Forthcoming in *European Journal for the Philosophy of Science* (December, 2022)

**Modeling Interventions in Multi-level Causal Systems: Supervenience, Exclusion and Underdetermination.**

James Woodward

jfw@pitt.edu

1. **Introduction**.

There is considerable recent interest in issues having to do with application of causal modeling techniques to contexts in which models and variables at different “levels” stand in non-causal relations of multiple realization or non-reductive supervenience. (In what follows, in order to avoid needless repetition, when I talk about relations between levels, I will assume that that it takes this multiple realization/non-reductive supervenience form. I will call this the *levels assumption*.) One issue that has received particular attention concerns how one should think about interventions (in the sense of unconfounded manipulations, as described in Pearl, 2009 Spirtes, et al., 2001, Woodward, 2003) in such multi- level contexts—this issue is left unaddressed by these publications. In the philosophical literature much of this discussion has focused on the implications of a broadly interventionist framework for so-called causal exclusion arguments. In particular, several writers (e.g., Baumgartner, 2010, Baumgartner et al. , 2018) have argued that if upper-level variables (e.g., mental variables) supervene non-reductively on lower level (e.g., physical variables) , it follows from a proper understanding of what an intervention involves that interventions on upper-level variables are impossible. Indeed, a recent paper by Baumgartner et al., 2018 claims:

recent discussions have *shown* that…there are no ideal interventions on upper-level phenomena that non-reductively supervene on their underlying mechanisms”. (my italics)

It has also been claimed in another recent paper by Baumgartner (2018) that it follows from the levels assumption and an interventionist treatment of causation that it is impossible in principle to obtain evidence that upper-level properties are causally efficacious.

These claims will be discussed in more detail below, but their guiding idea is that, given the levels assumption, attempted interventions on upper-level variables are inevitably ”confounded” by their supervenience bases, in a way that implies that they are not legitimate interventions for purposes of causal analysis. Since, within an interventionist framework of the sort described in Woodward, 2003, a necessary condition for a variable to be causally efficacious is that it be possible to intervene on it, it follows from the above claims (again given the interventionist framework) that such upper-level variables are not causally efficacious[[1]](#footnote-1). Obviously such claims rely on a assumptions about how we should understand interventions on upper-level variables when the levels assumption holds.

At the same time, some philosophers and researchers in computer science/machine learning have discussed a closely related problem. Suppose one has a lower or micro-level model in which certain causal relations obtain, where causal relations are understood in terms of responses to interventions, again along the lines described in Woodward, 2003 and others. Suppose that the variables in the micro-model satisfy the levels assumption with respect to the variables in a macro-model—in particular, assume that the variables in the macro-model arise from a coarse-graining of the variables in the micro model. (I make this more precise in Section 3.) Assume that the macro-model also postulates causal relations, again understood along interventionist lines. The problem is this: assuming that the micro-model is correct (it gives the “ground truth” about causal relations among the micro-variables) under what conditions are the causal relations postulated in the micro-model “preserved” in the macro-model? Put differently, what do the causal relations among the micro-variables imply about causal relations among the macro-variables? (This issue is discussed in, e.g., Rubenstein et al., 2017 and Beckers et al., 2019-- see Sections 3 and 6 .) More ambitiously, there is a closely related causal discovery /variable choice problem: suppose one knows the correct causal relations in the micro-model. How might one form a macro-model with more coarse-grained variables that recovers (for some set of phenomena of interest) the causal relations implied by the micro-model? (Discussed in Chalupka et al., 2017.) As we shall see, in this case too a central part of the solution to the problem involves finding an appropriate way of thinking about the relationship between interventions on macro -level variables and on their micro-level realizers that allows for a well-behaved mapping of micro to macro-level causal information. These researchers are looking for what might be described as a “compatibilist” (that is, non-exclusionist) solution to the relation between micro and macro-level interventions, according to which the two sets of interventions fit together in some appropriate way and interventions on upper-level variables are not impossible. Compatibilist solutions have also been advanced by philosophers working within an interventionist framework—see Woodward, 2015, 2017, 2021.

Clearly there is a close connection between the project just described and philosophical discussions of the exclusion argument. If interventions on upper-level variables are impossible, given the levels assumption, it does not make sense to investigate the conditions under which the causal relations and the results of interventions on micro-level variables can be coherently mapped into interventions involving macro-models. If exclusionism is correct, there are no genuine macro-interventions to relate to micro-interventions and upper -level variables are always causally inert.

It thus seems relevant to assessments of the exclusion argument that the philosophers and computer scientists who are working along compatibilist lines have proposed formal models which appear to consistently relate micro and macro causal models along the lines described above. These models do not characterize macro-interventions in such a way that they are confounded by their supervenience bases—thus, given the assumptions of these models, exclusionist arguments do not go through. Unless these models are incoherent or otherwise objectionable, they constitute an existence proof that it is possible to characterize macro-level interventions and macro-level causal claims in a way that avoids exclusionist conclusions.

My goals in this essay are severalfold. First, I want draw on the literature referenced above to set out a compatibilist account of how interventions work in multi-level contexts. This will allow us to make sense of interventions on upper-level variables and of causal claims relating such variables. Assuming that this account is coherent, it follows that a commitment to exclusionist conclusions within an interventionist framework is not, contrary to what many exclusionists seem to suppose, a “forced move “ or in some way dictated by non-negotiable commitments of interventionism. The fact that exist coherent compatibilist accounts of the sort described makes it clear that is not true that it can be “shown” or “proved” just from the assumptions characterizing interventionism, the Bayes net formalism, and other uncontroversial assumptions about causation that interventions on upper-level variables are impossible. In fact, as we shall see, the arguments for impossibility all rest on adding further claims to the interventionist framework -- claims that interventionists can and should deny. Very roughly these are claims that one needs to control for or hold fixed supervenience bases in characterizing how interventions behave in multi-level contexts or, put differently, that it is appropriate to think of supervenience bases as confounders. It is the presence of these additional claims that resolves the puzzle presented earlier concerning how it can be that some philosophers contend that exclusionist conclusions follow from interventionist premises while other researchers construct actual models in which this is not this is not the case. The exclusionist philosophers adopt these additional claims while the compatibilists do not.

This observation by itself does not show we should reject accounts that have exclusionist implications, but it does raise the issue of the grounds on which we should choose between compatibilist and exclusionist accounts. Should we accept the additional assumptions that when combined with standard interventionist premises imply exclusionist conclusions? I will argue that we should reject these additional assumptions and that the compatibilist account should be preferred because it fits best with goals associated with causal reasoning and that, moreover, the exclusionist account has no good motivation in terms of these goals, and indeed frustrates them. I thus claim that the compatibilist account can be justified on the basis of what Woodward, 2014 calls a functionalist account of causation—an account that understands causal reasoning (and the constraints that are appropriately imposed on causal reasoning) in terms of the function or purpose of such reasoning. The justification for adopting a compatibilist account is thus in a broad sense “methodological”. As I will explain, that this does not mean that adoption of the compatibilist account is an arbitrary or subjective choice. Compatibilist accounts and the causal claims they license are, as Baumgartner (2018) writes at one point, “modeling choices” but they are well-motivated modeling choices. By contrast, accounts with automatic exclusionist implications are (I will claim) “bad” modeling choices.

I acknowledge that this framing in terms of methodology is very different from the usual way of thinking about the conflict between exclusionist and compatibilist treatments of upper-level causation. As noted above, those favoring exclusionism often write as though this conclusion follows with inexorable logic from assumptions that are built into interventionism or perhaps the “metaphysics” of causation more generally. But once it is recognized that one can coherently characterize interventions in multi-level contexts in a way that exclusionist conclusions are avoided, it seems clear that such claims of unavoidability are mistaken. To decide between the two frameworks we thus need to appeal to other considerations, and this is what my invocation of functional/methodological considerations attempts to do.

This paper extends the arguments in Woodward, 2015, 2021 and other recent papers in several ways. It responds to recent discussions of interventionism and exclusionist arguments that have not previously been addressed by Woodward or others. These include Baumgartner's (2018) arguments that given interventionist premises it follows that it is underdetermined by all empirical evidence whether mental causes are efficacious and that interventions must be modeled as common causes of upper-level variables and their supervenience bases. I also distinguish the treatment of interventions proposed in this paper from alternative recent proposals due to Polger et al. (2018) and to Zhong (2020). In connection with the former, I emphasize the importance of not relativizing the notion of an intervention to a variable set. In addition, as noted above, this paper discusses recent work in the computer science/machine learning literature on the representation of interventions when supervenience relations are present and relates it to the proposals in Woodward, 2015. This work has not been discussed so far either by those who are sympathetic to interventionism or those who are critical of it, despite its relevance to current discussion

Let me add that the issue of how to think about interventions at different levels which I will be exploring below is an important problem in its own right, independently of exclusion arguments, and should be of independent interest for that reason. In the course of my discussion. I also make some novel suggestions about how multiple realizability is best understood that may be of some interest, stressing in particular that this should be understood as relation between values of variables rather than a relation between properties [[2]](#footnote-2).

To forestall a possible confusion, I also want to emphasize that the discussion that follows is mainly focused on the issue of what happens when interventions are performed in multi-level contexts -- what the connections between upper and lower levels *are* when one intervenes on an upper-level variable, in contrast to the epistemological issue of how we might come to *know* about such connections. For example, I assume below that the relationship between the values of upper-level variables and the values of their lower-level realizers can be represented by a many-to -one function satisfying certain other conditions, but this should not be understood as the claim that people must know this function when they reason with upper-level variables. On the contrary, I assume that in many cases in which we reason with or manipulate upper-level variables, we have little or no information about how these are realized at lower levels, or how the upper and lower-level variables are connected. This is not to say that the issue of discovering such connections is unimportant-- on the contrary[[3]](#footnote-3). But it is is not my topic in this paper.

1. **Preliminaries**.

Like others who have discussed these issues within an interventionist framework, I assume that causal relationships can be represented by directed graphs. A single –tailed arrow from *X* to *Y* (*X🡪Y*) in a graph represents that *X* is a *direct cause* of *Y.* (The notion of direct causation is characterized in more detail below.) I will use the expression *causal graph* to refer to a directed graph in which the only arrows or connections between vertices are those that represent direct causal relationships. That is, a causal graph does not contain representations of non-causal relationships such as supervenience relationships[[4]](#footnote-4). Thus by definition “mixed” graphs (such as Kim’s iconic graph in Figure 1 below) in which both causal relationships and non-causal supervenience relationships are represented—the former by means of single tailed arrows, the latter by means of double-tailed arrows-- are *not* causal graphs. The distinction between causal and mixed graphs is crucially important because, as we will see, these structures obey different requirements and conform to different inferential rules. We should not automatically assume that the two sorts of graphs behave in the same way. As I see it, arguments for exclusion in this context virtually always involves claims (either assumed or argued for, but not convincingly) that the two sorts of graphs behave in the same way.

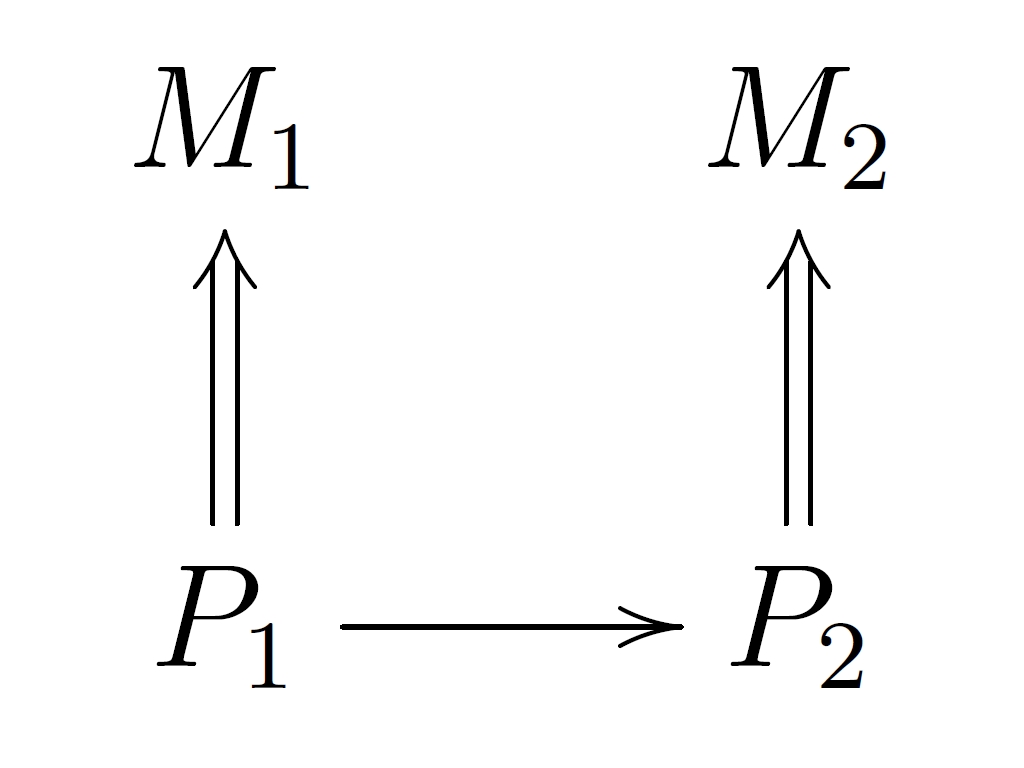


Figure 1

**Fig 1:** *P1* causes *P2*, *M1* supervenes on *P1*, *M2* supervenes on P2.

When we employ a causal graph or a mixed graph, it is crucial that we provide rules governing how the arrows are to be interpreted, when it is appropriate or not to draw them, and the inferences they warrant. It is not acceptable simply to draw arrows of various sorts without explaining what they mean or the constraints that they must satisfy. Call this the *interpretation constraint*. Woodward (2003) provided such rules for causal graphs (but not for mixed graphs). Confining ourselves for the moment to *causal* graphs, let us characterize an intervention *I* on a variable *X* with respect to a second variable *Y* in the following way: *I* changes *X* in such a way that *I* becomes the only cause of *X* and any change in *Y* occurs only through this change in *X* and not in some other way[[5]](#footnote-5). One implication of this last requirement is often expressed in terms of the idea that *I* should not affect *Y* via an “off-path” variable, where an off-path variable is one that is on a causal path[[6]](#footnote-6) to *Y* does not go through *X*. For example, *Z* is such an off-path variable in Figure 2 below and in this figure *I* does not count as an intervention on *X* with respect to *Y* since *I* affects *Y* via Z. Intuitively an intervention on *X* with respect to *Y* is an unconfounded manipulation of *X* with respect to *Y*. It is important to understand that the notion of an intervention just described is *not* relativized to any particular causal graph or variable set. *I* only counts as an intervention on *X* with respect to *Y* if there *is* no causal graph G that correctly represents the causal structure under discussion in which there is a causal path from *I* to *Y* that does not go through *X*. In other words, the requirement is that there are no causal factors that confound the intervention and not merely that no such factors are present in whatever graph or variable set we employ or that are known to us.

The reason for characterizing interventions in this way should be obvious: a manipulation of *X* that is confounded by unknown or unrepresented causes is still confounded and this fact can undermine causal inferences involving *X*. For example, *I*  is confounded in an inference designed to determine whether *X* causes *Y* if the true causal structure is represented by Figure 2, even if the investigator employs a graph in which *Z* is not represented.

*I X Y*

*Z*

Figure 2

**Fig 2**: *I* causes *Y* via a route (involving *Z*) that does not go through *X*

One of the virtues of a randomized control experiment is that if properly implemented it removes (or has least has a substantial probability of removing) confounders which are not known to or represented by the experimenter as well as those that are. Such randomization is one way of implementing an intervention. I stress this point because some (e.g. Polger et al. 2018) have argued that one can successfully counter pro-exclusion arguments within an interventionist framework by relativizing the notion of intervention to a graph or variable set[[7]](#footnote-7). For the reason just described I do not think such a response works. In any case, it is not the response that I will adopt.

The result of intervening on a variable *X* can be represented by breaking all other arrows directed into *X*, so that the *X* is caused only by the intervention *I*[[8]](#footnote-8). All other arrows not directed into *X* are preserved. For example, if *C* is a common cause of *X* and *Y*, as in Figure 3 with no causal relationship between *X* and *Y*, the result of intervening on *X* is to replace

*X*

*C*

*Y*

Figure 3

**Fig 3**: *C* is a common cause of *X* and *Y.*

with figure 4:

*I X , C Y*

Figure 4

**Fig 4**: The result of intervening on *X* in Figure 3

This “arrow-breaking” understanding of interventions plays an important role in reasoning with causal graphs: when *I* breaks the arrows directed into *X* and meets the other conditions for an intervention, *X* is given an independent unconfounded causal history. This allows us to infer that when changes in *X* that are caused in this way and are associated with changes in a second variable *Y*, it must be the case that *X* causes *Y.* A standard, if sometimes tacit assumption in the causal modeling literature and in interventionist interpretations of causal modeling is that causal relationships between variables can always be broken in this way—as noted in Section 3 this can be thought of as a consequence of a condition called Independent Fixability (IF) in Woodward, 2015. By contrast, this is not true for supervenience relations – they are unbreakable. This is one of several reasons why mixed graphs should not be assimilated to causal graphs.

We can define direct causation and provide an interpretation for the arrows in a causal graph as follows. (This is a slight modification of Woodward, 2003, the modification being that here **DC** is defined in a way that assumes determinism -- I’m following the literature on exclusion in assuming the context is one in which causal relations are deterministic .)

(**DC**) *X* is a direct cause of *Y* with respect to variable set **V** iff there are possible interventions on *X* that will the change the value of *Y* when all other variables in **V** are held fixed at some value by interventions.

Direct causation is thus defined relative to a variable set but other causal notions, such as the notion of a total cause or a contributing cause along a route that are not relative to a variable set can be characterized in terms of (**DC**) and additional conditions (Woodward, 2003). I also note, for future reference, that as explained in Woodward, 2003, pp. 41ff, 70) the condition in **DC** that there are possible interventions on *X* that change the value of *Y* should be interpreted in the following way: there are one or more values *xi* of *X* such that interventions that set *X* to those values are always followed by the same associated value of *Y* when those interventions are repeated. In other words, for some values *xi* of *X* there exists a *function* *f* from those values to associated values of *Y*, *yj*= *f*(*xi*) which describes how for these interventions on *X*, *Y* responds. It is important to understand that this does *not* require that the function *f* holds for all values of *X—*the reference to some possible interventions in **DC** is meant to be interpreted as the claim that for some but not necessarily all interventions setting values of *X*, *Y* responds to those interventions in a uniform way. For example, if *X* is the extension of a spring and *Y* the restoring force it exerts, with *Y=-kX*, the requirement is that for some values of *X*—e.g., *x1*, ..*xn* , *Y = -k x1* whenever *X* is set to *x1* via an intervention, *Y= -kx2* whenever *X*= *x2, Y= -k x2* and so on. This is compatible with *Y* responding in some other way that is not correctly described by *Y=-kX* to interventions that set *X* to values different from *x1*, ..*xn*. For example, the spring will respond in a non-linear fashion if you extend it too much.

This "uniformity" requirement is intended to ensure that for *X* to cause *Y*, interventions on *X* with respect to *Y* must be “unambiguous” for some values of *X* in the sense that *Y* does not respond differently or non-uniformly to interventions on different occasions that set *X* to the values in question —instead there must be a uniform or stable response to interventions that set *X* to the same value. The importance of this consideration (and some further motivation for it) will become clearer below, where we will see that it plays a central role in the characterization of upper-level causation.

We are now in a position to describe in somewhat more detail one version of the exclusion argument alluded to earlier. Following Baumgartner, 2010 and focusing on Figure 1 note that the interventionist test for whether *M1* causes *M2* (or *P2*) is whether some intervention on *M1* would change these variables. Baumgartner contends that in Figure 1 when one attempts to intervene on *M1*, *P1* must be treated as an “off-path” variable with respect to *P2* or *M2* . Thus in intervening on *M1* to determine whether it is a cause of *P2* or *M2*, we would need to change the value of *M1* while controlling or holding fixed the value of *P1*. However, this is impossible given the supervenience relation between *M1* and *P1*. Hence, one must conclude that *M1* is causally inert[[9]](#footnote-9). To use Baumgartner’s language, any manipulation of *M1* is inevitably “confounded” by *P1*, or “fat-handed”[[10]](#footnote-10), undercutting any conclusion about the causal efficacy of *M1*.

As I see it, this line of argument rests on the claim that off-path variables in a *mixed* graph should be treated the same way as off-path variables in a *causal* graph. That is, it is claimed that just as it is appropriate to control for off-path variables in a causal graph in assessing whether some variable *X* causes another variable *Y*, we should also treat *P1* as an off-path variable (a potential “confounder”) that needs to be controlled for in assessing whether *M1* causes *M2* (or *P2*) in a mixed graph like that pictured in Figure 1. Put differently, it is assumed that confounding by the presence of a supervenience base has the same significance for the presence of causal relationships as ordinary confounding by alternative causes when no supervenience relations are present.

The contrary view, defended in Woodward, 2015 and which I endorse, holds that we should not treat variables like *P1* in Figure 1 as potential confounders when there are interventions on variables like *M1*. As explained in Woodward, 2015, this means that we do not interpret the requirement that an intervention *I* on *X* with respect to *Y* must not affect *Y* via causal paths that do not go through *X* as applying to paths that involve the supervenience bases of *X* and *Y*. In other words the “no influence via off-path variables” requirement is interpreted as not applying to the paths or relations associated with the supervenience bases of *X* and *Y*. To explore how this might work, we first need to be more precise about how non-reductive supervenience relations should be characterized.

1. **Interventions and Supervenience Relations**

Unfortunately, the usual treatments of non-reductive supervenience relations in the philosophical literature are unperspicuous for a number of reasons. For example, they make it difficult to capture cases of multiple realizability involving arithmetical relations among values of variables[[11]](#footnote-11). The framework that follows is deliberately very simplified but I think it is nonetheless adequate for our purposes[[12]](#footnote-12) .

Suppose that we have two sets of variables {*Ui*} *i*= 1, 2, … (upper-level variables) and {*Lj*} *j*= 1, 2 …. ( lower-level variables). Possible values of each variable are represented by indexed lower cases letters: the possible values of *Ui* are *uik*—*u11,* *u12*  and so on and the possible values of *Lj* are *ljm*. The values of the *Ui* supervene on the values of the *Pj* and this involves “multiple realization” of the following sort: for each *Ui* there is a many-to-one surjective function *f* that maps a number of different values of the *Lj* into each value of *Ui*. (*f* may map values from different *Lj* into values of a *Ui* or alternatively different values from the same *Lj*  may be mapped into a value of *Ui.* )We require that *f* be a function because we want to exclude the possibility that same value of *Lj* is mapped into different values of *Ui.* (This is contrary to the assumption of supervenience.) We require that this function be surjective to capture the standard assumption that every value of each of the *Ui*  is realized by some value (typically many values) of the *Lj*s. Multiple realizability is captured by the many-to one-character of the function—that is, we assume that the function is *not* bijective.

Note also that multiple realizability is characterized as a relation between *values* of variables—I don’t speak of “properties or “kinds” being “multiply realized” by other properties or kinds since this framing does not naturally capture a number of cases of multiple realizability. As an illustration discussed below, suppose total cholesterol (*TC*) is the sum of *HDL* and *LDL* cholesterol, with single values *tc* of *TC* being multiply realized by pairs of values *hdl, ldl* of *HDL* and *LDL* that sum to *tc*. This is not a matter of some property associated with *HDL* (e.g., the property of having *HDL= hdl* ) by itself realizing *TC*[[13]](#footnote-13)*.* In general talk of properties or kinds realizing others seems (at best) to fit “binary” variables[[14]](#footnote-14), and does not capture lower to upper relations among variables that have a more complex quantitative structure and where these relations are represented by more complex mathematical operations like addition, averaging, integrating and so on.

An additional motivation for thinking of multiple realizability as having to do with the relation between values of upper and lower-level variables is that this seems necessary to capture one of the most important reasons for employing upper-level variables in causal claims—that these involve a reduction in dimensionality or degrees of freedom in comparison with lower- level variables. For example, the use of upper-level thermodynamic variables like temperature and pressure allows us to use a very small number of such variables in place of much higher dimensional information about the positions and momenta of the huge number of individual molecules comprising the gas. Note that this feature is lost (or at least not represented) in a diagram like Kim’s (Figure 1)—there is nothing about the diagram that represents whether the upper-level variables *M* involve a reduction in number or dimension or a “coarsening” of the lower-level variables.

We may think of the values of *Lj*s that are mapped into the same value of *Ui* as belonging to the same equivalence class; *f* thus induces a partition of the values of *Lj* into disjoint equivalence classes each of which corresponds to a single value of *Ui*. As is standard in discussions of supervenience, this functional relation is understood as an “unbreakable” constraint relation rather than a causal relationship[[15]](#footnote-15). Within an interventionist framework, a natural way of capturing the presence of such a non-causal constraint relationship is in terms of the notion of independent fixability (**IF**)[[16]](#footnote-16). When all relationships are causal, all possible combinations of values of distinct variables are permissible states in the state space or phase space of such variables. This corresponds to the standard assumption that interventions that set the values of variables to any value within their range are always possible, independently of the values taken by other variables. This in turn supports the standard assumption in causal modeling that causal relations are always “breakable” in principle in the sense that if *C* causes *E*, there is a possible intervention on *E* that will “break” the arrow from *C* to *E* and that such an intervention can set *E* to any value within its range. By contrast, when non-causal relations are present, as is the case with supervenience relations, certain combinations of values of variables are precluded, not for causal reasons but for reasons of some other sort—logical, conceptual or if you like, “metaphysical” reasons. For example, when *M* supervenes on *P*, it is not possible in this broader, non-causal sense, for *M* and *P* to take combinations of values such that the value taken by *M* is not in the supervenience base for *P* for that value.

As a working illustration of this picture and how one should think about interventions in such a context, assume that the target upper-level variable is the temperature *T* of a dilute gas and the lower level realizing variable *K*  describes the kinetic energy of each of its component molecules. That is, the values of *K* are n-tuples, each of which specifies such and such a kinetic energy for molecule 1, a kinetic energy for molecule 2 and so on. A different value of *K* specifies a different possible n-tuple of kinetic energies for each of the individual molecules. A given value *t* of the upper-level variable temperature *T* thus can be realized by a very large number of different values of *K*—all of the values that correspond to the same value of the average kinetic energy for the gas. In this sense the value *T=t* is multiply realizable by different values of *K*.



As noted above, advocates of exclusion arguments hold that interventions on *T* are impossible because “confounded” by the lower-level variable *P*. Let us bracket this claim for the moment and consider what an intervention on *T* might involve if such a thing were possible—that is, whether there is a coherent way of thinking about interventions on *T* in the situation under discussion. To explore this, consider the following operation that might naturally be regarded as an intervention on *T* if anything is: we set the temperature of the gas (in a container of course) to the value *T= t* by placing it within a heat bath. (The temperature of the heat bath might be determined exogenously by some random process.) When we do this, the value *T=t* (at an instant or some very short time interval) will be realized by some particular value of *K=k1* that is consistent with *T=t*. The intervention causes both *T=t* and *K*= *k1* but it is natural to think the intervention as “controlling” the value of *K* but not as controlling in the same way which value of *K* realizes *T= t*.[[17]](#footnote-17) What I mean by this is that the experimenter has a way of intervening (via the heat bath) on the temperature to reliably fix it to the same value on repeated occasions but no analogous operation available for reliably fixing *K* to the same value on different occasions. The latter is outside the experimenter’s control in the set-up we have envisioned. The value of *K* that realizes *T* could be any value of *K* that is consistent with *T=t*. Indeed, an instant later, the realizer of *T=t* will almost certainly be some different value of *K*, corresponding to some different combination of kinetic energies of the molecules.

Note that in carrying out this intervention on *T* we do not, so to speak, have to do two different things or to introduce two different, distinct causal relationships, one of which corresponds to the setting of *T= t* and other of which corresponds to setting *K=k1*. The same single intervention that fixes the value of *T* also ensures the presence of some realizer for that value in *K*. Because of this, as I will explain in more detail below, it does not seem appropriate to think of the intervention as operating on *T* and *K* via two causally independent paths, one affecting *T* and the other distinct path affecting *K —*i.e., the intervention is not a “common cause” of *T* and *K* as this expression is ordinarily understood

Assuming that interventions on *T* are possible, it is clear that they must satisfy certain consistency constraints with respect to interventions on the realizers of *T* in *K*. In particular, interventions on *T* and on *K* must not involve impossible violations of the supervenience relations specified by the relation *f* between *K* and *T.* For example, if *T=t* and *K=k1*  is the realizer of *T=t* that happens to be present at some particular time or occasion, we can’t intervene to change *T=t* to some different value *t’* while it remains true that *K=k1* or while we intervene to set *K= k1 .* Similarly we can’t intervene to set *T=t* while simultaneously intervening to set *K* to some value that is not in the equivalence class of realizers of *T=t.* Of course one might introduce a notion of intervention according to which interventions that change the values on upper-level variables like *T* leave the values of their supervenience bases unchanged. It would then immediately follow that such upper-level interventions are always impossible but the obvious question is why one would want to do that if an alternative characterization of upper-level interventions that does not automatically make them impossible is also available.

There are other constraints relating upper to lower-level variables that must be satisfied if upper-level causal relations are to be well-defined or appropriately behaved, given the facts about their realizers. We noted earlier that within an interventionist framework for *X* to cause *Y* interventions on *X* must result in a “uniform” or “unambiguous” response in *Y*. The presence of multiple realizability makes apparent one way in which this condition might fail. Consider the example mentioned earlier: total cholesterol *TC* is defined as the sum of *HDL* and *LDL* cholesterol. *HDL* cholesterol has a beneficial effect on heart health *H* while *LDL* cholesterol has a deleterious effect. Thus *TC* has a non-uniform or ambiguous effect on *H* depending on the precise mix of values of *HDL/LDL* by which a value of *TC* is realized. Suppose that this is true for all values of *TC*. We might say that a manipulation of an ambiguous variable such as *TC* with respect to *H* does not count as an intervention with the consequence that we cannot intervene on *TC*  and that it cannot cause *H* for this reason. Alternatively we can count this as a bona-fide intervention but deny that *TC* causes heart health because it has ambiguous effects on the latter. The upshot is the same under both alternatives (*TC* does not cause *H* in either case) but I will adopt the latter since it seems more natural.

Note how this contrasts with the manipulation of the temperature of a gas considered earlier. This manipulation will have a uniform/ unambiguous effect on such other thermodynamic variables such as pressure for almost all (all except a set of measure zero of) molecular realizations of the temperature variable. One of the intuitions we want to capture is that a well-behaved upper-level causal variable (with respect to some target effect) should exhibit this sort of behavior under interventions.

Putting all of this together suggests the following characterization of when the relationship between interventions on lower and upper-level variables and the causal relations in which they figure is consistent or well-behaved. (This roughly follows, with some emendations, the characterizations in Chalupa et al., 2017 and Rubenstein et al., 2017-- the latter is discussed in more detail below.) As before let *Ui* be the upper-level variables and *Lj* be variables the values of which realize values of *Ui* with *f* a many-to-one surjective function relating values of *Lj* to values of *Ui* .The model involving the lower-level variables is assumed to be causally correct in the sense that it correctly describes the response of the lower-level variables to interventions, understood in the usual way. Then define an upper-level intervention that sets that sets *Ui* =*ui* as any intervention that sets the value of *Lj* to any one of the realizers of *ui* in the equivalence class of its realizers. (Thus such an intervention fixes values for both the upper and lower-level variables in a way that is consistent with the supervenience relations.) Call this condition INTERVENTION.

We also require, for reasons noted above, that for an upper-level variable *Ui* to cause another variable *Y* (whether upper or lower-level) there must be values of *Ui* such that all interventions (understood as above) that set *Ui* to those values are associated with the same uniform response in *Y*. This NON-AMBIGUITY condition corresponds to what Woodward (2008) calls *realization-independence*: it is the requirement that for *Ui* to cause *Y* there must be values of *Ui* such that interventions that set *Ui* to those values are associated with the same response in *Y* regardless of how those values of *Ui* are realized at the lower level. This requirement fails for the relationship between *TC* and heart health, because for any value of *TC*, there will be different lower-level realizations, corresponding to different pairs of values of *HDL* and *LDL* which have non-uniform effects on heart health.

INTERVENTION makes it clear that the intervention that sets the realizer of *ui* to some value is at the same time an intervention (the same intervention) that sets the value of *ui* itself. It also follows immediately that there is no possible combination of interventions that sets *Ui* = *ui*  and also sets *Lj* to a value that is not among the realizers of *Ui* = *ui .* In general combinations of upper and lower-level interventions that are “inconsistent” are excluded by this characterization. This also provides one (of several possible) motivations for *not* requiring that for an upper-level variable *U1* to cause another upper-level variable *U2* it must be possible to intervene on *U1* to change its value from, say *u11* to *u12* , while at the same time holding fixed the realizer for *u11*. From the perspective of INTERVENTION requiring this amounts to requiring that the same intervention on the realizer of *U1* do two inconsistent things—both change the value of the realizer in such a way that the value of *U1* is changed from *u11* to *u12* and not change the value of this realizer. A natural thought is that the impossibility of doing something that is logically inconsistent does not show that *U1* is causally inert.

As several writers note (e.g., Rubenstein et al., 2017, see also Ellis, 2016 ) the INTERVENTION and NON-AMBIGUITY requirements, in conjunction with some natural additional assumptions, imply the following “commutivity” requirement concerning the causal relations involving upper and lower-level variables:

(COMM) Suppose F is a function describing the causal relation between the lower-level variables *L1* and *L2,* g1 describes the realizing relation that maps *L1*to *U1,*  g2the realizing relation between *L2* and *U2* and that H describes the putative upper-level causal relation between *U1* and *U2.* Then for *U1* to cause *U2* (with causation understood in the usual interventionist way) and for consistency across levels, the response of *U2* (say *U2*=*u21*) to performing an intervention setting *U1*= *u11* (again understood in terms of the INTERVENTION condition) should be the same as performing an intervention on any one of the realizers of *u11*, using F to give the resulting value of *L2* and then coarse- graining this value of *L2* via g2 to yield the value for *U2* which should be *u21*.

Woodward 2015 provided a characterization of interventions in contexts involving supervenience relations which he called **IV\*** and which is closely related to INTERVENTION. Adapted to the present context, an **IV\*** intervention *I* that sets an upper-level variable *U1* to value *u11* with respect to a second variable *Y* is understood as the performance of an operation in which *U1* = *u11* is realized by some value of a lower -level variable or variables within the equivalence class of realizers of *u11* and which conforms to conditions **IV1**, **2** and **4** for an intervention in Woodward, 2003 but adds the following modification (**IV3\***) of his condition **IV3:**

(**IV3\***) *I* affects *Y*, if at all, only via a route that goes through *U1* or only via a route involving the supervenience bases of *U1* and *Y.*

In other words the usual requirements in the definition of an intervention that *I* not affect *Y* via variables on paths that do not go through *U1* are not taken to apply to variables related to *U1* and *Y* via supervenience relations, although they apply, as before, to all other variables. As explained above, once we agree that an intervention on an upper- level variable also change values in its supervenience base (so that we can’t think of the intervention as changing the former while leaving the latter unchanged) we also need to exclude any requirement that the supervenience base for *U1* be treated as an ordinary confounding causes which is what **IV3\*** attempts to do[[18]](#footnote-18).

Once we understand interventions on upper level variables in the way described above it follows that interventions on upper-level variables are possible and if, under an intervention on such a variable *U* , a second variable *Y* changes, we can (on the usual interventionist grounds) conclude that *U* causes *Y*.

1. **Justification for Not Controlling for Supervenience Bases**

The discussion in previous sections shows that one can consistently describe frameworks for charactering interventions and causation in multi-level systems in which interventions on upper-level variables are possible. However the crucial question is not whether compatibilist frameworks having this feature are possible or consistent but whether there are good *arguments* or *rationales* that favor adoption of such frameworks over the alternatives with exclusionist implications. On my view there are such arguments for compatibilism. These have to do, among other considerations, with the goals or purposes that underlie causal reasoning and more specifically with what these goals imply about what should count as a confounder and why it is important to control for confounders—that is, what such control achieves.

One argument for compatibilism, developed in more detail in Woodward, 2015 is this: Consider an ordinary case—call it case 1-- of confounding in which no supervenience relations are present (or in which, for one reason or another, it is not thought relevant to model these) and in which the confounding is due to omitted causal factors. For example, suppose an association between *X* and *Y* is present which is due to some additional variable(s) *Z*, with no causal relation holding between *X* and *Y,*  as when *Z* is a common cause of *X* and *Y*. As a result, when we intervene on *X*, the association between *X* and *Y* disappears, and *Y* does not change. Thus by failing to control for Z, we are misled about whether manipulating *X* is a way of changing *Y*. By contrast, when *L* is the supervenience base of *X*, the association between *X* and *L* is unbreakable, with *L* automatically changing under manipulations of *X* in whatever way is required by the supervenience relation. Thus we don’t have to worry that failure to control for *L* will have the result that the association between *X* and *Y* will disappear when we manipulate *X*. Failing to control for or hold fixed the supervenience bases of *X* and *Y* will not result in the kind of error that results when we fail to control for Z in case 1.

As an illustration, suppose that we want to know whether aspirin ingestion *A* can cause a decrease in headache incidence *H.* We notice that *A* supervenes on a complicated set of physical facts *F* having to do with the chemical structure of aspirin, its introduction into a human body with a certain physiology and so on, with the latter varying somewhat from individual to individual so there is multiple realizability. Suppose that we manipulate *A* in a randomized controlled trial (so that the manipulation is unconfounded in any ordinary sense) without controlling for *F* and observe a corresponding change in *H*. This does not mean that this association misleads us about whether we can use *A* to change *H*. Failing to control for *F* does not lead to the kind of mistake that we make in ordinary cases of causal confounding, such as a case in which we give aspirin preferentially to subjects who are most likely to recover from headaches independently of whether they receive aspirin..

By contrast, if we do control for *F,* it will follow automatically and independently of the results of any experiment that *A* must be causally inert. This is so even if the results of the randomized experiment above are that there is a much higher incidence of recovery in the treatment group than in the control group. As this example illustrates, what we often care about in causal inference is the (apparent) effects of manipulation of upper - level variables. (These are often the variables that we can most readily manipulate and the lower-level variables on which they supervene are often unknown to us or not readily manipulable). Thus a requirement to control for supervenience bases in the context we are assuming (supervenience but the absence of type identities) would deprive causal thinking of much of its usefulness: it would deprive us of the ability to distinguish between ordinary causal confounding and (supposed) confounding by a variable’s supervenience base, which is not ordinarily regarded as confounding at all. Unless such a move is somehow required, there seems no reason to adopt it, when alternatives that do not have these limitations are available.

This is admittedly a “pragmatic” argument in the broad sense of that word according to which pragmatic considerations are those bearing on the effective achievement of our goals. But interventionism *is* a pragmatic (or “functional”) theory in this sense. It associates the discovery of causal relationships with the goal of discovering relationships that can be used for manipulation and control. If that is the goal, it seems self-defeating to control for supervenience bases in the way defenders of the exclusion argument advocate.

To this we may add the following consideration which is also broadly pragmatic: As we have seen, it is generally thought to that an interesting and important problem in causal modeling and in science more generally has to do with finding conditions under which an upper-level theory retains some portion of the causal truths that hold in a lower level theory when the variables of the upper-level theory involve a coarse-graining of the lower-level theory, and where the coarse-graining satisfies assumptions about multiple realizability and supervenience like those described above. As we have noted, a number of recent papers contain strategies for addressing this problem and their utility seems obvious. But if the variables of the upper-level coarse grained theory must automatically be regarded as causally inert, given the assumptions described above, it seems to follow that all of this work is otiose and, indeed, based on a conceptual confusion of some kind. If (as I have argued) there are ways of avoiding this conclusion, it seems worthwhile to do so[[19]](#footnote-19).

The considerations just described would carry less weight if there were other plausible theories of causation for which exclusion problems did not arise even if one controls for supervenience bases. However, although critics sometimes suggest that interventionism is uniquely susceptible to exclusion problems, this is far from the case. The basic observation here is that *all* plausible theories of causation need some strategy for controlling for confounders since failure to do so undermines causal claims. Different theories may disagree about what counts as confounder and how to appropriately control for them but whether one controls for confounders is not optional. For example, a simple version of a probabilistic theory might attempt to accomplish such control by conditioning on appropriate other variables—e. g.:

(4.1) *C* causes *E* iff Pr*(E/ C. Ki)* > *Pr (E/not C. Ki)* for all *Ki* where the *Ki* are possible combinations of confounding variables.

A naïve regression-based test for whether *X1* causes *Y* might involve regressing *Y* on *X1* and all other variables *X2*…*Xn* thought to be possible confounders and then determining whether *X1* has a non-zero coefficient[[20]](#footnote-20).

Suppose that in the case of the probabilistic theory *C* has values *ci* , for *i = 1.., n* and we take the *Ki* to include the full supervenience base for *C*, with each *Ki*  taking values *kij*, *j = 1,..m* . It then follows that Pr (*E*/ *ci.*  *kij*) = Pr *E*/ *kij*) for each *kij* that realizes *c i* and that Pr *E*/ *ci.*  *kij*) is undefined for those *kij* that are inconsistent with the realization *ci* since in this case Pr (*ci.*  *kij)*  = 0. Thus, conditioning on the supervenience base for *C,* each value *ci* of *C* is either irrelevant to *E* or the causal relation between that value and *E* is undefined. It thus follows that *C* is causally inert with respect to *E* when causation is assessed according to (4.1). It is also worth noting in this connection that it is usually (if tacitly) assumed in probabilistic theories of causation that all possible combinations of values of the variables) have positive support. Again, this will not be satisfied if we include the supervenience bases of *C* as possible confounders—another indication of how badly treating supervenience bases as confounders fits with this standard framework.

Similar conclusions follow on the regression-based test for causation: if we include variables representing the full supervenience base for *X1* in the regression equation their inclusion will automatically make the upper-level variables appear causally inert. Of course if one is convinced, antecedently, that this exclusionist conclusion is correct, the fact the conclusion follows for pretty much any account of causation which involves control for confounders and the assumption that supervenience bases are confounders may seem to simply strengthen the exclusionist conclusion. On the other hand, the above observations do show that exclusionist worries are not specifically a problem for interventionism. Moreover, in showing that exclusionist conclusions follow generically once it is assumed that one should treat supervenience bases as confounders, they focus attention on whether that assumption is justified. In other words, the crucial question becomes: what should be controlled for and why. If one holds that supervenience bases must be controlled for one needs to provide an argument for why this is appropriate rather than treating it as an inevitable default.

In this connection there appears to be a tendency to assume that interventionism is committed to a quite general requirement to control or “fix” variables in elucidating causal relations. For example, Zhong, 2020 writes:

many argue that interventionist supervenient causation is exempted from the fixability condition [ that is, the requirement to hold fixed supervenience bases—JW] . However, this approach looks ad hoc, inconsistent with the general interventionist requirement on fixation.

In this passage Zhong claims there is a general prima-facie interventionist requirement to hold variables fixed, and it is ad hoc to exempt supervenience bases from this requirement[[21]](#footnote-21).

I disagree that interventionism or any other plausible theory of causation should be committed to any such general requirement. Instead what one should “control for” depends on the structure under investigation, the causal question one is trying to answer as well as what is meant for by “control”. As a trivial illustration, if we are interested in whether *X* causes *Z* and *X* causes *Y* which causes *Z*, we should not control for *Y* in any sense of “control for”. In a “triangular” structure in which *X* causes *Y* which causes *Z* and *X* also causes *Z* by direct route, we should not control for *Y* if we want to capture the total or overall effect of *X* on Z but we should control for Z if we want to capture the direct effect on *X* on Z. (In other words, what we should control for depends on whether the causal question concerns an overall effect or a direct effect.) In a “collider” structure in which *X* causes *Z* and *Y* causes *Z*, with no causal connection between *X* and *Y*, no harm is done if we hold fixed Z via an intervention and then intervene on *X* to see whether there is a response in *Y,* using this to establish whether *X* causes *Y* . However, if, in this case, we hold *Z* fixed in the sense of conditioning on it, it is well known that that this will create a conditional dependence between *X* and *Y* and thus a misleading inference about whether there is a causal relation between *X* and *Y.* If we are making an actual cause judgment, according to many current theories, we are allowed, for any path between the candidate cause *c* and effect *e*, to fix off-path variables to any values that do not change the value of the effect or the value of any variables on that path (other fixings of off path variable values are not allowed) in determining whether an intervention on *c* will change *e* and hence whether *c* causes *e* [[22]](#footnote-22). This would not be an appropriate thing to do in answering other sorts of causal questions. In still other cases, the question of which variables one should “fix” or “control” for becomes even more subtle. As these examples illustrate, following a general requirement to control or fix variables willy-nilly can easily lead to mistaken causal inferences. There is thus no general argument from the need to “control for things in causal inference” that leads to the conclusion that one should control for supervenience bases—if there is such an argument it needs to focus specifically on what would be accomplished by controlling for supervenience bases (what question this would answer and why this is an appropriate way of answering the question). .

**5. A New Argument that Interventionism Implies Causal Exclusion**

In a recent paper (2018) , Baumgartner advances an argument for “causal underdetermination” when supervenience relations are present that differs somewhat from argument described in Section 2 but is broadly similar in spirit. It is worthwhile to consider this argument separately since it doesn’t just assume that it is appropriate to control for supervenience bases but instead presents an argument that we are required to do so within an interventionist framework. Moreover, Baumgartner claims that what follows from this argument is not that upper-level variables are inert but rather that it is underdetermined by all possible empirical evidence whether or not this is true.

Consider again Kim’s diagram (Figure 1) and suppose there is an intervention *I* on *M1* with supervenience base *P1*. Baumgartner considers two possibilities for how such an intervention should be represented . The first is that *I* causes *M1* and *P1* along one causal path—e.g., *I*🡪 *M1*🡪 *P1* or *I*🡪 *P1*🡪 *M1*. Baumgartner argues (entirely correctly) that this possibility is excluded because the relationship between *M1* and *P1* is not a causal relationship. The second possibility, represented by Figure 5, is that we should regard *I* as a common cause of *M1* and *P1*:

*M1 🡨 I🡪 P1*

Figure 5

**Fig 5**: *I* is a common cause of *M1* and *P1*

According to Baumgartner these are the only two possibilities and since the first is inadequate, we must adopt the common cause representation in Figure 5. Given this understanding of what an intervention on *M1* involves, Baumgartner argues (similarly to the argument in Section 2) that *I* is a “confounded” manipulation of *M1* with respect to *M2* (or *P2*) in Kim’s diagram. It is confounded because in that diagram there is an alternative route from *I* through *P1* to *P2* or *M2* which does not go through *M1*. As a consequence, if *M2* or *P2* changes (even uniformly) under this manipulation, we are not in a position to determine whether this is because (i) *M1* causes these variables or, alternatively, because (ii) *M1* is causally inert with all of the causation instead going through the *I*🡪 *P1* 🡪 *P2* ⇒ *M2* route (where the last thicker arrow represents supervenience). Thus, according to Baumgartner, there is no empirical basis for deciding between (i) and (ii). Baumgartner also suggests, however, that there may be a non-empirical basis for distinguishing between (i) and (ii)—I will take up this suggestion in Section 8.

**6. Response**

The argument described turns on the claim that an intervention *I* on *M1* (with supervenience base *P1*) must be represented as in Figure 5 with *I* as a “common cause” of *M1* and *P1* which has the consequence that intervention-based inferences involving *M1* are inevitably “confounded” by *P1*. There are two issues here. One is whether it is correct that, given the inadequacy of the single path treatment of interventions, the only alternative is the common cause account represented by Figure 5. I will take up this issue below and here just remind the reader that the treatment of interventions in Section 3 represents such an alternative to this common cause representation . The second, prior issue is whether thinking about interventions in terms of the common cause representation in Figure 5 is acceptable. If it is, the fact that there are alternatives may not matter. In fact, however, there are independent reasons for not thinking of interventions as common causes of upper-level variables and their supervenience bases—reasons in addition to the fact this leads to exclusionist conclusions that, at least from my point of view, are unwelcome.

To begin with, Figure 5 with *M1* and *P1* in a supervenience relation behaves very differently from structures ordinarily described as involving common causes. As noted in Section **2**, in ordinary common cause structures like Figure 3, one can intervene on each of the joint effects individually to *break* the arrow from the common cause *C* into that effect while leaving the arrow into the other effect undisturbed—this is an implication of the condition called **IF** there[[23]](#footnote-23). By contrast in the common cause representation in Figure 5, any breaking of the arrow from *I* to *M1* (say by means of a second intervention *I\** on *M1* ) must, because of the supervenience relation, also break the (distinct) arrow from *I* to *P1*, changing the value of *P1* appropriately and putting *P1* under the control of *I\**, as represented in Figure 6

*I*\*

*M1*

*I*

*P1*

Figure 6.

**Fig 6**: Suppose *I* was previously a common cause manipulation of *M1* and *P1* asrepresented in Figure 5. Intervention *I\** on *M1* breaks the arrow from *I* to *M1* but given the nature of the supervenience relation, *I\** must also break the arrow from *I* to *P1* (again assuming the correctness of the common cause representation). Presumably, again assuming the correctness of the original representation in Figure 5, this is also accompanied by the creation of a new arrow from *I\** into *P1* so that *I\** is now a new common cause of *M1* and *P1*.

This is a violation of the standard requirement in the interpretation of causal DAGs that interventions on a variable break only arrows directed into the variable intervened on, leaving all other arrows intact and not creating any new, additional arrows.

Put differently in a structure like Figure 5 with *M1* supervening on *P1* the possibility of a certain pattern of independent interventions on each of the joint effects *M1* and *P1*  is ruled out: one can’t intervene on *M1* without also intervening on *P1.* This feature (possibility of independent interventions) is characteristic of ordinary common cause structures without supervenience relations (and for that matter, other causal structures without supervenience relations). Relatedly, whatever the arrows drawn from *I* to *M1* and *P1* mean in Figure 5, they cannot be interpreted in terms of **DC** or along other standard interventionist lines. Interpreting the arrow from *I* to *M1* in Figure 5 in terms of **DC** requires that changing *I* changes *M1* when all other variables in the graph are held fixed by independent interventions, but (once again) it is impossible for *M1* to change when *P1* is fixed. In the absence of some alternative account of what the arrows in Figure 5 mean we thus have a violation of what I called the interpretation constraint (Section 2). Of course one can *draw* separate arrows from *I* into *M1* and *P1*  but this by itself is unhelpful if one doesn’t provide an interpretation what those arrows mean.

As noted earlier, the plausibility of Baumgartner’s argument that interventions in the context of supervenience relations should be thought of in terms of a common cause structure like Figure 5 rests on the assumption that this is the only alternative to a representation in terms of chain structure which is obviously inadequate. But this provides no reason to adopt the representation in Figure 5 if, as I have argued, it is also inadequate. Instead the conclusion we should draw is that neither the common cause structure nor the chain structure is acceptable. This conclusion should not be surprising. Non-causal supervenience relations behave differently from causal relationships and capturing the former and how they interact with the latter requires an expanded framework-- representational tools and accompanying interpretative assumptions -- that goes beyond the directed graph framework used to represent causal relationships alone[[24]](#footnote-24).

Section 3 presented one version of such an expanded framework for thinking about interventions when supervenience relations are present. According to this framework the intervention *I* above does not have two independent effects, one on *P1* and one on *M1* (as proposed in Figure 5)). Instead *I* has a single effect on both *M1* and *P1*—the intervention *I* that changes *M1* is the same intervention as the intervention that changes *P1* and (assuming that *I* is a genuine intervention) it obeys the requirements for an **IV\***-intervention. Woodward (2015) represented this interpretation of what an intervention does by means of a bracket, as in Figure 7:

*M1*

*I* {

*P1*

Figure 7

**Fig 7**: The bracket represents the fact that the intervention *I* has a single effect on both *M1* and *P1*.

The idea here is *not* that the issues having to do with the behavior of interventions (and what needs to be controlled for) when supervenience relations are present can all be resolved just by introducing a new symbol (or that exclusionist conclusions can be avoided just by introducing a new symbol). As with other representational devices what matters is the interpretation given to this symbol and the rules governing its use and whether those rules are well-motivated. The point of the bracket is just to capture or represent the idea that we should think of the intervention represented by this symbol as operating in the way described above-- as not having two independent effects[[25]](#footnote-25) in the way in which an ordinary common cause does and as obeying the requirements for an **IV\*** intervention.

That said, there is an alternative way of capturing the role of interventions, developed in Rubenstein et al. 2017, which builds on ideas described in Section 3. This is in some respects less of a departure from the directed graph framework. In this approach, causal relations involving upper-level variables and those involving their supervenience bases are not represented in a single graph. Rather there are two ordinary directed graphs, one representing the upper-level causal relations and one representing the lower-level relations. The relationship between the two graphs is encoded in the manner described in Section 3 by means of transformations specifying the relationships between the variables in the graphs and transformations specifying the relationships between interventions on the variables in the graphs. That is, there are transformations (meeting certain conditions) specifying how interventions on lower-level/supervenience base variables *L* “correspond” to interventions on upper level variables *U*, where “correspond” means that carrying out the lower level intervention just amounts to or is tantamount to carrying out the upper-level intervention[[26]](#footnote-26). (Thus again additional structure—in this case in the form of these transformations—is introduced to deal with supervenience relations, but there is no need to employ Woodward’s bracket notation or his accompanying ideas about how to understand interventions when variables and their supervenience bases are represented in a single graph.) Because this approach does not employ a single graph in which both *U* and *L* are represented, it avoids the common cause interpretation (5.1) of how interventions operate in the context of supervenience relations and the conclusions about confounding and causal exclusion that Baumgartner and others draw from these. I see (i) the bracket interpretation of interventions in Woodward (2015) and the accompanying assumptions about not controlling for supervenience bases when one intervenes on upper-level variables and (ii) the treatment just described in terms of transformations between graphs as entirely compatible—they are motivated by very similar underlying ideas and have the same consequences for the exclusion argument.

Before leaving this section, let me emphasize again that neither of the treatments of interventions just described relativize the notion of an intervention to a graph or to a variable set. The **IV\*** intervention notion described in Woodward, 2015 requires that an intervention that changes the value of an upper-level variable also change the value of some variable in its supervenience base. It also requires that we do not think of the latter variable as a potential confounder but in other respects interventions behave as they do (and are characterized in the same way as) when supervenience relations are not present. In particular we still require that an intervention *I* on *X* with respect to *Y* not affect *Y* via a causal path that does not go through *X*, that *I* not be correlated with variables that affect *Y* through causal paths that do not go through *X* (and not merely that we do not know of or represent such paths), although for reasons explained above we do not treat supervenience bases of *X* and *Y* as violating this requirement. In the case of a representation like Rubenstein’s, although two separate graphs are employed, the notion of an intervention is not relativized to these graphs. At the lower, ground truth level, interventions are understood as satisfying the usual non-relativized requirements and the transformations employed carry this non-relativity forward to interventions on the upper-level variables.

**7. Baumgartner on Underdetermination**.

I noted above that one of Baumgartner’s conclusions in his (2018) is that it is impossible to get empirical evidence that differentially supports the claim that upper- level variables are causally efficacious, as opposed to the claim that they are causally inert. His argument is that for any structure like Figure 1 in which *M1* causes *M2* or *P2,* there will be an alternative, empirically indistinguishable structure in which (i) *M1* is causally inert with (ii) all the causation instead running through the *P1*🡪 *P2* ⇒ *M2* route.

Suppose that we adopt the account of **IV\***-interventions and what needs to be controlled for in Woodward, 2015 as described in previous sections (see especially Sections 3 and 6). According to this account, if, in Figure 1, *M2* changes under an **IV\***-intervention on *M1,* this would license the conclusion that *M1* causes *M2.* However (Baumgartner claims) this amounts to an unreasonable *apriori* preference for an account that assigns causal efficacy to the mental over an empirically equivalent account according to which *M1* is causally inert and causation runs only along the *P1*🡪 *P2* ⇒ *M2* route. (Baumgartner describes accounts of the latter sort as “epiphenomenalist”—a label I will return to below). On this basis Baumgartner criticizes philosophers like Woodward (he calls them “evidentialists”) on the grounds that they are mistakenly committed to the claim that one can resolve issues about the causal efficacy of the mental by appealing to empirical considerations.

I have several responses to these claims. First, a clarification: evidentialists do not (or at least need not) claim that the issue of the causal efficacy of the mental is *purely* empirical. What they claim is that (i) given what they take to be the correct account of causation, including an appropriate account of interventions and the sorts of confounders that need to be controlled for (that is an account along the lines sketched above), (ii) empirical evidence can be brought to bear on the causal efficacy issue. Obviously determining whether the account sketched in previous sections is correct has an important conceptual, *apriori* non-empirical, or if you like, “philosophical” component. The claim of evidentialists is not that this philosophical component can be bypassed or ignored or replaced by purely empirical arguments, but rather that *if* an account of the sort described above can be established as correct, then we can appeal to empirical evidence to decide whether and in what circumstances upper-level variables are causally efficacious. I see it as a virtue rather than a defect of compatibilist accounts of the sort that I favor that they can be used in this way—that they have the consequence that empirical evidence can be brought to bear on issues about when upper-level variables are causally efficacious. By contrast, other things being equal, it seems to me to be defect in an account of causation that it implies that even in very familiar circumstances it is impossible in principle to get empirical evidence regarding which of two competing causal claims is correct[[27]](#footnote-27).

To further explore this issue, suppose, as before, that *M2* changes under some **IV\***-intervention on *M1* with the supervenience relations being as in Kim’s diagram and compare the following representations of this situation:

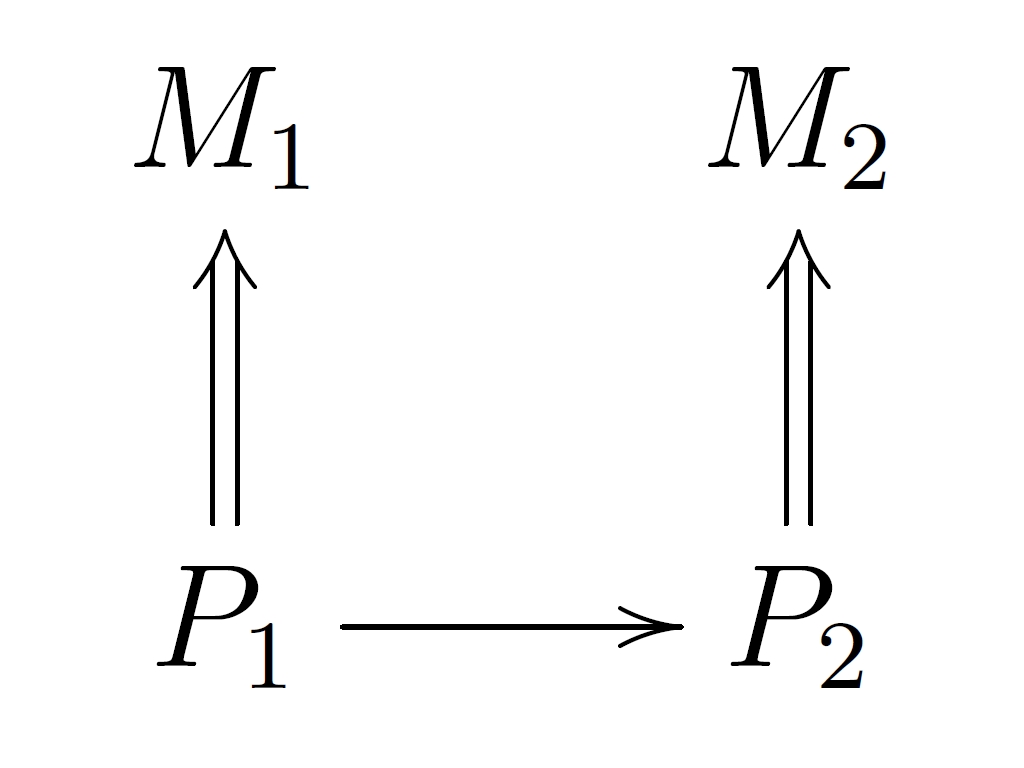


Figure 8

**Fig 8:** *P1* causes *P2*, *M1* supervenes on *P1*, *M2* supervenes on *P2* and *M1* causes *M2*

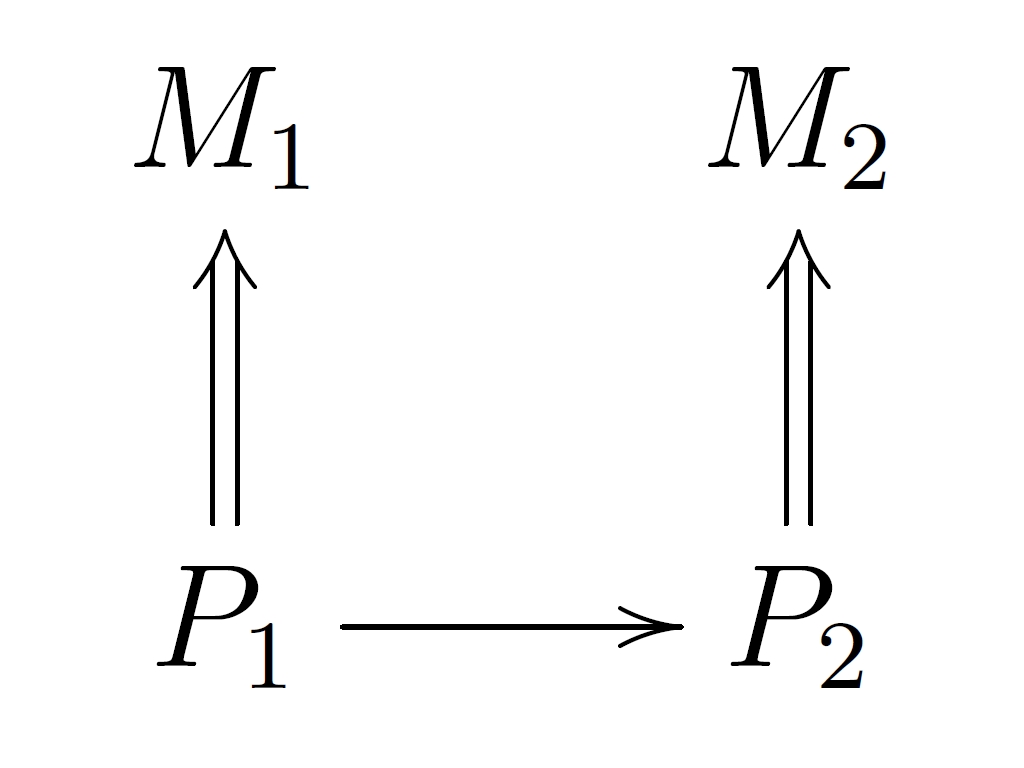


Figure 9

**Fig 9**: *P1* causes *P2*, *M1* supervenes on *P1*, *M2* supervenes on *P2* but *M1* is merely correlated with and does not cause *M2.*

The arrow between *M1* and *M2* in Figure 8 represents that *M1* causes *M2.* The undirected line (and absence of an arrow) between *M1* and *M2* in Figure 9 is my version of Baumgartner’s depiction of what he regards as an “epiphenomenalist” alternative: the undirected line represents that *M1* and *M2* are correlated under an **IV**\*-intervention on *M1* but claims that this correlational relation is not causal.

Of course the *diagrams* in Figures 8 and 9 are different but for the difference between them to correspond to a genuine case of empirical underdetermination, the figures must have interpretations that correspond to different possible ways the world might be. Showing this requires some account of what the presence or absence of causal arrows in a diagram means and of the conditions under which it is appropriate to either draw them or omit them—an account that makes it clear how the two figures represent distinct empirical possibilities. (Again this is the interpretive constraint.) Merely writing down two representations, one with an arrow from *M1* to *M2* and the other in which this arrow is absent but a non-causal correlation is said to be present does not provide such an interpretation. Here we encounter the same issue that we noted in connection with Baumgartner’s common cause representation of interventions: diagrams with and without arrows are presented but we are not given any story about how we are to understand the presence (or absence) of these arrows or the rules governing their use.

Since Figure 9, as understood by Baumgartner, corresponds to a case in which *M1* and *M2* remain correlated under an **IV**\*-intervention on *M1,* the version of interventionism that I have been defending judges that *M1* causes *M2* and hence that Figure 8 rather than Figure 9 is the correct description of the situation. Put differently, if the relation between *M1* and *M2* is non-causal and merely correlational, as Figure 9 claims, this requires that it not be the case that *M1* and *M2* remain correlated under an **IV**\*-intervention on *M1*, which is contrary to what we are assuming about the example.

I noted that Baumgartner claims that interventionism builds in an *apriori* preference against epiphenomenalism and in favor of mental (or more generally upper- level) causation. However, interventionism does distinguish between (what we might call) *ordinary* *causal* *epiphenomenalism* and alternatives to it. Ordinary causal epiphenomenalism involving the mental takes the relationship between the mental and physical to be like the relationship between a moving object and its shadow, with physical state *P1* causing mental state *M1*, *P1* causing physical state *P2* and *P2*  causing mental state *M2* but with no causal relation from *M1* to *M2*. (In other words, the supervenience relations in Figure 1 are replaced by ordinary causal relations.) In such a case, intervening on *M1* (in the sense of intervention described in Woodward, 2003 which does not take account of supervenience considerations) will not change *M2* so we can obtain evidence that *M1* does not cause *M2*. The situation depicted in Figure 6 is very different—from an interventionist perspective, relying on an **IV\***-type notion of intervention, the behavior of *M1* and *M2* is exactly "as if" mental causation is present but (according to Figure 9) mental causation is not “really” present. To distinguish this from ordinary causal epiphenomenalism I will call it epiphenomenalism\*.

Seen from this perspective the *apriori* preference for the causal efficacy of the mental built into the compatibilist version of interventionism described above is arguably not so unreasonable: what it amounts to is the claim that if two variables behave “exactly as if” a causal relationship is present between them (where “exactly as if” “means that the second changes in response to suitably defined interventions on the first—that is, interventions characterized by **IV\*** ) we are entitled to take that relationship as causal. Again one can think of the justification for this as broadly pragmatic—one won’t go wrong in any respect that matters (at least as interventionists understand what matters[[28]](#footnote-28)) if one adopts this preference.

**8**. **A Role for Non-empirical Considerations?**

Baumgartner also suggests that given (what he claims to be) the empirical equivalence of models that attribute causal efficacy to the mental and those that do not, we might consider looking to non-empirical considerations to choose between these models. For example, he suggests that epiphenomenal\* models (that is, models like that in Figure 6) will generally be simpler and will involve fewer redundancies (since they postulate fewer arrows) than empirically equivalent models that attribute causal efficacy to mental variables[[29]](#footnote-29). The former might be preferred on those grounds.

There are indeed proposals in the causal modeling literature for choosing among alternative models on the basis of what might broadly described as simplicity and avoidance of redundancy considerations. But typically these proposals are intended to apply to choices among models that (i) represent different ways the world might be and (ii) are *not* empirically equivalent[[30]](#footnote-30). For example, one anti-redundancy condition is the criterion of minimality. Suppose that we have a DAG **G** with variables **V** and an associated probability distribution **P** which is strictly positive and which satisfies the Markov condition[[31]](#footnote-31). Then **G** is minimal if P does not satisfy the Markov condition for any proper subgraph of **G**. For example, given the following (in)dependence relations X\_/|\_Z, Y\_/|\_Z, X\_/|\_Y, X\_|\_Y/Z, (with \_|\_ meaning independence and \_/|\_ meaning dependence) and the two DAGs in Figures 10 and 11

X Y

Z

Figure 10

**Fig 10**: A non-minimal DAG

X Y

Z

Figure 11

**Fig 11**: A minimal DAG

the DAG in Figure 11 is minimal and the DAG in Figure 10 is not. Note, however, that the structures in Figure 10 and Figure 11 are not equivalent with respect to their behavior under interventions. According to Figure 10 some intervention on *X* will change *Y*, while according to Figure 11 this is not the case. It is thus straightforward how to understand the claim that Figures 10 and 11 represent different possible ways the world might be.

For this reason, a preference (like that exhibited by epiphenomenalism\*) for representations containing fewer arrows over those containing more in a context in which those representations are equivalent under all possible interventions and yet are claimed to represent different possible ways the world might be causally is going to require a very different motivation from that which underwrites minimality or which underwrites other simplicity conditions which are claimed to guide choices among empirically inequivalent alternatives. Again we need an interpretation of the representations in Figures 8 and 9 which spells out both how they represent different ways the world might be and yet are equivalent under all possible interventions. We need to understand the alternatives that we are choosing among and how they differ from each other. Put differently: What exactly would it mean for a representation like that in Figure 9 to be “correct” and that in Figure 8 to be “wrong”, assuming that both describe the behavior of *M2* under an **IV\***-intervention on *M1*? This is an issue that arises for any exclusion-friendly account that, like epiphenomenalism\*, claims that, at least as far as responses to interventions (that is, **IV\*** interventions) go, matters are just as if upper-level variables have causal efficacy but in reality they do not.

**9. Realization Independence: Which Variables Need to be Included in a Causal Model?**

Let me conclude with an additional observation which may help to clarify how we should understand DAGs with arrows between upper-level variables when supervenience relations are present. We may distinguish two different purposes associated with the use of such DAGs. In one the goal is to model dependency relations among some limited set of variables, where this may include just upper-level variables. For example, a political psychologist might be interested in representing the causal relations between party identification *D*, openness to new experience *O*, religiosity *R* and support *S* for politically conservative candidates, where these are understood as psychological variables.

In this case, the researcher need not (and should not) deny that underlying each of these psychological variables are more-fine grained neurobiological and perhaps other physical variables on which the psychological variables supervene. Suppose, however, given her interests and what she can measure, the researcher does not include these lower-level variables in her model. When, if ever, is this justifiable? A sufficient condition is that the cause/ effect relations postulated by the model framed in terms of upper-level variables be intervention- supporting in the sense that the cause/effect relations among these variables should be such that interventions (understood as above) on the cause variables are stably followed by changes in the effect variables of the claimed sort, regardless of how these variables are realized at the lower level. This is the realization-independence conditional described in Section 3. It can also be formulated somewhat more explicitly in in terms of a notion that Woodward, 2020, 2021 calls *conditional causal independence*.

Suppose, as before, that we have an upper-level variable *U* the values of which are multiply realized by a lower level variable *L* (this can be extended to sets of variables in an obvious way but to simplify things I focus on single variables.) *L* is unconditionally causally relevant to (alternatively, causally irrelevant to or independent of) effect *E* if there are some (no) changes in the values of *L* when produced by interventions that are associated with changes in *E*. (Thus unconditional causal relevance is what is captured by a standard interventionist criterion for causation). *L* is *conditionally* *causally irrelevant* to or *conditionally causally independent* of *E* conditional on *U* iff *L* is unconditionally causally relevant to *E*, *U* is unconditionally causally relevant to *E*, and conditional on the

values of *U,* changes in the value of the *L* produced by additional interventions that are consistent (given the supervenience relations) with these values for *U* are irrelevant to *E*. In other words, we are to imagine a situation in which in which *U* and *L* are causally relevant to *E*, *U* is set to some value – say *ui* -- via an intervention and then *L* is set via independent interventions to each of the various values that are consistent with this value *U* = *ui*. If under all such variations in *L* for fixed values of *U,* the value of *E* does not change, and this is true for all values *ui*. of *U*, *L* is causally independent of *E* conditional on *U.* This condition would be satisfied if, conditional on values for the temperature of a gas, further variations in the kinetic energies of the individual molecules that are consistent with that value for the temperature make no difference to other thermodynamic variables like temperature.

As noted in Woodward, 2021, this condition of strict conditional independence can be relaxed in various ways to yield notions of restricted or approximate conditional independence, which may be more appropriate for some systems. For example, one might require that conditional independence holds only for some restricted range of values for the variables *U*, *L* and *E* where the restricted range corresponds to values that are most likely to be realized for some system of interest. Or one might require that the condition hold for almost all (e.g., all except a set of measure zero) variations in values of the lower level variable—a condition that is satisfied by real gases. Or one might require that conditional independence holds only approximately, say, in the sense that most even if not all of the variance in some target effect *E* is explained by the variance in *U,* given variations in *L* that are consistent with the values of *U*.

When strict conditional independence (or arguably some approximation to it) is satisfied, it is permissible or legitimate to use only the upper-level variables in the construction of a causal graph—roughly because the impact of the lower-level variables in the supervenience bases of the upper level-variables *with respect to various other upper- level variables*  is fully absorbed into the values of the upper-level variables themselves. (Again, because further variations in the lower-level variable *L 1* that forms the supervenience base for, say, *U1*, make no difference to value of some other upper-level variable *U2* , conditional on the values of *U1*, we can just use *U1*, to predict what the results of intervening on this variable will be for *U2*—we don’t need the information in *L1*. ) As noted earlier, this is what the transformations between lower and upper-level variables described by Rubenstein et al, 2017 accomplish: When the appropriate conditions on the transformations are met, this ensures that we may model the system of interest solely in terms of the upper-level variables without loss of causal information among the variables represented. For example, if our interest is in the causal relations among upper-level thermodynamic variables like temperature, volume and pressure, then the causal impact of a given change *ΔT* in temperature on volume will almost never depend on the details of the way in which *ΔT* is realized at the molecular level. Note however that this does not mean that causal relations among the lower- level variables do not exist— they are there, but we don’t *need* to represent them for the purpose at hand.

On the other hand, there are legitimate reasons for sometimes graphically representing the relations between upper-level variables and their supervenience bases in addition to the causal relations among the upper-level variables themselves. One might do this simply to make the former relations explicit or as part of a causal discovery strategy in which appropriate upper -level variables and causal relations among them are discovered from information about the relations among lower-level variables, as in Chalupka et al., 2017. As argued above, such representations of supervenience relations can be accomplished either by including upper and lower-level variables and their supervenience relations in the same graph, as in Woodward, 2015 or by employing two distinct graphs, with a specification of how the variables in and interventions associated with each graph are related, as in Rubenstein et al., 2017. In either case, we will need to specify additional rules and interpretative guidelines (besides those needed for characterizing conventional causal models). These additional rules are not ad hoc additions but simply reflect the fact that the representation of supervenience relations adds complexities that are not present in ordinary causal modeling.

1. **Conclusion**

I conclude with a more general observation. A number of readers of this article who are sympathetic to the exclusion argument have responded to my discussion by drawing a sharp distinction between the argument as a methodological proposal and the argument as metaphysics. They agree that to the extent we are interested in "practical" problems of modeling causal relations among upper-level variables and in causal explanations framed in terms of such variables, it is not appropriate to control for supervenience bases and that talk of causal relations involving upper-level variables is completely legitimate. In other words, in such contexts we can ignore the exclusion problem. It is suggested, however, that this is merely a claim about methodology or perhaps epistemology. When it comes to *metaphysics*, our assessment should be different-- the exclusion argument is correct qua metaphysics even if we have good grounds for not reasoning in accord with it when faced with practical tasks of causal modeling. For what is is worth, my view is that we should not be satisfied by this response. From my perspective it seeks to protect the metaphysical version of the exclusion argument by drawing a sort of protective boundary around it that blocks it from having implications for anything else we may be interested in -- implications that might otherwise force us to reconsider the argument. In addition it invokes a puzzling dualism-- claims can be true or legitimate in metaphysics but need not be adhered to elsewhere. We are provided with no real understanding aside from a vague invocation of pragmatics of how it can be the case that the exclusion argument can be evaded when it comes to methodology even though it is correct from the point of view of metaphysics. (If there is no coherent notion of intervening on upper-level variables, how exactly are the problems of interpreting causal claims in terms of interventions on such variables avoided when we engaged in modeling and methodology?) I will add, though, that even if I am wrong about this, it still should be of interest to philosophers of science and methodologists with "practical" interests to understand how we should think about interventions and their use in causal reasoning when supervenience relations are present. This is what this paper has tried to do.

**References**

[Baumgartner](http://philpapers.org/s/Michael%20Baumgartner), M. (2010). “[Interventionism and Epiphenomenalism”](http://www.uni-konstanz.de/FuF/Philo/baumgartner/docs/inter_epi.pdf) *Canadian Journal of Philosophy* 40: 359-383.

Baumgartner, M. (2018) “The Inherent Empirical Underdetermination of Mental Causation” *Australasian Journal of Philosophy* 96:335-350.

Baumgartner, M., Casini, L. and Krickel, B. (2018) “Horizontal Surgicality and Mechanistic Constitution” *Erkenntnis*: 1-14

Beckers, S., Eberhardt, F., and Halpern, J.Y. (2019). “Approximate Causal Abstraction” In Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI 2019).

Blanchard, T. , Murray, D. and Lombrozo, T. (Forthcoming) “Experiments on Causal Exclusion” *Mind and Language*

Chalupka, K., Eberhardt, F., Perona, P. (2017). “Causal Feature Learning: An Overview” *Behaviormetrika*, 44:137-164.

Gebharter, A. (2017). "Causal exclusion and causal Bayes nets".

 Philosophy and Phenomenological Research, 95 :353–375.

Ellis, G. (2016) *How Can Physics Underlie the Mind: Top-Down Causation in the Human Context*. Berlin: Springer.

Eronen, M. (2012) “Pluralistic Physicalism and the Causal Exclusion Argument” *European Journal for the Philosophy of Science* 2: 219-232.

Forster, M., Raskutti, G., Stern, R., & Weinberger, N. (2018). “The Frugal Inference of Causal Relations”. *The British Journal for Philosophy of Science,* 69: 821-848.

Hitchcock, C. (2001) "The Intransitivity of Causation Revealed in Equations and Graphs" *Journal of Philosophy* 98:273-299

Pearl, J. (2009) *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.

Polger, T., Shapiro, L., Stern, R. (2018). “In Defense of Interventionist Solutions to Exclusion”*. Studies in History and Philosophy of Science* Part A 68: 51-57.

Rubenstein, P., Weichwald, S. , Bongers, Mooij, J., Janzing, D. , Grosse-Wentrup, M. and Scholkopf , B. (2017). “Causal Consistency of Structural Equation Models.” In *Proc. 33rd Conference on Uncertainty in Artificial Intelligence (UAI 2017)*.

Shapiro, L. (2012) “Mental Manipulations and the Problem of Causal Exclusion” *Australasian Journal of Philosophy* 90: 507-524.

Spirtes, P., Glymour, C. and Scheines, R. (2000) *Causation, Prediction and Search*. Cambridge, MA: MIT Press.

Spirtes, P. and R. Scheines (2004). “Causal Inference of Ambiguous Manipulations”. *Philosophy of Science* 71:833–845.

Woodward, J. (2003) *Making Things Happen*. New York: Oxford University Press.

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

Woodward, J. (2014) “A Functional Account of Causation; or, A Defense

of the Legitimacy of Causal Thinking by Reference to the Only Standard

That Matters—Usefulness (as Opposed to Metaphysics or Agreement

with Intuitive Judgment)” *Philosophy of Science* 81: 691–713.

Woodward, J. (2015). “Interventionism and Causal Exclusion”. *Philosophy and Phenomenological Research*, 91(2): 303-3.

Woodward, J. (2017) “Intervening in the Exclusion Argument”. In Beebee, H., Hitchcock, C. and Price, H. (eds.) *Making a Difference: Essays on the Philosophy of Causation*. Oxford: Oxford University Press, pp 251-68.

Woodward, J. (2020) “Causal Complexity, Conditional Independence and Downward Causation” *Philosophy of Science* 87: 857-67.

Woodward, J*.*  (2021) “Downward Causation Defended” in Voosholz, J. and Markus, G. (eds.) *Topdown Causation and Emergence.* Springer: Cham, Switzerland.

Woodward, J. (Forthcoming) “Levels, Kinds and Multiple Realizability: The Importance of What Does Not Matter” In Shenker, Hemmo, Ioannidis, and Vishne (eds) *Levels of Reality in Science and Philosophy: Re-examining the Multi-level Structure of Reality* Springer

Zhong, L. (2020) “Intervention, Fixation, and Supervenient Causation” *Journal of Philosophy* 117: 293-314.

1. Some (e.g., Baumgartner, 2018) may hold that what follows from the impossibility of intervening on upper-level variables is merely (i) that is impossible to tell, at least by appealing to interventions, whether upper-level variables are causally efficacious, rather than (ii) that such variables are not causally efficacious. However, within the interventionist framework I am assuming (and which is the target of the criticisms under discussion), the impossibility of intervening establishes (ii) rather than (i). For additional criticism of (i) as an option, see Section 7. [↑](#footnote-ref-1)
2. Since this is already a long essay, one set of issues that, for reasons of space, I will not take up concerns the general claims, made by several authors, that the Bayes net formalism (and for some, interventionist ideas themselves) can be used to capture non-causal relations such as constitutive relevance as well as causal relationships. As should be clear from my discussion I think that such claims are mistaken. Among other considerations, they face serious technical difficulties since in the presence of constitutive (or for that matter supervenience relations) the requirement the probability distribution accompanying a Bayes net have positive support for all values of the variables involved is violated. However, this is not the place for a detailed discussion of this issue. [↑](#footnote-ref-2)
3. In fact it is addressed in one of the papers discussed below-- Chalupka et al., 2017. [↑](#footnote-ref-3)
4. I say more below about why supervenience relations of the sort considered in this essay should not be regarded as causal relations. [↑](#footnote-ref-4)
5. To forestall a possible conclusion, this is *not* a characterization that relativizes the notion of an intervention to some particular causal graph-- see discussion immediately below. At this point I am just reprising a standard notion of intervention that is defined with respect to causal systems that do not involve supervenience relations. The point is to distinguish these from systems that do involve supervenience relations, in which case the representation of interventions requires additional structure --again as discussed below. [↑](#footnote-ref-5)
6. The notion of a causal path is only defined for causal graphs: A causal path from *X* to *Y* in a *causal* graph G is a directed path from *X* to *Y* in G. The path from *P1* to *P2* to *M2* in Figure 1 is *not* a causal path. [↑](#footnote-ref-6)
7. To be more precise, Polger et al., 2018 consider graphs in which some but not all variables of the sort characterized by Kim's diagram are represented and "interventions" are understood as relative to these graphs. For example, they consider a graph G in which *M1* and *M2*from Kim's diagram are represented but not *P1* and *P2*. They then characterize a notion of intervention with respect to G. They add (p. 54, footnote 14) that they are assuming that the graphs they consider do not involve omitted common causes (presumably this is in recognition of the observation made above that such relativization yields mistaken conclusions when there are such omitted common causes.) They claim that this assumption is a "presupposition" of the interventionist framework. However, this is not a presupposition of the interventionist framework, if that framework is characterized as in Woodward, 2003. That framework is intended to apply to graphs in which there are omitted common causes. For example, the framework applies to a case in which the true structure is represented by figure 2 but the graph we employ omits the variable *Z*. In such a case if we were to carry out genuine interventions on *X* we would observe no changes in *Y* even though the graph we employ does not represent *Z*. In other words within the interventionist framework, whether it is appropriate to draw an arrow from *X* to *Y* depends on whether it is true that intervening on *X* will change *Y*  and this is not something that depends on whether we are working with a graph in which *Z* is represented. More generally, Polger et al.'s proposal greatly restricts the applicability of the interventionist framework in a way that seems unmotivated: we don't want whether *X* causes *Y* or whether an intervention on *X* with respect to *Y* has been performed to turn on whether we are operating with a graph in which all common causes are represented-- causation and intervention are not graph-relative in this way. Again, when we conduct a randomized experiment to determine whether *X* causes *Y*, the point of the randomization is to remove the influence of any common causes of *X* and *Y*, any common causes of the manipulation of *X* and *Y* and so on even though we may possess no representation of what the candidates for such causes may be. It is also worth noting that if the correct way of representing an intervention *I* on *M1* in Kim's diagram takes *I* to be a common cause of *M1* and *P1* (as Baumgartner, 2018 claims-- see below), then it is question-begging to claim, as Polger et al do, that a graph in which only *I*, *M1* and *M2* are represented does not omit common cause relations. Instead if Baumgartner's claim is correct, Polger et al's representation omits a common cause. My own view is that Baumgartner's claim is mistaken (see Section 7) but this is something that needs to be argued for, not assumed. [↑](#footnote-ref-7)
8. The notion of an intervention can be generalized in various ways to include “soft interventions” among other possibilities. Soft-interventions are not arrow-breaking but instead supply the variable intervened on with an exogenous source of variation. See Eberhardt and Scheines, 2007. I will ignore this possibility in what follows. [↑](#footnote-ref-8)
9. Recall that **DC** and the other interventionist criteria for causation require that interventions on the cause variable be possible. [↑](#footnote-ref-9)
10. A word about the phrase "fat handed" is appropriate here. To the best of my knowledge this phrase first came into use in the 1990s to describe a confounded manipulation. Unfortunately (and confusingly) it is now sometimes used (particularly in the literature I am discussing) to describe cases in which a manipulation has any effect on more than one variable. If (as seems uncontroversial) a manipulation of an upper-level variable *U* also affects its supervenience base *L* and *U* and *L* are not identical, it follows automatically from this new usage that such a manipulation is" fat handed". Moreover if, e.g., I administer a medication to an experimental control group, this manipulation will also count as "fat handed" if (as will be the case) it also disturbs the surrounding air molecules, even if this disturbance has no effect on recovery. Obviously this usage deprives the notion of fat handedness of any usefulness since pretty much any physically realizable manipulation will now count as fat handed. The crucial question ought to be not whether a manipulation affects more than one variable but whether it does so in a way that introduces confounding. In the medication example, the motion of the air molecules is presumably not a confounding variable if the effect of interest is recovery from an illness (this motion will not affect recovery), and so this manipulation is not usefully described as "fat handed". In the case in which the manipulation affects both *U* and its supervenience base *L*, whether this is to be regarded as involving confounding is exactly the point at issue. This is not something that can be settled just by adopting an expansive notion of fat handedness and assuming that fat handedness necessarily implies confounding. [↑](#footnote-ref-10)
11. See discussion below. [↑](#footnote-ref-11)
12. I acknowledge that in many realistic cases relations between lower and upper-level variables will be far more complicated than the simple possibility assumed here. Also, although my discussion is framed around the notion of supervenience, readers who are skeptical of this notion should substitute whatever other relation (besides type- identity) they think characterizes lower to upper-level relationships that are non-causal. My reasons for taking the lower to upper relation to be something other than type identity are that this is a background assumption in current discussion, which concerns whether the conjunction of forms of physicalism that do not require type identities and interventionism lead to exclusionist conclusions. Although this is not central to my discussion, I will add that in my view it is simply a fact that in present science, theorizing about relations between levels rarely takes the form of identity claims. One reason for this is that the variables and entities that figure in theories at different levels rarely line up with one another in a way that permits such identifications—see Woodward, forthcoming for additional discussion. For this reason, I do not think that exclusionist worries can be avoided simply by adopting type-identity accounts of interlevel relations. [↑](#footnote-ref-12)
13. Similarly *TC* is not a “disjunctive” property in the sense in which philosophers typically use that notion: *TC* is not equivalent to the disjunction of *HDL* and *LDL* and values of *TC* do not correspond to disjunctions of values of *HDL* and *LDL*. Values of *TC* might be thought of as equivalence classes of pairs of values, one member in the pair an *HDL* value and the other an *LDL* value, with each pair in the same equivalence class summing to the same value of *TC* but this is not captured by talk of disjunctive properties. [↑](#footnote-ref-13)
14. And even then this fit is very imperfect for reasons described in Woodward, forthcoming. [↑](#footnote-ref-14)
15. Philosophical discussions of supervenience and realization often assume that such relations obtain as a matter of “metaphysical” necessity. For my purposes, no particular assumptions of this sort are required. What is crucial is that the relations in question are unbreakable, whatever the source of this unbreakability might be. [↑](#footnote-ref-15)
16. This condition and its motivation are discussed in more detail in Woodward, 2015. One way of motivating the basic idea and the notion of non-causal “possibility” involved is to consider, outside of a causal modeling context, a standard way of presenting physical theories like Newtonian mechanics. Here one first describes the possible states that a system can be in, independently of the dynamical laws governing it. For example, in a system of N particles, it may be assumed that each particle can take any possible combination of values of the three variables along each spatial axis specifying its position and the three variables specifying its momentum. Moreover, each particle can take any possible combination of such values independently of the values for these variables taken by other particles. (This amounts to the assumption that each such variable value is “independently fixable”.)The notion of possibility invoked here is not causal possibility. The latter is specified independently by the dynamical laws, which characterize the causally possible relations for the system. In causal modeling, the analogous relations are specified by structural equations. [↑](#footnote-ref-16)
17. This observation raises an important issue which was highlighted by one of the referees in comments on an earlier version. The definition of an intervention in Woodward, 2003 and subsequent papers requires only that the intervention variable *I* "cause" the variable intervened on, *X*, to assume a particular value, where "cause" means "actual cause". If we assume an account according to which "cause" works in such a way that one can legitimately say that the upper-level intervention causes whatever particular value of *Lj* is realized, it follows from this understanding that an upper-level intervention that sets *Ui=ui* also sets *Lj* to whatever particular value that realizes *ui* on the particular occasion of this intervention. This is the way I think about interventions in this paper. However, as noted above, there is an obvious sense in which this upper-level intervention does not "control" which value of *Lj* is realized -- this in the sense that setting *Ui=ui* is not a reliable or repeatable way of setting any particular value of *Lj* All that is "controlled" is that the realizer in *Lj* is some member or other of the equivalence class of realizers of *Ui=ui*. So "control" and "cause" come apart. One could certainly imagine changing the understanding of what an intervention does to require that it must control and not just cause the realized value of the variable intervened on. It would then follow that an upper- level intervention like placing a container of gas in a heat bath is not an intervention on the lower-level molecular realizers of the temperature even though it is accompanied by some change in these. Of course it would still be possible to intervene on the molecular realizers of temperature but this would require a very different technology and intervention variable than the heat bath.

    Exploring this idea systematically would be very worthwhile but would require a different and even longer paper. Here I will just observe that one way of developing this idea might be to think in terms of a proprietary set of interventions associated associated with each "level" of variables -- a perhaps natural idea when one thinks of thermodynamics or folk psychology. Arguably this would also fit better with the approach to interventions involving different levels taken in Rubenstein et al., 2017 as described below. As nearly as I can see, such an alternative approach would not lend any new support to the exclusion argument.

    [↑](#footnote-ref-17)
18. Some writers (e.g., Baumgartner and Gebharter, 2016) also modify **IV4** in Woodward, 2003 but this change seems superfluous. [↑](#footnote-ref-18)
19. A recent paper by Blanchard et al. (forthcoming) is also relevant here. These authors report the results of a series of experiments to determine whether ordinary people endorse “exclusionist” causal judgments in contexts in which multiple realization is present. They find that people do not and instead endorse compatibilist judgments. Of course the fact that they do so does not by itself show that they are correct to do so but it does put pressure on the claim that exclusionist conclusions are built into ordinary thinking about causation. [↑](#footnote-ref-19)
20. This is a terrible way of testing for whether *X1* causes *Y*, but put that aside. [↑](#footnote-ref-20)
21. I lack space for a detailed discussion of Zhong's interesting paper. However of particular relevance to this paper is his claim (4.2) that there are cases in which it is appropriate to think in terms of an intervention that changes *M* from its actual value *m1* to *m2* while holding fixed the realizer of *m1*, say *p1,* at its actual value-- a claim that I have rejected. Zhong claims that we need to allow for this possibility when we test whether an upper-level property and not just one of its lower level realizers is causally efficacious. For example, if, in a case in which Sophie is presented with a scarlet target and pecks, we want to test whether the redness of the target causes Sophie to peck, we should consider (among others) (4.3) cases in which the target is red but not scarlet. I agree but do not think this requires (4.2). Within my framework, the appropriateness of considering (4.3) follows from my non-ambiguity condition that requires that if redness causes pecking, pecking should follow for all interventions that set the color of the target to red, regardless of whether red is realized by scarlet, crimson etc. Or put in terms of the conditional causal independence requirement described below (Section 9) , we set the target color to red and then via different, independent interventions, set the color to various specific realizations of red and see whether pecking follows. Neither of these tests involve a single intervention that changes the color of the target from, say, red to non-red, while keeping the realizer of red (scarlet in this case) fixed at is actual value -- something which I have taken to be impossible. In other words the appropriate test is not one in which the upper level property *M* is changed while whatever realizes the original value of *M* in its supervenience base is held fixed. Rather the appropriate test is the other way around. One considers interventions where *M* is fixed at some value *m1* and the realizers of m are allowed to vary, either "naturally" as will happen when *m1* is realized on different occasions or via independent interventions that fix the realizers to different values consistent with *m1*. [↑](#footnote-ref-21)
22. See, for example, Hitchock, 2001. [↑](#footnote-ref-22)
23. Again failure of independent fixability is one of the features that distinguishes supervenience relations from causal relations. [↑](#footnote-ref-23)
24. Some readers have worried that the use of the bracket (or anything similar such as the use of transformations between interventions at different levels as in Rubenstein at al.) is ad hoc and/or that such additional structure should be rejected because it complicates the standard directed graph representation. Note, however, that once we introduce supervenience relations and thick vertical arrows to represent them as in Figure 1, we have already introduced additional structure. Again it shouldn’t be surprising that if we want to talk about interventions operate in such contexts, we need additional representational devices of some kind to capture how they operate. As I see it, there is nothing sacrosanct about directed graphs; it is perfectly appropriate to modify these if the need to do so arises. [↑](#footnote-ref-24)
25. My suggestion that the intervention *I* does not have "independent effects" on *M1* and *P1* means simply that these effects and the relationships to *I* in which they figure are not independently disreputable. It is of course true that given the assumptions with which we are working *M1* and *P1* are not identical-- this does not imply, however, that they are "independent" in the independent disruptability sense. (Non-identity is a necessary condition for such independence but it is not a sufficient condition.) If this seems puzzling, consider that it is built into the notion of non-reductive supervenience that the relata of the supervenience relation are not identical but also not fully independent in the sense of being capable of varying fully independently of each other. [↑](#footnote-ref-25)
26. In a bit more detail, and slightly simplified, their proposal is this. Suppose that we have an lower-level causal model M*X* formulated in terms of structural equations involving variables *X* and an upper-level model M*Y* formulated in terms of structural equations involving variables *Y*. Let *f* be a function from *X* to *Y*. Let *I X* be the set of interventions on the *X* variables and *IY* be the interventions on the *Y* variables. Then M*Y* is an *exact f-transformation* of M*X* if the exists a surjective map *g*: *I X* -- > *IY* such that the result of intervening on the *X* variables with *I X* and then transforming that result via *f* to the corresponding result for the *Y* variables is the same as the result of carrying out the interventions *IY* on the *Y* variables that correspond to *I X*  as given by *g*. (I have omitted an additional requirement which is that *g* must be "order-preserving" in a sense that they specify--roughly that the compositional behavior of *I X* and *IY* must be coherent. This is not needed to convey their underlying idea.) When such an exact transformation exists this ensures that interventions on M*X* and M*Y* fit together in a way that yields consistent results and that interventions on M*Y* are well defined from the point of view of M*X* . As an illustration, suppose that M*X* is formulated in terms of the positions and momenta of the individual molecules making up a gas and M*Y* in terms of thermodynamic variables like temperature temperature and pressure. Then an intervention from *IY* on a thermodynamic variable like temperature will correspond to a set of many different compound interventions *I X*  on the positions and momenta of the gas molecules that are mapped into *IY* via g. If M*Y* is an *exact f-transformation* of M*X* and we perform such an intervention from *I X*  and calculate the results *X\** via the equations in M*X* and then transform *X\** to the corresponding *Y\** variables as given by *f*, the result should be the same as if we performed the corresponding interventions from *IY* as specified by g and then calculated the results *Y\** according to the equations in M*Y*. This condition-- the existence of an exact *f*-transformation -- will not be satisfied if, for example, different *X*-level interventions *I X*  are mapped into a *Y*-level intervention *IY* in such a way that (according to the equations of M*X* ) performing *IY* has different results on the *Y* variables variables depending on on which such *I X*  intervention realizes *IY*, as in the total cholesterol example.

    . [↑](#footnote-ref-26)
27. For more in defense of this assessment, see Woodward, 2021. [↑](#footnote-ref-27)
28. That is, one won’t go wrong regarding such matters as whether the correlation between *M1* and *M2* will disappear when one **IV\*** - intervenes on *M1*. Put differently, from an interventionist perspective, systematic changes in a variable under **IV\***-interventions on another just is causation; it is not mere "as if" or "ersatz" causation. Someone who wishes to contrast "as if" causation (understood in terms of appropriate behavior under **IV\*** interventions) with "real" causation needs to provide some alternative characterization of what this contrast consists in.

    . [↑](#footnote-ref-28)
29. In this connection, I should also note that there are several mistaken claims in Baumgartner 2018 about whether and in what respects standard approaches to causal inference implement non-redundancy requirements. (Baumgartner appeals to such claims to motivate his own arguments concerning avoiding redundancies.) For example, Baumgartner writes:

    The theory underlying Bayes-net procedures for causal inference [Spirtes, Glymour and Scheines, 2000] defines causes to be non-redundant probability-changers of their effects. Viz. probability-changes for which no off-screeners exist. (p. 12)

    But first of all, Spirtes et al. do not define “cause” at all. Second, were they to do so, they certainly would not define it in the way Baumgartner describes. This because in many cases there are a number of different non-equivalent causal structures that satisfy their Causal Markov and Faithfulness conditions (the conditions which underlie their search procedures) with respect to a given probability distribution—that is, even given Markov and Faithfulness, in many cases the independence/dependence information in the probability distribution underdetermines causal structure, so that one cannot use this information to define what it is for *C* to cause *E*. (See Spirtes et al. 2000, pp 59ff—note that while screening-off considerations are employed in the search procedures in this framework, that does not mean that causation itself is characterized in terms of such considerations.) Finally even if it is the case that there is some kind of definitional connection between causation and the Causal Markov condition there is no such connection between causation and Faithfulness—everyone agrees that it is possible for causal structures to violate Faithfulness. I mention this only because claims like these provide a misleading picture of the extent to which standard treatments of causation implement non-redundancy conditions of the specific sort that Baumgartner invokes. I’ll add that there are other non-redundancy conditions besides those mentioned above employed in current causal discovery procedures These include minimality and frugality (in the sense of Forster et al., 2018). I comment briefly on the these immediately below but neither provides support for epiphenomenalism\*. [↑](#footnote-ref-29)
30. In this connection it is worth recalling the contrast between what might be called (i) simplicity of representation of a single theory or model and (ii) more substantive notions of simplicity which are used to compare different theories—for example, a notion of simplicity according to which, ceteris paribus, we should prefer theories with fewer free parameters. In (i) we compare two different empirically equivalent representations of what is acknowledged to be the same situation and claim that one of these representations is simpler—as when a representation in polar coordinates is claimed to be simpler than an empirically equivalent representation of the same target in Cartesian coordinates. In (ii) we have competing hypotheses that make different claims about what the world is like (they are not empirically equivalent) and one of these hypotheses is preferred on the grounds that is “simpler”. I assume that the notion of simplicity which is in play when it is claimed that epiphenomenal\* models should be preferred because they are simpler is notion (ii)—the substantive notion. Baumgartner is not claiming that Figure 8 and Figure 9 are just alternative representations of the same causal structure. [↑](#footnote-ref-30)
31. The Causal Markov condition says that given a graph **G** and an associated probability distribution **P**, every variable in the graph is independent of its non-descendants, conditional on its parents. The positivity condition says that every value for the variables in **V** has non-zero probability. Of course positivity will fail in the presence of deterministic relationships and when positivity fails the minimality condition is arguably not a plausible constraint on model choice. The examples discussed in the text above and elsewhere in the literature on causal exclusion typically assume determinism, so for this reason alone (and independently of my other criticisms above) one cannot appeal to minimality as a criterion for model choice in such cases or in support of exclusionist conclusions.

    [↑](#footnote-ref-31)