing H and H_2 without H_1. I_2 satisfies the requirements for an intervention variable for H in $\mathscr{M}_{H3}$ and I_1 satisfies the requirements for an intervention variable for H in $\mathscr{M}_{H4}$. And it may seem that we can conjure up intervention variables for H_1 in $\mathscr{M}_{H3}$ and H_2 in $\mathscr{M}_{H4}$ as follows: I_3 reflects ways of changing H_1 that involve simultaneous compensating changes to H_2, and I_4 reflects ways of changing H_2 that involve simultaneous compensating changes to H_1. But neither of these intervention variables satisfy *(I_3*)*, since for each there is a model in which there is an independent path from the intervention variable to T. For example, in a model containing I_3, H_1, H_2 and T there will be a path from I_3 to H_2 to T. Moreover, because of the mathematical relationship between H, H_1 and H_2 it is metaphysically necessary that a change to H_1, holding H fixed, will have an effect on H_2. So there is no possible intervention variable for H_1 in $\mathscr{M}_{H3}$, and therefore no such model $\mathscr{M}_{H3}$. A reason to believe *Isolation Failure* would ground an argument of the same form with respect to model $\mathscr{M}_{PxM}$ invoked in the interventionist argument for epiphenomenalism.

So, who is entitled to believe *Isolation Failure*? Not the non-reductionist in general, as I will now argue. Consider first the following argument offered by Block (1990*a*, §4) and Bennett (2003, p. 488) on behalf of the role functionalist. If role functionalism is true, then it is false that any mental properties supervene on P_x alone, for mental properties also supervene on the causal role filled by P_x. This means that any change to M_x holding P_x fixed would have to change the causal role filled by P_x. But necessarily, if the causal role of P_x were changed, P_x would no longer cause A_x and therefore A_x would change. So there is no change to M_x, holding P_x fixed, that does not have an independent effect on A_x. So *Isolation Failure* is true.

The problem is with the premise that necessarily, if the causal role of P_x were changed, P_x would no longer cause A_x[58]. This requires the claim that the *only* way to change the causal role of P_x in a way that makes a difference to M_x is by changing whether P_x causes A_x, and this in turn requires the highly implausible claim that there exists no alternative value of M_x that supervenes on a causal role that can be played by P_x and that includes having effect A_x, *but differs in myriad other ways*. All realistic forms of role functionalism entail that there are many causal relations that make up typical causal roles, and that distinct mental properties will typically be identified with distinct causal roles that nevertheless overlap with respect to particular causal relations. Given this feature of role functionalism, all we need to suppose in order to reject this argument for *Isolation Failure* is that there exists some way of changing enough *other* of the causal relations into which P_x enters, without independently changing A_x, so as to change the causal role definitive of M_x. I see no reason to deny that this is in general possible[59].

The faulty premise is stronger than is needed, however. It isn't required for the truth of *Isolation Failure* that every change to M_x, holding P_x fixed, makes for a change to A_x. Rather, it is only required that every such change be made through a variable that lies on an independent path to A_x. Recall the hose example. The problem was not that it was impossible to change H_1 without changing T. This was perfectly possible, by changes that involved altering H_2. The problem was that changes of this sort worked by cancellation along two paths to T, meaning they did not satisfy the requirements for interventions. Here then is what the role functionalist must believe, in order to believe *Isolation Failure*. Call the background variables in virtue of which P_x has the causal role it has Z_i. *Isolation Failure* is true *iff* every subset of Z_i, for which some possible change of state would be sufficient to change the value of M_x, contains a variable that is a contributing cause of A_x. That is, the role functionalist must believe that every possible way of changing the causal role definitive of a given mental state involves some change to another state that is itself a cause of the same effects. While properly evaluating this claim raises issues far too extensive to deal with in this paper, there are two points worth briefly noting. First, on the face of it only a very behaviouristically oriented analytic functionalist will be in a position to reject *Exclusion* in this way. Second, it is just this sort of functionalist that is particularly susceptible to the problem of metaphysically necessary effects (Rupert

[58] This problem with the argument is also pointed out by Rupert (2006, p. 267).

[59] Don't say it is impossible because the changed causal role would impact other mental states, and it is impossible that these wouldn't in turn make some difference to A_x. The one to many relationship between behaviours and the mental states that can produce them is one of the many basic problems with logical behaviourism.

2006). Better to find a general argument against *Exclusion* that does not turn on whether such a position can be defended.

The only other argument for *Isolation Failure* I have seen is also suggested by Bennett (2003, p. 496 fn. 29), this time on behalf of the content externalist. If content externalism is true, then it is false that relevant mental properties supervene on P_x alone, for those mental properties also supervene on external conditions P_y. Suppose our mental property is *desiring-water*. According to content externalism, the actual value of P_x is equally metaphysically compossible with the alternative mental property *desiring-twater*. But as the interventionist argument for epiphenomenalism urges, switching between these properties would make no difference to behaviour. Bennett replies that, on her model of overdetermination, it is a mistake to think that the desiring-twater world is the closest world in which the mental property is different and the physical property the same. So it is a mistake to think that the existence of the desiring-twater world shows that her (O_2) is non-vacuously true. This is because when we are evaluating counterfactuals in causal contexts, we must employ deletion readings (on which if an event had not occurred, no relevantly similar event would have occurred) rather than replacement readings (on which if an event had not occurred, some relevantly similar event might have occurred). Bennett suggests that with respect to desiring-water worlds, desiring-twater worlds are replacement worlds. Moreover once we rule out replacement worlds, Bennett suggests, the closest world in which M_x varies is likely to be one in which A_x also varies. Her only argument for this last claim is the sheer possibility of worlds where for example the causally sufficient physical property is instantiated "in a Petri dish"—that is, in worlds where there is no thinker left whatsoever, and therefore where it is obvious that A_x will be different.

There are problems with this argument even if we stay within Bennett's framework. For we have been given no reason to think that in general the only way in which content properties could vary holding the relevant internal state fixed involves possibilities ruled out by replacement readings. Perhaps desiring water is relevantly similar to desiring twater, though we certainly need to hear more about the principles governing this verdict. However, many externalist arguments have a form that allows us to vary mental content in ways that obviously allow non-replacement possibilities consistent with no behavioural differences. For instance, on some forms of externalism there are worlds where there is no content had by the state in question at all, consistently with the total internal state of the thinker remaining fixed[60]. More

[60]Consider for instance a world in which there is no single natural kind underlying the watery appearances. See Pryor (2007) for a discussion of these possibilities.

generally, Yablo (1997, §11) argues that two people could be wholly intrinsically identical but intentionally vary in pretty much any way we care to imagine, depending on their external context[61]. Moreover, keeping in mind that the relevant counterfactuals only involve holding a *single* internal state fixed, once we allow possibilities involving other internal states shifting their content it is hard to see how mental content could not be made to vary in close to arbitrary ways, consistent with that single state and the associated behaviour being preserved. If we want to defend *Isolation Failure* in this way then, we need to find a way either to dramatically restrict the scope of these externalist arguments, or to show that all of the relevant possibilities are more distant than Petri dish worlds. I do not regard this as a promising strategy for the compatibilist.

If we switch back to the interventionist framework, things initially look even worse. For then, as I argued above, considerations of similarity become irrelevant and it looks like we are charged with denying the possibilities invoked by the very thought experiments that motivated externalism in the first place. However, for reasons elaborated above, we are also constrained regarding the internal states we are permitted to change, since we must not alter any that are themselves causes of the effect in question. Perhaps some mixed strategy, denying certain externalist possibilities on the one hand and appealing to strong functionalist constraints on possible contents on the other, could be made to work. But the work remains to be done[62,63].

I conclude that the non-reductionist has no general reason to believe *Isolation Failure*. The possible theories of mental properties that entail *Isolation Failure* are either unattractive or await development. We have already seen that the non-reductionist has no reason to believe *Subvenience Sufficiency*. So far we have identified some elegant possibilities for denying *Exclusion*, neither of which most who believe *Nonreductionism* have any reason to endorse.

[61] See also Fisher (2007).

[62] See Rescorla (2011) for some interesting suggestions along these lines.

[63] There are complexities here that I have glossed over, for strictly what is established by externalist arguments is the possibility of internal physical duplicates with different mental properties. This does not entail the claim that a single thinker could change mental properties without changing any internal physical properties. Moreover, externalists typically do not argue in ways that commit themselves to this latter possibility, for they typically suppose that there must be some causal interaction between the environment and the thinker in order for content properties to change (see for example Block 1990*b*). In this case, the thinker goes through a process of interaction with the environment, after which they return to the same internal state as before the process. I am supposing that processes of this type qualify as interventions in the relevant sense—this might be contested, but I would be surprised if the defence of compatibilism turned out to rest on these details.

4.5 A General Argument against Exclusion

It is time to identify the real problem with the interventionist argument for epiphenomenalism. That argument made plausible that if *Subvenience Sufficiency* is false and M_x supervenes on P_x and P_y together, there exists a model $\mathscr{M}_{P_xM}$ containing P_x, M_x and A_x in which M_x is not a contributing cause of A_x. It then inferred that M_x is not a contributing cause of A_x *simpliciter*. I grant the existence of $\mathscr{M}_{P_xM}$, but reject the inference.

Recall the de-relativised definition of contributing cause from §3.1: X is a *contributing cause* of Y *simpliciter iff* there exists a model in which X is a contributing cause of Y. This entails an asymmetry between the way in which a claim that a variable is a contributing cause *simpliciter* can be justified, and the way in which a claim that variable is not a contributing cause *simpliciter* can be justified. To show the former, we merely need to identify a single model in which it is so represented. To show the latter, we need to argue there does not exist *any* model in which it is so represented. So the inference from the existence of $\mathscr{M}_{P_xM}$, in which M_x is not a contributing cause of A_x, to the conclusion that M_x is not a contributing of A_x *simpliciter*, is invalid.

Moreover, seeing this flaw in the argument puts us into position to appreciate how it could be that there are causal models in which mental variables are causes when *Subvenience Sufficiency* is false. The interventionist argument for epiphenomenalism involved a model $\mathscr{M}_{P_xM}$ in which P_x was included and P_y was excluded. This allowed M_x to be changed holding P_x fixed, since M_x could be changed by varying P_y. In this model M_x is not a contributing cause of A_x. But consider the reverse case, a model $\mathscr{M}_{P_yM}$ in which P_y is included and P_x excluded. This allows M_x to be changed holding P_y fixed, since M_x can be changed by varying P_x. There is every reason to believe that in models of this sort changes to M_x will result in changes to A_x, and certainly no general reason for believing that they will not. The basic claim enabled by the interventionist framework, then, is extremely intuitive: it is both true that holding our internal physical state fixed and varying our mental states would not change our behaviour, and true that holding external physical states fixed and varying our mental states would change our behaviour—and the truth of the latter is sufficient to establish the causal relevance of our mental states.

Having diagnosed the mistake in the interventionist argument for epiphenomenalism, we are now in a position to provide a general argument against *Exclusion*, given interventionism and *Nonreductionism*. Either *Subvenience Sufficiency* is true or it is false. If it is true, then as the argument in §4.1 showed, *Exclusion* is false since the mental variables and causally sufficient physical variables cannot appear in the same

model. If it is false, then as the argument in §4.3 showed, *Exclusion* is false since every model in which the mental and causally sufficient physical variables appear together is one in which the mental variables are not represented as causes. This does not entail that they are *not* causes, for this merely requires that there exists some causal model in which they are so represented. So interventionism and *Nonreductionism* entail that *Exclusion* is false, whether or not *Subvenience Sufficiency* is true or false.

5 Conclusion

I have argued that interventionism and *Nonreductionism* entail that *Exclusion* is false. If interventionism is true than mental properties can cause physical events without overdetermining them, even if there is a sufficient physical cause for every physical event. I will close with a comment on the relationship between the relativised and de-relativised definitions of contributing causation.

Woodward is well aware that the relationship between the relativised definition of contributing causation and the de-relativised definition of contributing causation has the consequence that two models can differ on whether a variable is a contributing cause of another variable. For example, as Woodward (2003, p. 56) observes, when there are two paths between variables X and Y that cancel each other out, a model that does not include any variables along these paths will be one in which X is not a contributing cause of Y, while a model that does include at least one variable along these paths may be one in which X is a contributing cause of Y. So the fact that $\mathscr{M}_{P_x M}$ fails to represent M_x as a contributing cause of A_x is not a peculiarity of this case, but a possibility that also arises in more mundane cases. Woodward (2008*b*, §7) however suggests that the relativised definition of contributing causation is nevertheless *conservative* in the following sense. Call a model that has the variables in $\mathscr{M}$ as a subset an *expansion* of $\mathscr{M}$. For a given causal concept, if for every model $\mathscr{M}$ in which X satisfies the concept with respect to Y there exists no expansion of $\mathscr{M}$ in which X does not satisfy the concept with respect to Y, then the causal concept is conservative. On the face of it, $\mathscr{M}_{M_I}$ and $\mathscr{M}_{P_x M}$ provide a counterexample to the claim that contributing causation is conservative. For $\mathscr{M}_{P_x M}$ and $\mathscr{M}_{P_y M}$ are both expansions of $\mathscr{M}_{M_I}$, and yet M_x is a contributing cause of A_x in $\mathscr{M}_{M_I}$ but not in $\mathscr{M}_{P_x M}$. I suggest that the claim that a model is conservative implicitly makes reference either to background conditions or intervention variables. On the first option we take $\mathscr{M}_{M_I}$ to really be equivalent to $\mathscr{M}_{P_y M}$, with background conditions P_y left implicit. We don't typically represent these background conditions since we never in fact manipulate the mental states of others by engaging in externalist switching or functionalist

rewiring. On the second option we take conservatism to apply not to expansions of models *simpliciter*, but to models and their associated intervention variables. Again, the way in which we intervene on M_x in $\mathscr{M}_{M_I}$ will reflect the intervention variables of $\mathscr{M}_{P_y M}$ but not $\mathscr{M}_{P_x M}$. Either way, $\mathscr{M}_{P_y M}$ will count as an expansion of $\mathscr{M}_{M_I}$, $\mathscr{M}_{P_x M}$ will not, and a form of conservatism is retained.

References

Baker, Lynne Rudder. 1993. "Metaphysics and Mental Causation", in *Mental Causation*, edited by John Heil and Alfred Mele, Oxford University Press, Oxford, pp. 75–96. URL: http://www.people.umass.edu/lrb/files/bak93metS.pdf.

Baumgartner, Michael. 2009. "Interventionist Causal Exclusion and Non-reductive Physicalism", in *International Studies in the Philosophy of Science*, Vol. 23, No. 2, July 2009, pp. 161–178. URL: http://dx.doi.org/10.1080/02698590903006909.

————. 2010. "Interventionism and Epiphenomenalism", in *Canadian Journal of Philosophy*, Vol. 40, No. 3, September 2010, pp. 359–383.

Bedau, Mark A. and Paul Humphreys. 2008. *Emergence: Contemporary Readings in Philosophy And Science,* edited by Mark A. Bedau and Paul Humphreys. MIT Press, Cambridge MA.

Bennett, Karen. 2003. "Why the Exclusion Problem Seems Intractable, and How, Just Maybe, To Tract It", in *Noûs*, Vol. 37, No. 3, September 2003, pp. 471–497. URL: http://dx.doi.org/10.1111/1468-0068.00447.

Björnsson, Gunnar. 2007. "How Effects Depend on their Causes, Why Causal Transitivity Fails, and Why we Care About Causation", in *Philosophical Studies*, Vol. 133, No. 3, April 2007, pp. 349–390. URL: http://dx.doi.org/10.1007/s11098-005-4539-8.

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, Cambridge University Press, Cambridge, pp. 196–225. Reprinted in Blackburn (1993, pp. 229–254).

————. 1993. *Essays in Quasi-Realism,* Oxford University Press, Oxford.

Block, Ned. 1980. *Readings in Philosophy of Psychology,* edited by Ned Block. Vol. 1. Harvard University Press, Cambridge MA.

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

————. 1990*b*. "Inverted Earth", in *Philosophical Perspectives*, Vol. 4: Action Theory and Philosophy of Mind, January 1990, pp. 53–79. Reprinted in Block (2007, pp. 511–532). URL: http://dx.doi.org/10.2307/2214187.

————. 2003. "Do Causal Powers Drain Away?", in *Philosophy and Phenomenological Research*, Vol. 67, No. 1, July 2003, pp. 133–150. URL: http://dx.doi.org/10.1111/j.1933-1592.2003.tb00029.x.

————. 2007. *Consciousness, Function, and Representation,* Vol. 1. Collected Papers. MIT Press, Cambridge MA.

Bontly, Thomas D. 2002. "The Supervenience Argument Generalizes", in *Philosophical Studies*, Vol. 109, No. 1, May 2002, pp. 75–96. URL: http://dx.doi.org/10.1023/A:1015786809364.

Burge, Tyler. 1993. "Mind-Body Causation and Explanatory Practice", in *Mental Causation*, edited by John Heil and Alfred Mele, Oxford University Press, Oxford, pp. 97–120. Reprinted with postscript in Burge (2007*a*, pp. 344–362).

————. 2007*a*. *Foundations of Mind,* Vol. 2. Philosophical Essays. Oxford University Press, Oxford.

————. 2007*b*. "Postscript: Mind-Body Causation and Explanatory Practice", in *Foundations of Mind*. Vol. 2, Philosophical Essays, Oxford University Press, Oxford, pp. 363–382.

Clark, Andy and Peter Millican. 1999. *Connectionism, Concepts, and Folk Psychology: The Legacy of Alan Turing,* edited by Andy Clark and Peter Millican. Vol. 2. Oxford University Press, Oxford.

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, Oxford University Press, Oxford, pp. 176–195.

Davidson, Donald. 1963. "Actions, Reasons, and Causes", in *The Journal of Philosophy*, Vol. 60, No. 23, November 1963, pp. 685–700. Reprinted in Davidson (2001, pp. 3-19). URL: http://dx.doi.org/10.2307/2023177.

————. 1967. "Causal Relations", in *The Journal of Philosophy*, Vol. 64, No. 21, November 1967, pp. 691–703. Reprinted in Davidson (2001, pp. 149-162).

————. 1970. "Mental Events", in *Experience and Theory*, edited by Lawrence Foster and Joe William Swanson, University of Massachusetts Press, Amherst MA, pp. 79–101. Reprinted in Block (1980, pp. 107–119) and Moser and

Trout (1995, pp. 11–126) Davidson (2001, pp. 207-227). URL: http://dx.doi.org/10.1093/0199246270.003.0011.

————. 1995. "Laws and Cause", in *Dialectica*, Vol. 49, No. 2-4, June 1995, pp. 263–279.

————. 2001. *Essays on Actions and Events,* 2nd edition. Oxford University Press, Oxford. URL: http://dx.doi.org/10.1093/0199246270.001.0001.

Feigl, Herbert and May Brodbeck. 1953. *Readings in the Philosophy of Science,* edited by Herbert Feigl and May Brodbeck. Appleton-Century-Crofts, New York.

Field, Hartry. 2003. "Causation in a Physical World", in *The Oxford Handbook of Metaphysics*, edited by Michael J. Loux and Dean W. Zimmerman, Oxford University Press, Oxford, pp. 435–460.

Fisher, Justin C. 2007. "Why Nothing Mental is Just in the Head", in *Noûs*, Vol. 41, No. 2, June 2007, pp. 318–334. URL: http://dx.doi.org/10.1111/j.1468-0068.2007.00649.x.

Fodor, Jerry A. 1974. "Special Sciences (Or: The Disunity of Science as a Working Hypothesis)", in *Synthese*, Vol. 28, No. 2, October 1974, pp. 97–115. Reprinted in Block (1980, pp. 120–133), Fodor (1983, pp. 127–145), Moser and Trout (1995, pp. 53–67) and Bedau and Humphreys (2008, pp. 395–410). URL: http://dx.doi.org/10.1007/BF00485230.

————. 1983. *RePresentations: Philosophical Essays on the Foundations of Cognitive Science,* MIT Press, Cambridge MA.

Gibbons, John. 2006. "Mental Causation without Downward Causation", in *The Philosophical Review*, Vol. 115, No. 1, January 2006, pp. 79–103. URL: http://dx.doi.org/10.1215/00318108-115-1-79.

Glymour, Clark et al. 2009. "Actual Causation: A Stone Soup Essay", in *Synthese*, Vol. 175, No. 2, March 2009, pp. 169–192. URL: http://dx.doi.org/10.1007/s11229-009-9497-9.

Godfrey-Smith, Peter. 2005. "Folk Psychology as a Model", in *Philosophers' Imprint*, Vol. 5, No. 6, November 2005, pp. 1–16. URL: http://hdl.handle.net/2027/spo.3521354.0005.006.

Goldman, Alvin I. 1969. "The Compatibility of Mechanism and Purpose", in *The Philosophical Review*, Vol. 78, No. 4, October 1969, pp. 468–482.

Hall, Ned. 2007. "Structural Equations and Causation", in *Philosophical Studies*, Vol. 132, No. 1, January 2007, pp. 109–136. URL: http://dx.doi.org/10.1007/s11098-006-9057-9.

Halpern, Joseph Y. and Judea Pearl. 2001. "Causes and Explanations: A Structural-Model Approach. Part I: Causes", in *Uncertainty in Artificial Intelligence: Proceedings of the Seventeenth Conference*, edited by Jack Breese and Daphne

Koller, Morgan Kaufmann, San Francisco, pp. 194–202. Revised version published as Halpern and Pearl (2005*a*). URL: http://arxiv.org/abs/cs/0011012v2.

———. 2005*a*. "Causes and Explanations: A Structural-Model Approach. Part I: Causes", in *British Journal for the Philosophy of Science*, Vol. 56, No. 4, December 2005, pp. 843–887. Revised version of Halpern and Pearl (2001). URL: http://dx.doi.org/10.1093/bjps/axi147.

———. 2005*b*. "Causes and Explanations: A Structural-Model Approach. Part II: Explanations", in *British Journal for the Philosophy of Science*, Vol. 56, No. 4, December 2005, pp. 889–911. URL: http://dx.doi.org/10.1093/bjps/axi148.

Heckman, James J. 2005. "The Scientific Model of Causality", in *Sociological Methodology*, Vol. 35, No. 1, August 2005, pp. 1–98. URL: http://dx.doi.org/10.1111/j.0081-1750.2006.00163.x.

Heil, John and Alfred Mele. 1993. *Mental Causation,* edited by John Heil and Alfred Mele. Oxford University Press, Oxford.

Hiddleston, Eric. 2005. "Causal Powers", in *The British Journal for the Philosophy of Science*, Vol. 56, No. 1, March 2005, pp. 27 –59. URL: http://dx.doi.org/10.1093/phisci/axi102.

Hitchcock, Christopher. 2001. "The Intransitivity of Causation Revealed in Equations and Graphs", in *The Journal of Philosophy*, Vol. 98, No. 6, June 2001, pp. 273–299.

———. 2004. "Routes, Processes, and Chance-Lowering Causes", in *Cause and Chance: Causation in an Indeterministic World*, edited by Phil Dowe and Paul Noordhof, Routledge, London, pp. 138–152.

———. 2007. "Prevention, Preemption, and the Principle of Sufficient Reason", in *Philosophical Review*, Vol. 116, No. 4, October 2007, pp. 495–532. URL: http://dx.doi.org/10.1215/00318108-2007-012.

———. 2009. "Structural Equations and Causation: Six Counterexamples", in *Philosophical Studies*, Vol. 144, No. 3, June 2009, pp. 391–401. URL: http://dx.doi.org/10.1007/s11098-008-9216-2.

———. 2011. "Trumping and Contrastive Causation", in *Synthese*, Vol. 181, No. 2, July 2011, pp. 227–240. URL: http://dx.doi.org/10.1007/s11229-010-9799-y.

Hitchcock, Christopher and Joseph Y. Halpern. 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, College Publications, London, pp. 383–406.

Horgan, Terence. 1997. "Kim on Mental Causation and Causal Exclusion", in *Noûs*, Vol. 31, Supplement: Philosophical Perspectives, 11, Mind, Causation, and World 1997, pp. 165–184.

Jackson, Frank and Philip Pettit. 1988. "Functionalism and Broad Content", in *Mind*, Vol. 97, No. 387, July 1988, pp. 381–400. Reprinted in Pessin and Goldberg (1996, pp. 219–237) and Jackson, Pettit, and Smith (2004, pp. 95–118). URL: http://dx.doi.org/10.1093/mind/XCVII.387.381.

———. 1990*a*. "Causation in the Philosophy of Mind", in *Philosophy and Phenomenological Research*, Vol. 50, Supplement, Autumn 1990, pp. 195–214. Reprinted with postscript in Clark and Millican (1999, pp. 75–100) and Jackson, Pettit, and Smith (2004, pp. 45–68). URL: http://dx.doi.org/10.2307/2108039.

———. 1990*b*. "Program Explanation: A General Perspective", in *Analysis*, Vol. 50, No. 2, March 1990, pp. 107–117. Reprinted in Jackson, Pettit, and Smith (2004, pp. 119–130). URL: http://dx.doi.org/10.2307/3328853.

Jackson, Frank, Philip Pettit, and Michael Smith. 2004. *Mind, Morality, and Explanation: Selected Collaborations,* Oxford University Press, Oxford.

Kallestrup, Jesper and Jakob Hohwy. 2008. *Being Reduced: New Essays on Reduction, Explanation and Causation,* edited by Jesper Kallestrup and Jakob Hohwy. Oxford University Press, Oxford.

Kim, Jaegwon. 1973. "Causes and Counterfactuals", in *The Journal of Philosophy*, Vol. 70, No. 17, October 1973, pp. 570–572. Reprinted in Tooley (1999, pp. 190–193).

———. 1997. "Does the Problem of Mental Causation Generalize?", in *Proceedings of the Aristotelian Society*, Vol. 97, No. 3, pp. 281–297. URL: http://dx.doi.org/10.1111/1467-9264.00017.

———. 1998. *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation,* MIT Press, Cambridge MA. URL: https://cognet.mit.edu/library/books/view?isbn=0262611538.

———. 2003. "Blocking Causal Drainage and Other Maintenance Chores with Mental Causation", in *Philosophy and Phenomenological Research*, Vol. 67, No. 1, July 2003, pp. 151–176. URL: http://dx.doi.org/10.1111/j.1933-1592.2003.tb00030.x.

———. 2005. *Physicalism, Or Something Near Enough,* Princeton University Press, Princeton.

Lewis, David. 1973. "Causation", in *The Journal of Philosophy*, Vol. 70, No. 17, October 1973, pp. 556–567. Reprinted in Lewis (1986, pp. 159–171) and Tooley (1999, pp. 178–189). URL: http://dx.doi.org/10.2307/2025310.

Lewis, David. 1986. *Philosophical Papers,* Vol. II. Oxford University Press, Oxford. URL: http://dx.doi.org/10.1093/0195036468.001.0001.

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, Oxford University Press, Oxford, pp. 149–163. URL: http://rci.rutgers.edu/~loewer/papers/Why_There_is_Anything_Except_Physics.pdf.

————. 2009. "Why is There Anything Except Physics?", in *Synthese*, Vol. 170, No. 2, September 2009, pp. 217–233. URL: http://dx.doi.org/10.1007/s11229-009-9580-2.

McCain, Kevin. 2010. "Interventionism Defended". Unpublished manuscript. June 2010. URL: http://sites.google.com/site/kevinmccain/home/research/interventionism-defended.

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

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, Imprint Academic, Exeter, pp. 195–224. Reprinted in Monnoyer (2004).

————. 2004. "Causal Models, Token Causation, and Processes", in *Philosophy of Science*, Vol. 71, No. 5, December 2004, pp. 820–832. URL: http://dx.doi.org/10.1086/425057.

Monnoyer, Jean-Maurice. 2004. *La Structure du Monde: Objets, Propriétés, États de Choses,* edited by Jean-Maurice Monnoyer. Vrin, Paris.

Moser, Paul K. and J. D. Trout. 1995. *Contemporary Materialism: A Reader,* edited by Paul K. Moser and J. D. Trout. Routledge, London.

Noordhof, Paul. 1997. "Making the Change: the Functionalist's Way", in *British Journal for the Philosophy of Science*, Vol. 48, No. 2, pp. 233–250. URL: http://dx.doi.org/10.1093/bjps/48.2.233.

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

Pereboom, Derk and Hilary Kornblith. 1991. "The Metaphysics of Irreducibility", in *Philosophical Studies*, Vol. 63, No. 2, August 1991, pp. 125–145. URL: http://dx.doi.org/10.1007/BF00381684.

Pessin, Andrew and Sanford Goldberg. 1996. *The Twin Earth Chronicles: Twenty Years of Reflection on Hilary Putnam's "The Meaning of 'Meaning'",* edited by Andrew Pessin and Sanford Goldberg. M. E. Sharpe, Armonk NY.

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 University Press, Oxford.

Pryor, James. 2007. "What's Wrong With McKinsey-Style Reasoning?", in *Internalism and Externalism in Semantics and Epistemology*, edited by Sanford C. Goldberg, Oxford University Press, Oxford, pp. 177–200.

Rescorla, Michael. 2011. "The Causal Relevance of Content to Computation". Unpublished manuscript. URL: http://www.philosophy.ucsb.edu/people/profiles/faculty/cvs/papers/causal2.pdf.

Robb, David and John Heil. 2008. "Mental Causation", in *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Stanford University, Stanford. URL: http://plato.stanford.edu/entries/mental-causation/.

Ross, Don and David Spurrett. 2005. "What to Say to a Sceptical Metaphysician: A Defense Manual for Cognitive and Behavioral Scientists", in *Behavioral and Brain Sciences*, Vol. 27, No. 5, October 2005, pp. 603–627. URL: http://dx.doi.org/10.1017/S0140525X04000147.

Rupert, Robert D. 2006. "Functionalism, Mental Causation, and the Problem of Metaphysically Necessary Effects", in *Noûs*, Vol. 40, No. 2, June 2006, pp. 256–283. URL: http://dx.doi.org/10.1111/j.0029-4624.2006.00609.x.

Russell, Bertrand. 1912–1913. "On the Notion of Cause", in *Proceedings of the Aristotelian Society*, Vol. 13, pp. 1–26. Reprinted in Russell (1918, pp. 142–164) and Russell (2003, pp. 163–182). Simplified version published as Russell (1914).

————. 1914. "On the Notion of Cause, with Applications to the Free Will Problem", in *Our Knowledge of the External World: As a Field for Scientific Method in Philosophy*, Open Court, Chicago, pp. 214–246. Reprinted in Feigl and Brodbeck (1953, pp. 387–407). URL: http://www.hist-analytic.org/Russellcause.pdf.

————. 1918. *Mysticism and Logic,* Longmans Green, London.

————. 2003. *Russell on Metaphysics: Selections from the Writings of Bertrand Russell,* edited by Stephen Mumford. Routledge, London.

Shapiro, Lawrence A. 2010. "Lessons from Causal Exclusion", in *Philosophy and Phenomenological Research*, Vol. 81, No. 3, November 2010, pp. 594–604. URL: http://dx.doi.org/10.1111/j.1933-1592.2010.00382.x.

Shapiro, Lawrence A. and Elliott Sober. 2007. "Epiphenomenalism—The Do's and the Don'ts", in *Thinking about Causes: From Greek Philosophy to Modern Physics*, edited by Peter Machamer and Gereon Wolters, Pittsburgh-Konstanz Series in the Philosophy and History of Science, University of Pittsburgh Press, Pittsburgh, pp. 235–264. URL: http://philosophy.wisc.edu/sober/shapiro%20and%20sober%20epi%201%202%2006.pdf.

Strevens, Michael. 2007. "Review of Woodward, *Making Things Happen*", in *Philosophy and Phenomenological Research*, Vol. 74, No. 1, January 2007, pp. 233–249. URL: http://dx.doi.org/10.1111/j.1933-1592.2007.00012.x.

————. 2008. "Comments on Woodward, *Making Things Happen*", in *Philosophy and Phenomenological Research*, Vol. 77, No. 1, July 2008, pp. 171–192. URL: http://dx.doi.org/10.1111/j.1933-1592.2008.00180.x.

Thomson-Jones, Martin. 2007. "Overdetermination, Mental Causation, and the Staggered Firing Squad: Problems with the Case for Compatibilism". Unpublished manuscript.

Tooley, Michael. 1999. *Laws of Nature, Causation, and Supervenience,* edited by Michael Tooley. Vol. 1. Metaphysics. Garland, New York.

Van Gulick, Robert. 1992. "Three Bad Arguments for Intentional Property Epiphenomenalism", in *Erkenntnis*, Vol. 36, No. 3, May 1992, pp. 311–332. URL: http://dx.doi.org/10.1007/BF00204132.

Weslake, Brad. 2010. "Explanatory Depth", in *Philosophy of Science*, Vol. 77, No. 2, April 2010, pp. 273–294. URL: http://dx.doi.org/10.1086/651316.

————. ms. "A Partial Theory of Actual Causation".

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

————. 2008*a*. "Mental Causation and Neural Mechanisms", in *Being Reduced: New Essays on Reduction, Explanation and Causation*, edited by Jesper Kallestrup and Jakob Hohwy, Oxford University Press, Oxford, pp. 218–262.

————. 2008*b*. "Response to Strevens", in *Philosophy and Phenomenological Research*, Vol. 77, No. 1, August 2008, pp. 193–212. URL: http://dx.doi.org/10.1111/j.1933-1592.2008.00181.x.

————. 2011. "Interventionism and Causal Exclusion". Unpublished manuscript. June 2011. URL: http://philsci-archive.pitt.edu/8651/.

Yablo, Stephen. 1992. "Mental Causation", in *The Philosophical Review*, Vol. 101, No. 2, April 1992, pp. 245–280. Reprinted in Yablo (2009, pp. 222–248).

————. 1997. "Wide Causation", in *Noûs*, Vol. 31, No. Supplement: Philosophical Perspectives, 11, Mind, Causation, and World, pp. 251–281. Reprinted in Yablo (2009, pp. 275–306).

————. 2002. "De Facto Dependence", in *The Journal of Philosophy*, Vol. 99, No. 3, March 2002, pp. 130–148.

Yablo, Stephen. 2004. "Advertisement for a Sketch of an Outline of a Prototheory of Causation", in *Counterfactuals and Causation*, edited by John Collins, Ned Hall, and L. A. Paul, MIT Press, Cambridge MA, pp. 119–138.

————. 2009. *Thoughts: Papers on Mind, Meaning, and Modality,* Oxford University Press, Oxford. URL: http://dx.doi.org/10.1093/acprof:oso/9780199266463.001.0001.