To Appear in *Making A Difference* (eds.) Beebee, Hitchcock, and Price

Intervening in the Exclusion Argument

James Woodward
History and Philosophy of Science
University of Pittsburgh

1.

Peter Menzies' work on causation includes a number of important contributions to our understanding of mental causation and the causal exclusion argument. I share with Peter the conviction that an interventionist account of causation can cast new light on this complex of issues, but our views diverge in detail at several points. In this essay, I would like to briefly clarify my own views about mental causation and the exclusion argument, respond to some recent criticisms of those views, and then contrast those views with somewhat different approach favored by Peter.

2.

I assume that the exclusion argument is sufficiently well- known that a detailed summary is unnecessary. Following Jaegwon Kim's iconic diagram (Figure 1) , assume that M_1 and M_2 are mental properties or events and P_1 and P_2 their respective physical supervenience bases, with these supervenience relations represented by double-tailed arrows. These supervenience relations are understood in the way that is standard in non-reductive physicalism: P_1 is a realizer of M_1 but not identical with it and P_2 is a realizer of M_2 but is not identical with it. Assume also that P_1 causes P_2, as represented by the single-tailed arrow from P_1 to P_2. The questions of interest have to do with the causal role of M_1 . Given the above assumptions is it correct to regard to M_1 as causing M_2 or as causing P_2?

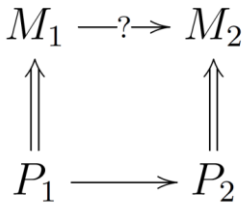

Figure 1

According to the exclusion argument, as usually formulated, the answer to both questions is "no". Focusing for simplicity just on the issue of whether M_1 causes M_2, one version of the argument runs roughly as follows: By the principle of the causal closure of the physical, P_2 must have a "sufficient" cause that is purely physical—assume without loss of generality that P_1 is such a cause. It follows that "all the causal work"

that is required for the occurrence of P_2 is accomplished by the occurrence of P_1. Moreover, once P_2 occurs, it guarantees the occurrence of M_2 in virtue of the supervenience relationship between P_2 and M_2. Thus P_1 is by itself sufficient for M_2 . It thus appears that there is nothing "left over" for M_1 to do in the way of causing M_2. Put slightly differently, the argument is that given the causal relationship between P_1 and P_2, it is unnecessary and superfluous to suppose that *in addition* there is a also a causal relationship between M_1 and M_2 ; indeed if we were to draw an arrow from M_1 to M_2 we would be introducing an implausible kind of "over-determination" into our picture of the relationships among P_1, P_2, M_1 and M_2. M_2 and P_2 would be over-determined by two sets of causes: P_1 and M_1. Yet another way of putting the argument is in terms of a claim about what it is appropriate to "control for" or "hold fixed" in assessing the causal impact of M1: since P_1 is an alternative cause of M_2, we must hold P_1 fixed in assessing whether M_1 causes M_2 When we do so we see that M_1 has no causal impact on M_2 under this condition. Similar arguments can be deployed to support the conclusion that M_1 does not cause P_2 .

We thus seem lead to the conclusion that M_1 is causally inert, with all of the real causal action taking place at the level of P_1 and P_2. Moreover, if the above arguments are correct, this conclusion seems to generalize to all of the special sciences: to the extent that non-reductive physicalism is the right account of the relationship between the states or properties that figure in sciences like biology, economics or psychology and the underlying physical realizations of these, then strictly speaking there are no causal relationships among these states and no true causal generalizations relating them.

3.

Like Peter, I think that this conclusion is mistaken and that an interventionist approach to causation can give us some insight into why it is mistaken. In this section I briefly reprise my version of the interventionist account and lay the groundwork for its application to the exclusion problem.

In the version of interventionism presented in my 2003, a number of different causal notions are distinguished and characterized. However, as far as the exclusion problem goes, it will be sufficient to make use of a simple, generic notion of causation. Assume that we have a set of variables **V** representing properties or states that stand in causal relationships. Variables are capable of taking a range of different "values" which in this context we may think of corresponding to more specific properties or their absence. For example, we might think of the variable *R* as possessing two possible values which correspond to the presence of absence of the color red. Initially I will assume that the variables with which we are concerned are "fully distinct" in the sense that they are *not* connected by dependency relations that are non-causal in nature. I include in this category relationships that hold for logical, conceptual, or mathematical relationships as well as supervenience relationships. I take the mark of the presence of non-causal dependency relations among a set of values **V** to be that the values of some of the variables impose constraints on the possible values that can be taken by some of the other variables, but not because of causal relations among those variables (Cf. condition **IF** below) . In other words, certain combinations of values for the variables in **V** are "impossible", but not because of causal relations among those variables. For

example, if we have a variable *X* one of whose values is "saying hello" and another variable *Y* one of whose values is "saying hello loudly" , then *Y*'s taking this value constrains the value taken by *X*, but in a non-causal way. Similarly if a variable *M* supervenes on a variable *P*, then certain combinations of values for *P* and *M* are ruled out as "metaphysically impossible. Later I will relax this assumption of non-distinctness, in order to deal with structures in which supervenience relations are present.

Given variables that satisfy the requirement of distinctness, we can characterize what it is for *X* to cause *Y* as follows:

(**M**) *X* causes *Y* if and only if there is a possible intervention on *X* such that if that intervention were to occur, the value of *Y* or the probability distribution of *Y* would change.

We can also characterize a closely related notion of a variable *X* taking a certain value $X=x$ causing a second variable *Y* to take a value $Y=y$ as follows:

(**M***) $X=x$ causes $Y=y$ if and only if there is a possible intervention on *X* that changes the value of *X* so that $X \neq x$ such that if that intervention were to occur, the value of *Y* would change to $Y \neq y$.

The intuitive notion of an intervention on *X* is that of an unconfounded experimental manipulation of *X* which is of such a character that if an association between *X* and *Y* remains after this manipulation, this shows that *X* causes *Y*. In effect, the intervention "controls for" other possible sources of association between *X* and *Y* besides whatever association is due to *X*'s causing *Y*. I will assume that an intervention *I* on *X* brings *X* completely under the control of *I* so that *I* "breaks" any previously existing causal connections directed into *X*, supplying *X* with an independent, exogenous causal history—a consequence that my be represented graphically as the severing of previously holding or endogenous edges directed into *X*[1]. Since the details will matter for my later discussion, I quote from my (2003):

> (**IV**) *I* is an intervention variable for *X* with respect to *Y* iff
> 1. *I* causes *X*;
> 2. *I* acts as a switch for all other variables that cause *X*. That is, certain values of *I* are such that when *I* attains those values, *X* ceases to depend on the values of other variables that cause *X* and instead depends only on the value taken by *I*;
> 3. Any directed path from *I* to *Y* goes through *X*. That is, *I* does not directly cause *Y* and is not a cause of any causes of *Y* that are distinct

[1] This assumption is made for convenience and may be relaxed by replacing the notion of an "arrow-breaking" intervention with the notion of a "soft intervention", which merely supplies the variable intervened on with an exogenous source of variation, but does not break previously existing causal connections. See Eberhardt and Scheines, 2007 for details. The arguments below will go through, mutatis mutandis, with this weaker notion of intervention.

> from *X* except, of course, for those causes of *Y* , if any, that are built into the *I* → *X* → → *Y* connection itself; that is, except for (a) any causes of *Y* that are effects of *X* (i.e., variables that are causally between *X* and *Y*) and (b) any causes of *Y* that are between *I* and *X* and have no effect on *Y* independently of *X*;
> 4. *I* is (statistically) independent of any variable *Z* that causes *Y* and that is on a directed path that does not go through X. (Woodward 2003, 98)

The references to "directed paths" and so on refer to causal graphs in which a direct causal relation from *X* to *Y* is represented by means of an arrow from *X* to *Y* (*X*→*Y*). Woodward (2003) shows how one may characterize a notion of direct causation (and hence causal graphs) in interventionist terms; the details of this will not matter in what follows.

I will assume in what follows that if we are dealing with variables that do not stand in non-causal dependency relations, arrow-breaking interventions are possible on each variable in the graph—that is, that we can set any variable to any of its possible values via an intervention independently of the values taken by variables elsewhere in the graph. I will call this the assumption of Independent Fixability of values:

> (**IF**): A set of variables **V** satisfies independent fixability of values if and only if for each value it is possible for a variable to take individually, it is possible (that is possible in terms of assumed definitional, logical or metaphysical relations) to set the variable to that value via an intervention, concurrently with each of the other variables in **V** also being set to any of its individually possible values by independent interventions.

We may think of satisfaction (**IF)** as a necessary condition for a graph to count as a *causal* graph – that is a graph relating variables that may stand in causal relations but do not stand in any relations of non-causal dependency. A structure in which some variables stand in non-causal dependency relations (such as supervenience relations, as in Kim's diagram) will not satisfy **IF** and will not be a causal graph.

As I have observed elsewhere (**M**) (and **M***) characterize a rather weak notion of causation— (**M**) implies that *X* causes *Y* as long as there is *some* change in the value of *X* that leads to a change in the value of (or the probability distribution of) *Y* and parallel remarks apply to (**M***). As formulated, (**M)** and (**M*)** say nothing about *which* changes in *X* lead to changes in *Y*. One simple way of making (some of) this information explicit employs "rather than" or similar locutions to capture the contrastive structure of causal claims:

(**M****) *X=x* rather than *X=x'* causes *Y=y* rather than *Y=y' (*where *x≠x"* and *y≠y'* of course*)* if and only if there is a possible intervention on *X* that changes the value of *X* from *X=x to X=* such that if that intervention were to occur, the value of *Y* would change from *y* to *y'*.

I will say more about this issue in section 7 below, when I compare my view with Peter's.

4.

If we adopt the understanding of causation embodied in (**M**-**M***), and if neglect any possible complications arising from the supervenience of the mental on the physical, there appears to be no problem with attributing causal efficacy to mental states and properties. All that is required for a mental property M_1 to cause another mental property M_2 or to cause behavior B is that there be values of M_1 such that are possible interventions that change M_1 from one of those values to the other and such that under those interventions the values of M_2 or B change. Prima-facie at least, it looks as though this requirement is often met: it appears we often intervene on one another's mental states -- for example, when we successfully persuade or threaten some one in such a way as to change their beliefs and desires in such a way that the upshot is changes in other mental states or behavior? To the extent this is so, we have mental causation according to (**M**) and (**M***).

It has been argued by several writers (Baumgartner, 2009, 2010, Marcellesi, forthcoming), however, that this conclusion is undermined when we add back in the consideration that we have so far been neglecting: the supervenience of the mental on the physical and the way in which this interacts with the interventionist framework. In what follows I focus on Baumartner's version of this objection since his is the most fully developed. Focusing again on Kim's diagram, Baumgartner's argument is that far from supporting the claim that M_1 causes P_2 , interventionism is actually *inconsistent* with this claim. (I focus on the $M_1 \rightarrow P_2$ relationship in order to follow Baumgartner's exposition, but I think, in contrast to Baumgartner that if his objection is valid, it also rules out causation form M_1 to M_2 . Baumgartner's reasoning is as follows: Consider (**IV**). This requires that an intervention on X with respect to Y change the value of X in a way that is statistically independent of all other causes of Y that are not on a path from X to Y. Suppose one attempts to intervene on M_1 with respect to P_2. Then (Baumgartner claims) in Kim's diagram P_1 is such an off-path variable which is a cause of P_2. Moreover, it is built into the nature of the supervenience relation that M_1 cannot change in value without a change in P_1. But then it is impossible to change M_1 in a way that it is statistically independent of P_1 (or to change M_1 with P_1 being held fixed). Hence, according to Baumgartner, it is impossible to intervene on M_1 with respect to P_2. If there is no such possible intervention on M_1, then according to the interventionist account of causation, M_1 cannot cause P_2, since **M** requires that for such causation, interventions on M_1 must be possible. A similar line of reasoning seems to lead to the conclusion that M_1 cannot cause M_2 . In effect, Baumgartner's argument is that (**IV**) requires that in assessing the causal efficacy of M_1, one hold fixed or control for its supervenience base, and that this requirement implies that M_1 cannot cause M_2 (or P_2) .

In assessing these contentions, there are two issues that should be kept separate. One is an interpretive issue about what **IV** as formulated in Woodward, 2003 requires. The other, far more interesting issue is whether, regardless of what **IV** implies, one *ought* to "control for" supervenience bases in assessing the causal efficacy of supervening variables.

As far as the interpretive issue goes, Baumgartner's claims about what **IV** requires seem mistaken As explained above, **IV** is intended to apply to systems of causal relationships satisfying **IF** – that is, to *causal* graphs constructed according to the rules (governing direct causation etc.) given in Woodward, 2003. Baumgartner apparently

assumes that Kim's diagram is such a causal graph for the purposes of applying (**IV**) and that P_1 is the relevant sense, an "off-path variable "in such a graph. But this is not what **IV** says, assuming (as virtually everyone involved in this discussion, including Baumgartner, agrees) that supervenience relationship represented by the double-tailed arrow is a non-causal dependency relationship. **IV** says nothing about controlling for variables that are "off-path" in the sense that they stand in non-causal supervenience relations to the variable intervened on . Instead, in the context of **IV**, "off path" means "variable that is off-path in the causal graph that represents the causal structure of the system of interest". P_1 is not an off-path variable in this sense. So if we simply follow the letter of **IV**, we should not understand it as requiring that an intervention on M_1 must hold fixed the value of P_1.

Of course, this leaves open the possibility that **IV** is simply mistaken and that contrary to what it implies one ought to control for supervenience bases or at least that such control is required by the overall commitments of interventionism. After all, it might be said, interventionists agree that one should "control for" *some* other causes of Y in assessing whether X causes Y. Why shouldn't these "other causes of Y" that need to be controlled for include the supervenience basis $SB\ (X)$ for X, as the exclusion argument claims? If so, we should reformulate **IV** and other elements of the interventionist framework in such a way that they require such control. In the following section I address this issue.

5.

I think that there are several reasons why one should not control for supervenience bases in assessing the causal role of supervening variables. First, the arguments and considerations that support requirements for controlling for the effects of appropriate "other causes" (as detailed in **IV** and **M- M*)** when all of the relationships represented in the graph are between distinct variables and no non-causal dependency relations are present do not transfer to contexts in which non-causal relations of dependence are present. That is, it is a mistake (indeed one might say that it is *the* central mistake made by advocates of the exclusion argument) to assume that because it would be appropriate, in assessing the causal impact of M_1 on M_2, to control for or hold fixed variable P_1 in Kim's diagram in contexts in which P_1 *causes* M_1, it must also be appropriate to do this in contexts in which the relationship between P_1 and M_1 is one of non-causal supervenience. A second consideration is that control for supervenience bases leads to what, intuitively speaking, are mistaken causal inferences. In the absence of any positive reasons for such control, we should not require it.

Let me begin with the first point. Suppose that we are dealing with a structure like that in Kim's diagram but in which the double headed arrows are all replaced by single arrows representing causal relations—that is, P_1 causes M_1 and P_2 causes M_2 (and P_1 causes P_2 as before). In this sort of structure, if we wish to assess whether M_1 causes M_2, there is an obvious rationale for controlling for the presence of such other causes of M_2 as P_1 and P_2. This is because even if M_1 does not cause M_2 , the operation of P_1 may produce a "spurious association" between M_1 and M_2 in virtue of the fact that P_1 causes M_1 and causes M_2 via P_1. In this case, on the interventionist picture, a properly performed experiment (an intervention which satisfies **IV**) for assessing whether M_1

causes M_2 will be one which "breaks" the causal connection from P_1 to M_1, supplying M_1 with some independent source of variation besides P_1. If M_1 and M_2 remain correlated under this intervention, M_1 has an independent causal impact on M_2. If not, M_1 does not cause M_2 and their association is spurious. Here the notion of a spurious association has a clear meaning and a clear practical import. When the relation between P_1 and M_1 is causal, it follows from (**IF**) that it is causally possible for M_1 to occur in the absence of P_1 -- indeed this is what happens when an intervention on M_1 occurs. In such a situation, if the association between M_1 and M_2 is spurious, the association between M_1 and M_2 will disappear. In this case, the test for whether M_1 causes M_2 or instead their relation is spurious has to do with what will happen to M_2 if one were in fact to manipulate M_1.

When M_1 supervenes on P_1 the logic of the situation is quite different. In this case, by hypotheis, it is "metaphysically impossible" to break the relationship between P_1 and M_1 in the sense of changing M_1 independently of changing P_1 or supplying M_1 with a source of variation that is independent of P_1 . Instead of arguing, as we do in the case in which $P_1 \rightarrow M_1$ relation is causal, that M_1 does not cause M_2 (or P_1) on the grounds that if were we to change M_1 independently of P_1, there would be no change in M_2 (or P_1), any conclusion we reach about the causal inertness of M_1 when M_1 supervenes on P_1 is instead based on the impossibility of changing M_1 without changing P_1. That is, we in effect conclude that M_1 is causally inert from the *impossibility* of changing M_1 while holding P_1 fixed. By contrast, in the case in which the relation between P_1 and M_1 is causal, we reason to the causal inertness of M_1 on the assumption that it is *possible* to change M_1 independently of P_1 and that under such a change M_2 would not change .

In circumstances in which the supervenience of M_1 on P_1 makes it impossible to change M_1 without changing P_1, the claim that any association between M_1 and M_2 (or between M_1 and P_2) is "spurious" or non-causal does not have the significance or implications that it has in cases in which P_1 causes M_1 and also acts via some independent path as a cause of P_2 or M_2. When the presence of a supervenience relation assures us that it is impossible to change M_1 without changing P_1, we don't have to worry about the possibility of situations arising in which P_1 causes M_2, M_1 does not, and in which we or nature alter M_1 but P_1 remains unchanged leading to no change in M_2 or P_2. Instead, any change in M_1 must also at the same time change P_1 and whatever downstream effects of P_1 result from this change. So we don't have to worry about the kind of possibility that arises in more ordinary cases in which the association between M_1 and M_2 is spurious, which does carry with it the implication that attempting to exploit the relationship between M_1 and M_2 will simply cause the association between M_1 and M_2 to disappear.

To drive this last point home, consider the following contrast. Researcher 1 is interested in whether administration A of a drug produces recovery R from a certain disease. She is worried that the previously observed association between A and R may be spurious and in particular due to the fact that the drug has been given preferentially to healthier patients with better immune response, where the goodness of immune response is measured by a variable I. Researcher 1 does a randomized control trial (which among other things, balances values of I across the treatment and control group) and finds an association between A and R. She takes this to be strong evidence against the possibility that the A-R association is spurious.

This conclusion is challenged by researcher 2 on the following grounds: The values of *A* (whether or not the drug is administered) for any individual patient supervene on certain microphysical features *P* having to do with the molecular structure of the drug, whether these are present in the patient's body etc. Assuming that non-reductive physicalism is the correct picture of the relation between *P* and *A*, these features *P* are a possible cause of recovery which is distinct from *A*. Hence (following the logic of the exclusion argument) in order to demonstrate that *A* really causes recovery, one would have to do an experiment in which *A* is manipulated while *P* is held fixed. Researcher 2 notes that given the nature of the supervenience relation any such experiment is impossible and infers from this impossibility that *A* cannot cause recovery.

On any generally accepted conception of experimental methodology, there is a difference in the cogency of the reasoning of the two researchers. Researcher 1 reasons from the results of an actually performed experiment. Researcher 2 reasons to a conclusion about the causal inertness of *A* on the basis of the impossibility of performing a certain experiment. The worry about spuriousness addressed by the first experimenter is a live, practical worry—if the previously observed association between *A* and *R* is due entirely to the fact that the drug has been given differentially to those with stronger immune systems, then administering the drug is not an effective strategy for promoting recovery, a fact that would soon be discovered if the drug were given to populations of patients in which *A* and immune response *I* are not correlated. No similar worry is being addressed by the second researcher. Since anyone who receives the drug must also satisfy some appropriate version of the microphysical description *P* and by hypothesis *P* causes recovery, there is no possibility that anyone to whom the drug is administered will fail to have a higher probability of recovery. While *I* is an alternative cause of recovery that is independent of *A* in the sense that *I* can fail to be correlated with *A* (or indeed can have any arbitrary correlation with *A* between 0 and 1), the corresponding claim is not true of the relation between *A* and *P* since these variables cannot fail to be correlated—the exclusion argument's judgment that *A* does not "really" cause recovery does not license any corresponding worry that manipulating *A* will fail to be a successful strategy for promoting recovery. This is reflected in the fact that, as a matter of ordinary scientific methodology, the randomized trial described above is commonly regarded as one of the best possible sources of evidence for whether *A* causes recovery; that the causal exclusion argument conflicts with this judgment is an indication of the extent to which it relies on assumptions about causation and causal inference that are inconsistent with those ordinarily made about experimental design.

Yet another consideration in favor of this conclusion is suggested by how it seems appropriate to reason in analogous cases involving logical or mathematical dependency. Suppose, following Spirtes and Scheines, 2005 that we have a graph (*not* a causal graph) containing variables representing total cholesterol (*TC*), high density cholesterol (*HDL*) and low density cholesterol (*LDL*) where as a matter of definition, $TC = LDL + HDL$. Suppose we are interested in the effect of these variables on a health-outcome variable *H*. Consider the following reasoning: both *LDL* and *TC* are (or may imagined to be) causes of *H*. Therefore, in assessing the effect of *HDL* on *H*, we should hold fixed the values of *LDL* and *TC*. Of course we find that it is impossible, for definitional reasons, to manipulate *HDL* while holding fixed *LDL* and *TC*. Therefore, it is impossible to intervene on *HDL* and we conclude that *HDL* has no effect on *H*.

This reasoning seems perverse. The (definitionally grounded) impossibility of manipulating *HDL* while holding *LDL* and *TC* fixed is not a good reason for concluding that *HDL* is causally inert with respect to *H*. Intuitively, it seems unreasonable to demand that the notion of an intervention on *HDL* be understood in such a way that this requires changing *HDL* while holding fixed *LDL* and *TC*. Instead, it seems much more natural to understand the notion of an intervention in this context as operating in such a way that an intervention that changes *HDL* by amount Δ*HDL* will at the same time amount to (that is, it is the same intervention as) an intervention that changes *TC* by amount Δ *HDL*. Moreover, in determining the effect of *HDL* on *H*, we need to avoid "double counting" of the effects of the same intervention twice. If we draw arrows from *HDL* to *H* and from *TC* to *H* to represent causal relations and, say, a double-tailed arrow from *HDL* to *TC* to represent definitional dependence, then it is a mistake to reason that if *HDL* is changed by Δ*HDL*, this will also change *TC* by Δ *HDL,* with the result that the change in *H* will reflect the change in both of these variables – a change in *H* due to *HDL* and an additional similar change due in *H* due to *TC*. Instead, when an intervention occurs on *HDL* there is just one change of magnitude Δ*HDL* which affects *H*.

The supervenience relation thought to be present in non-reductive physicalism is of course not the same as mathematical/definitional relation present in the cholesterol example. Nonetheless the same basic points concerning causal inference seem to apply. When logical/mathematical relations are present, it seems uncontroversial that we will make mistaken causal inferences if we attempt to "control for" variables that are related as a matter of definition. That is, we get into difficulties when we treat non-causal relations of definitional dependency as though they are ordinary causal relations, and import demands for control and holding fixed that are appropriate to the latter into the former context. I claim that parallel points apply when supervenient relations are present.

6.

These considerations seem to support the conclusion that one should not control for (condition on) supervenience bases in assessing the causal efficacy of supervening properties but they do not give us a positive account of how we should think about interventions and causal relations in structures in which supervenience or other sorts of non-causal dependency relations are present. A full account of this sort is beyond the scope of this essay but let me suggest a few principles which will play a role later in my discussion.

First, in parallel with the treatment of definitional dependence above, when supervience relations are present, an intervention on *X* should be treated as automatically changing (indeed as also the same intervention on) the supervenience base *SB(X)* of *X*, with *SB(X)* changing in whatever way is required by the supervenience relation between *X* and *SB (X)* (or as changing in some way or other that is consistent with the supervenience relation if there are multiple possibilities). For example, an intervention that changes the value of some variable *M* representing a mental property, should be treated as at the same time changing the value of the physical variable *P* on which it supervenes.

Second, and consistently with this, when an intervention occurs on X, its supervenience base *SB(X)* should *not* be regarded as one of those "off route causes" in **IV** that one needs to control for or hold fixed in intervening on X. To be more explicit, when supervenience relationships are present, the characterization **IV** should be interpreted in such a way that in condition (I3) a directed path counts as "going from I to Y through X" even if I also changes (as it must) the supervenience base *SB (X)* of X, as well as the value of X. Similarly, the reference in (I4) to "any variable Z" should be interpreted as "any variable Z other than those in the supervenience base *SB(X)* of X". Put slightly differently the requirements in the definition (**IV**) should be understood as applying only to those variables that are causally related to X and Y or are correlated with them but and not to those variables that are related to X and Y as a result of supervenience relations or relations of definitional dependence.

Third and again consistently in applying the characterization (**M)** (and for that matter the characterizations of other causal notions in Woodward, 2003) we should consider as well-defined only those interventions or combinations of interventions that involve setting variables to combinations of values that are "co-possible" in the sense of satisfying whatever non-causal dependency relations are assumed to be present. Combinations of "interventions" that are "impossible" in the sense of violating whatever non causal dependency relations are assumed to be present are not well-defined interventions, and claims about what would happen under such impossible "interventions" have no bearing, one way or another, on which causal relations obtain. In other words, in assessing whether X causes Y we should consider only possible interventions on X (and where relevant other variables), where "possible" here means consistent with whatever non-causal dependency relations are assumed) and ask whether Y would change under any of these interventions.

Finally, because interventions on some variables may amount to interventions on other variables to which they bear non-causal dependency relations, we need to take account of this fact in tracking causal relationships, so that we avoid double-counting.

To illustrate these ideas, let me apply them to a structure like that in Kim's diagram but with some additional assumptions made explicit. Suppose that we think of N_1, N_2, M_1, and M_2 as variables, with the values of M_1 and M_2 supervening on values of N_1 and N_2 respectively. In particular, M_1 can take two different values- m_{11} and m_{12}. N_1 can take three different values, with n_{11} and n_{12} realizing m_{11} and n_{13} realizing m_{12}. M_2 can take values m_{21} and m_{22}, with n_{21} and n_{22} being different realizations of m_{21} and n_{23} realizing m_{22}. Suppose that n_{11} leads to n_{21}, n_{12} leads to n_{22}, and n_{13} leads to n_{23}, so that m_{11} is followed by m_{21} and m_{12} is followed by m_{23}. Then N_1 causes N_2 because there is a change in the value of N_1 (from e.g. n_{12} to n_{13}) that when produced by an intervention is associated with a change in N_2 (from n_{22} to n_{23}). It can also easily be checked that M_1 causes M_2, and that N_1 causes M_2 and M_1 causes N_2. So under these assumptions the associated causal graph looks like this (with supervenience relations omitted) :

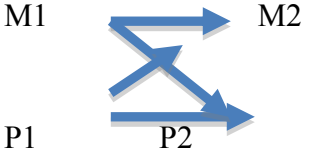

Figure 2

In interpreting this graph, we need to take care to follow the conventions described above and to avoid double counting. For example, when an intervention I changes M_1 from m_{11} to m_{12}, we should think of this intervention as at the same time producing a consistent change in N_1 (e.g. from n_{11} to n_{13}). And in assessing the causal impact of this change on M_2 we need to avoid counting it twice, as two independent changes in M_2 , one of which occurs along the M_1—> M_2 route and the other through the N_1→ N_2 -- M_2 route. Instead there is just one intervention on M_1/ N_1 and just one associated change in M_2. Finally, it is important to bear in mind what an arrow from, say, M_1 to M_2 means within the interventionist framework: *all* that it means is that there are possible interventions on M_1 that will change M_2. There is no inconsistency between this claim and the claim that interventions on M_1 will also change N_2, interventions on N_1 will change M_2 and so on. The arrows from M_1 to M_2 and to N_2, do not mean that there are independent, separately disruptable causal processes linking M_1 to M_2 and linking M_1 to N_2,.

Before leaving this example, let me make a further observation about Kim's diagram and about the exclusion argument which I hope will illustrate the power of the interventionist framework and the way in which it can afford novel insights. Both adherents and critics of the exclusion argument often seem to suppose that the information explicitly represented in Kim's diagram (that N_1 causes N_2, M_1 supervenes on N_1, M_2 supervenes on N_2) is enough to "fix" or determine the overall pattern of association between M_1 and M_2, with the only question being whether this association should be regarded as causal. Within a broadly interventionist framework, this assumption is mistaken. The information explicitly represented in the diagram does not even determine whether M_1 and M_2 are *correlated* or instead *independent*, when causal claims are understood along interventionist lines.

As one possible illustration of how this may happen, consider the following variation on the example immediately above. As before, M_1 has two possible values, m_{11} and m_{12}. M_1 supervenes on N_1, which now has four possible values n_{11}, n_{12}, n_{13} and n_{14}. m_{11} may be realized by either of the values n_{11}, n_{12} and m_{12} may be realized by either n_{13} or n_{14}. M_2 has two possible values m_{21} and m_{22}. M_2 supervenes on N_2, with possible values n_{21}, n_{22}, n_{23} and n_{24}. n_{21} and n_{22} are the realizers of m_{21} and n_{23} and n_{24} are the realizers of m_{22}. The causal relationship linking N_1 to N_2, is such that the values n_{11}, n_{12}, and n_{13} are followed by n_{21} and the value n_{14} is followed by n_{22}. This counts as a causal relationship between N_1 and N_2 within the interventionist framework since there are interventions that change N_1 (e.g., from n_{11} to n_{14}) that are followed by changes in N_2 (from n_{21} to n_{22}). However, it is easily verified that changes in M_1 from m_{11} to m_{12} (or vice-versa) are not associated with (do not lead to) changes in M_2. This is because all of the realizations of the different possible values of M_1 lead via the pattern of dependence of N_2 on N_1 to realizations of the same value of M_2 (namely m_{21}).

This example illustrates how different accounts of causation (and in particular whether one thinks of causation in terms of something like nomological sufficiency or instead along interventionist lines) have different implications for how one interprets and assess the exclusion argument. In Kim's diagram, any particular occurrence of a

value of M_1 will have some physical realizer – say n_{11}-- which is a value of N_1. If we interpret the claim that N_1 causes N_2 as implying simply that this realizer n_{11} is nomologically sufficient for whatever value is taken by N_2 then, in virtue of the supervenience relation between N_2 and M_2, n_{11} is also sufficient for the value taken by M_2 on this occasion. However, it does not follow that the values taken by M_2 *causally depend* on the values of M_1 in the sense that there are different possible values of M_1 which will lead to different values of M_2, or that changing the value of M_1 is a way of changing the value of M_2, which is the notion of causation the interventionist account attempts to capture. Nomological sufficiency is different from causal dependence, as many other examples in the explanation literature illustrate[2].

This in turn has an important additional consequence. Since (when causation is understood along interventionist lines) in some cases in which N_1 causes N_2 and the supervenience relations in Kim's diagram are present, M_1 will cause M_2 and in other cases of this sort M_1 will not cause M_2, we convey important additional information when we draw or omit to draw an arrow from M_1 to M_2. Whether M_2 changes under an intervention on M_1 when interventions are understood along the lines described above, depends not just on whether N_1 causes N_2 and M_1 (M_2) supervenes on N_1 (N_2) but on the further details of the way in which these variables are related to one another, matters which are not specified in Kim's diagram. So in this sense, the suggestion in standard formulations of the exclusion argument that it is redundant or superfluous to suppose that M_1 causes M_2, given the other information in Kim's diagram, is mistaken.

7.

Peter and Christian List are also critical of the exclusion argument, on grounds that have some similarity to but are nonetheless distinct from those advanced above. Their argument relies on the idea that causes are "proportional difference makers" for their effects. They propose the following "truth conditions" for a property F "making a difference" to a second property G:

> (**P**) The presence of F makes a difference to the presence of G in the actual world if and only if it is true in the actual world that (i) F is present > G is present and (ii) F is absent > G is absent. (List and Menzies, 2009, here > is the counterfactual conditional)

They then argue as follows: Consider (following Woodward, 2008) the claim possession of an intention I_1 causes a monkey to perform action A_1. Suppose that I_1 is "realized" on this particular occasion by some neural structure N_{11} but because I_1 is multiply realizable, it might also have been realized by neural structure N_{12}. It is plausible that (or at least the situation may be imagined to be one in which) the following two counterfactuals are true:

(1a) If I_1 were present, then A_1 would be present
(2a) If I_1 were not present, then A_1 would not be present

[2] Cf. Salmon's (1984) example about the male who takes birth control pills.

Thus the presence of I_1 is a difference-maker for A_1 .

Now compare:

(1b) If N_{11} were present, A_1 would be present
(2b) If N_{11} were not present, A_1 would not be present

Menzies and List claim that (2b) is false, on the following grounds: At least some worlds that are "closest" to the actual world in which N_{11} fails to occur will be worlds in which an alternative realizer N_{12} of I_1 occurs. In these worlds, A_1 will occur, rendering (2b) false. They conclude that under these conditions, with "causes" understood as "difference-makers" in the sense captured by (**P**), that the claim

(3) I_1 causes A_1

is true and the claim

(4) N_{11} causes A_1

is false. We thus have a case of what Menzies and List call "downward exclusion", in which the truth of the "upper level" causal claim (3) excludes the truth of the lower level claim (4).

As explained above, my view is different from this. Switching from property-talk to variable-talk, call the relevant neural variable N (where N takes values corresponding to n_{11}= presence of N_{11}, n_{12} = presence of N_{12} etc.) and the action variable A (its values include, say, presence of A_1 and presence of A_2). Then as long as there are some possible changes in the value of N which, when produced by interventions, are associated with changes in A , (4) will be true. (4) will be true if, for example, N has (in addition to its values n_{11} and n_{12}) a value n_{13} = presence of N_{13} that leads to the absence of A_1. Then, because a change in N from n_{11} to n_{13} leads to a change from the presence of A_1 to the absence of A_1, N_{11} causes A_1. If, in addition, N_{11} and N_{12} are the only two realizers of I_1, it will also be true that I_1 causes A_1. On my view, under these conditions, there is neither downward nor upward exclusion.

As remarked at the end of Section 3, I agree that there is an important respect in which (4) by itself is less perspicuous or informative than one would like: it fails to tell us under specifically which alternatives to N_{11}, A_1 would occur or fail to occur. One might remedy this by specifying explicitly what the pattern of dependence of A on N is – by, for example, using the "rather than" locution along the lines of **(M**)**: the occurrence N_{11}, rather than N_{13} caused A_1 rather than A_2 but if N_{12} rather than N_{11}, had occurred, A_1 (still) would have occurred. Or, more simply, one might simply specify what would happen to the A variable under each of the three values that N can take. Call this causal claim (6). I see no obvious reason to regard (6) as inferior in terms of informativeness to (3), the claim that attributes causal efficacy to the intention I_1 and certainly no reason to think that (6) is false and (3) true. This suggests to me that it is wrong to suppose that in the example as described there is, so to speak, no causation of A_1 going on at the neural level

characterized by N. Rather what is going on is simply that (4) by itself, is less informative and perspicuous than one might like, when compared with (3), a problem that can be remedied by spelling out (4) along the lines of (6).

I will add that I agree with Menzies and List that the counterfactual (2b) is false, although my grounds for thinking that it is false are different from theirs[3]. But the more important point, from my point of view, is that the truth of (2b) is not a necessary condition for the truth of (2). As claimed above, the truth of (2) requires only that some counterfactual of the form "If N were to take a different value from n_{11} (or if N were to take value N_{13} rather than n_{11}) then A_1 would not occur" be true. Indeed if we were to require, as a necessary condition for the truth of causal claims, that the second counterfactual in (**P**) be true, a very large number of causal claims that we ordinarily take to be true would turn out to be false[4].

---

[3] This issue deserves more attention than I can give it here, but very roughly I do not think that it matters in evaluating (2b) whether some or all of the worlds in which N_{11} does not occur and which are *closest* to the actual world are worlds in which A_1 does not occur. On my view, (2b) is false if *any* of the realizers of its antecedent (regardless of their closeness to their actuality in comparison with other realizers) are followed by the occurrence of A_1. Thus to argue that (2b) is false, we do not need to show, as Menzies and List attempt to do, that some worlds in which N_{11} does not occur are worlds in which N_{12} occurs instead. The falsity of (2b) does not require that counterfactuals of the form "if N_{11} had not occurred, then such and such an alternative to N_{11} would (or even *might*) have occurred" be true. One way in which this issue matters concerns the validity of the inference from (AvB) > C to (A>C). (B>C). I take this inference to be valid within an interventionist framework; I believe that Menzies' and List's treatment follows Lewis in judging it to be invalid.

[4] Consider a variant on Glymour's (1986) example of S, who smokes 4 packs of cigarettes a day and develops lung cancer. Assume that if S had smoked any amount in excess of 2 packs, S would have developed lung cancer. Is it true that (7) S's smoking 4 packs a day caused his lung cancer? Applying (P) and evaluating counterfactuals in the way List and Menzies suggest, (7) comes out false, since S will develop lung cancer in close-by worlds, such as those in which he smokes 3.9 packs. I find it more natural to follow (M) in regarding claims like (7) as true, albeit less informative than one might like. In general, it seems that our usual practice is not to follow (P) in requiring that for C causes E to be true, the counterfactual (8) "if C had not occurred, then E would not have occurred" must be true, at least when this counterfactual is evaluated in the way that Menzies and List suggest. It is worth noting in this connection that Lewis (1986) tells us that in evaluating the counterfactual "if C had not occurred, then E would not have occurred" in connection with the claim "C causes E" we would should consider worlds in which C is "wholly excised", rather than worlds which, so to speak, involve very small departures from C. Presumably this means that the relevant counterfactual for evaluating (7)is one whose antecedent has S not smoking at all (or smoking very little) rather than smoking 3.9 packs, which in turn leads to (7) be regarded as true in Lewis' framework. I am grateful to Chris Hitchcock for helpful discussion of this issue.

References

Michael Baumgartner (2009). "Interventionist Causal Exclusion and Non-Reductive Physicalism". *International Studies in the Philosophy of Science* 23 (2):161-178.

Baumgartner, M. (2010). "Interventionism and Epiphenomenalism"**.** *Canadian Journal of Philosophy* 40: 359-383.

Glymour, C. (1986) "Comment: Statistics and Metaphysics" *Journal of the American Statistical Association* 81: 964-966.

Eberhardt, F. & Scheines, R. (2007) "Interventions and Causal Inference". *Philosophy of Science*, 74:981-995.

Lewis, D. (1986) Postscripts to "Causation" in *Philosophical Papers II.* Oxford: Oxford University Press.

List, C. and Menzies, P. (2009). "Nonreductive Physicalism and the Limits of the Exclusion Principle." *Journal of Philosophy* 106: 475-502.

Marcellesi, A. Forthcoming. "Manipulation and Interlevel Causation"

Salmon, W. (1984) *Scientific Explanation and the Causal Structure of the World*.

Spirtes, P., and Scheines, R. (2005). "Causal Inference of Ambiguous Manipulations". *Philosophy of Science* 71: 833-45.

Woodward, J. (2003) *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

Woodward, J. (2008) "Mental Causation and Neural Mechanisms " in Hohwy and Kallestrup, eds. *Being Reduced: New Essays on Reduction, Explanation, and Causation*. Oxford University Press: 218-262.