Antireductionist Interventionism

Reuben Stern* Benjamin Eva†

February 7, 2020

Forthcoming in *The British Journal for the Philosophy of Science*‡

Abstract

Kim's causal exclusion argument purports to demonstrate that the non-reductive physicalist must treat mental properties (and macro-level properties in general) as causally inert. A number of authors have attempted to resist Kim's conclusion by utilizing the conceptual resources of Woodward's (2005) interventionist conception of causation. The viability of these responses has been challenged by Gebharter (2017a), who argues that the causal exclusion argument is vindicated by the theory of causal Bayesian networks (CBNs). Since the interventionist conception of causation relies crucially on CBNs for its foundations, Gebharter's argument appears to cast significant doubt on interventionism's antireductionist credentials. In the present article, we both (1) demonstrate that Gebharter's CBN-theoretic formulation of the exclusion argument relies on some unmotivated and philosophically significant assumptions (especially regarding the relationship between CBNs and the metaphysics of causal relevance), and (2) use Bayesian networks to develop a general theory of causal inference for multi-level systems that can serve as the foundation for an antireductionist interventionist account of causation.

1 Introduction

According to non-reductive physicalism, mental properties are not identical to physical properties, but nevertheless supervene on physical properties. The rough idea is that mental and physical properties are non-identical because the mental is multiply realized by the physical, but that everything is nevertheless physical in the sense that fixing something's physical properties fixes its mental properties. Non-reductive physicalism has struck many philosophers as plausible, but Jaegwon Kim (1989, 2000, 2003, 2005) has argued that it has an untoward consequence — namely, that mental properties are causally inert.

*Kansas State University, 66506, Manhattan, Kansas – https://sites.google.com/view/reubenstern/home – reuben.stern@gmail.com.

†University of Konstanz, 78464, Konstanz, Germany – https://be0367.wixsite.com/benevaphilosophy – benjamin.eva@uni-konstanz.de.

‡Both authors accept full and equal responsibility for what follows.

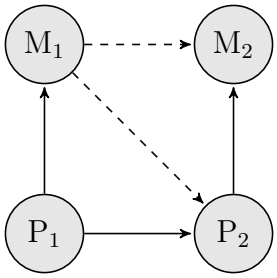

Figure 1: Informal Exclusion Argument

To illustrate Kim's argument, consider the following toy example from Kim (2005). Let P_1 and P_2 represent an agent's physical states at times t_1 and t_2, respectively. Similarly, let M_1 and M_2 represent their mental states at those times. Now, let us follow the non-reductive physicalist in assuming (i) that M_1 and M_2 supervene on P_1 and P_2 (respectively) and (ii) that P_1 is a sufficient cause of P_2.[1] The question at issue is whether these assumptions are compatible with regarding M_1 as a cause of M_2 or P_2.

Suppose we know that P_1, P_2, M_1 and M_2 are instantiated and are curious about what causally explains M_2's instantiation. Since the non-reductive physicalist contends that the occurrence of P_1 is sufficient for the occurrence of P_2, and that the occurrence of P_2 is sufficient for the occurrence of M_2, there is no causal work for M_1 to accomplish that goes over and above the causal contribution of P_1. Hence, were M_1 to cause M_2, then M_1 and P_1 would causally *overdetermine* M_2.[2] But according to Kim, this can't be right because effects are not systematically overdetermined by their causes, and we therefore must either reject non-reductive physicalism or accept that M_1 is not a cause of M_2. Moreover, because the same argument applies when it comes to explaining the occurence of P_2 (since P_1 is likewise a sufficient cause of P_2), M_1 cannot cause P_2, and it thus seems that we must either reject non-reductive physicalism or accept that mental properties are causally inert, period. Crucially, it's easy to see that the exclusion argument, as presented here, can be straightforwardly applied to demonstrate the causal inefficacy of any macro-level properties that are multiply realizable by micro-level counterparts.[3]

There is a vast literature analyzing the soundness of this informal version of the exclusion

[1] This assumption is often referred to as the "causal completeness of the physical" or the "causal closure of the physical." Interestingly, as we'll see in Section 4, the causal completeness assumption is superfluous when Kim's argument is viewed through the lens of CBNs.

[2] The same goes for P_2 since M_1 and P_1 threaten to causally overdetermine P_2 in exactly the same way.

[3] This aspect of our informal reconstruction of Kim's argument may lend some reason to doubt that it's not faithful to Kim's original argument since (as an anonymous referee helpfully points out) Kim (2003, p. 167) maintains that his argument does not imply the general thesis that a whole object cannot have causal powers over and above those had by its parts. (This suggests that Kim would take issue with the claim that his argument applies to the general class of multiply realizable macro-level properties.) If there is legitimate reason for this concern, this need not trouble the reader. For even if our reconstruction is not faithful to Kim's original argument, it *is* faithful to many descendants of Kim's argument that have occupied the literature — see, e.g., Baumgartner (2010), Gebharter (2017a), Polger et al. (2018), Sober and Shapiro (2000), and Woodward (2008). Any reader who shares this concern is welcome to view our paper as a response to the descendants of Kim's argument that apply to the general class of multiply realizable macro-level properties, rather than Kim's argument itself.

argument. Of particular interest here is the recent strand of literature in which a number of authors (e.g., Hitchcock (2012), List and Menzies (2009), Polger et al. (2018), Raatikainen (2010), Shapiro and Sober (2007), Shapiro (2010), Weslake (2015), and Woodward (2008, 2014)) assess Kim's argument through the lens of Woodward's (2005) interventionist account of causation. While some authors (e.g., List and Menzies (2009), Polger et al. (2018), and Woodward (2014)) have contended that an interventionist understanding of causation undermines some crucial premises of the exclusion argument, others (e.g., Baumgartner (2010) and Gebharter (2017a)) have argued that an interventionist conception of causation actually vindicates Kim's argument. Our aim in this article is to take the first steps towards developing a formally rigorous interventionist theory of multi-level causation by providing its foundations in terms of causal Bayesian networks (CBNs). The resulting framework not only reconciles the interventionist conception of causation with non-reductive physicalism, but also fills some significant theoretical lacunae in extant interventionist theories. More specifically, the plan is this.

We begin (§2) by providing a concise overview of the Spirtes et al. (2000) theory of CBNs and describing its relation to Woodward's (2005) interventionist account of causation. We then (§3) generalize the theory of CBNs so that it allows for the consideration of variables that enter into synchronic (non-causal) asymmetric supervenience relations. With this generalized framework in place, we subsequently (§4) reconstruct and criticize Gebharter's (2017a) vindication of the exclusion argument in terms of CBNs (by arguing that it relies on some highly contentious hidden premises concerning, first, what counts as an appropriate variable set in the context of multi-level causal inference, and, second, the relationship between CBNs and the metaphysics of causal relevance). We then (§5) develop a general approach to causal inference in multi-level settings that can be used to ground a plausible and formally rigorous interventionist theory of multi-level causation, and argue (§6) that the resulting theory can be used to make progress on some extant problems in the causal inference literature. Finally, we conclude (§7) by considering our approach against the backdrop of some examples from classical discussions of antireductionism.

2 Causal Bayesian Networks and Interventionism

The axiomatic theory of causal Bayesian networks (CBNs), as developed by e.g. Spirtes et al (2000) and Pearl (1988, 2009), provides the formal foundations for Woodward's (2005) interventionist theory of causation. We begin by providing a brief overview of the CBN formalism and its relationship to Woodwardian interventionism.

To start, suppose that there exists a set $\mathbf{V}$ of variables whose causal relationships we are interested in studying. Each variable $V \in \mathbf{V}$ has some discrete set of mutually exclusive and jointly exhaustive possible values.[4] For example, we might consider the variable M_t whose possible values represent the possible mental states m_i of an agent at a fixed time t.[5] The causal structure over $\mathbf{V}$ is the set of direct causal relationsips that obtain among the variables in $\mathbf{V}$.[6] This structure can be depicted as a directed acyclic graph (DAG) in which the nodes represent the variables in $\mathbf{V}$ and

[4]The framework extends to continuous variables, but we limit our discussion to the discrete case for ease of exposition.
[5]Throughout the paper, italicized capital letters refer to variables, and italicized lowercase letters refer to their values.
[6]See Woodward (2005) for a philosophical analysis of what constitutes a 'direct cause' in this framework.

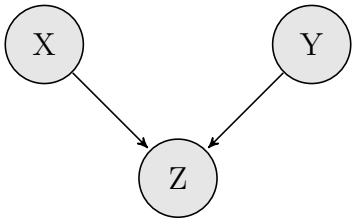

Figure 2: Common Effect Structure

the edges represent the direct causal relationships that obtain between pairs of variables in $\mathbf{V}$.[7] For example, the DAG in Figure 2 represents a causal structure in which X and Y are both direct causes of Z and in which neither X nor Y is a direct cause of the other.

If X is a direct cause of Y, we say that X is a *parent* of Y and that Y is a *child* of X. A *directed path* between two variables X and Y is an ordered sequence of variables $D = \langle X, ..., Y \rangle$ such that each variable in the sequence is a child of the variable that comes before it. If there exists a directed path from X to Y, we say that Y is a *descendant* of X and that X is an *ancestor* of Y.[8] If there does not exist a directed path from X to Y, we say that Y is a *non-descendant* of X.[9]

Among other things, the theory of CBNs provides the beginnings of a recipe for inferring causal structure from observational data. Specifically, suppose that you are interested in describing the causal structure over some variable set $\mathbf{V}$. Suppose also that you have observational data regarding the ways in which the values of the variables in $\mathbf{V}$ are correlated with one another. We can formalize this supposition by assuming that you have access to some full probability distribution Pr over the variables in $\mathbf{V}$, where Pr is informed by the observational data about these variables. The CBN axioms provide rules for interpreting the implications of Pr for the causal structure of $\mathbf{V}$. In particular, the axioms rule out many possible causal structures as incompatible with the empirical evidence encoded in Pr. The first and most fundamental axiom is the *Causal Markov Condition* (CMC), which Hausman and Woodward (1999) argue is "implicit in the view that causes can be used to manipulate their effects," and thus implicit in Woodward's (2005) interventionist account of causation.

Causal Markov Condition (CMC): A graph G and a probability distribution Pr satisfy the Causal Markov Condition if and only if every variable X in $\mathbf{V}$ is probabilistically independent of its nondescendants conditional on its parents according to Pr.[10]

The CMC encodes the assumption that causes screen off their effects. It is a generalization of Reichenbach's (1956) Principle of the Common Cause, which says that if variables X and Y are (unconditionally) correlated, then either X (directly or indirectly) causes Y, Y (directly or indirectly) causes X, or X and Y are (direct or indirect) joint effects of a common cause. To illustrate, suppose that $\mathbf{V} = \{X, Y, Z\}$ and that we know that X and Y are (unconditionally)

[7] A graph is acyclic if it does not contain any causal loops. The restriction to acyclic graphs encodes the idea that causal relevance is asymmetric in the sense that X cannot be both a cause of Y and an effect of Y.

[8] For technical reasons that need not concern us here, X is also considered a descendant of itself.

[9] Again, the one exception is when X and Y denote the same variable. We neglect this case in the body for ease of exposition.

[10] Where two variables X and Y are said to be *probabilistically independent* (or simply *independent*) when for any values x, y of X and Y, $P(y|x) = p(y)$.

correlated. The CMC entails that the causal structure depicted in Figure 2 cannot be right, since it posits neither any (direct or indirect) causal relationship between X and Y nor any common cause of X and Y. Thus, the CMC by itself somewhat restricts the range of candidate causal structures that are compatible with a given body of empirical evidence.

Still, the CMC does not narrow down the set of possible DAGs very much. For example, any fully connected DAG (in which there exists a directed edge between every pair of variables) is always consistent with the CMC. This underscores the fact that the CMC sticks its neck out with with respect to which edges must be included given Pr, but does not stick its neck out with respect to which edges should be absent given Pr. Contrapositively, the CMC dictates which probabilistic independencies must obtain given the absence of edges, but does not say which dependencies must obtain given the inclusion of edges. This means that the CMC must be supplemented with some additional condition in order to render the inclusion of a directed edge informative.

The weakest (and therefore least controversial) axiom that is standardly assumed in addition to the CMC is the *Causal Minimality Condition* (CMIN).[11] In order to state the CMIN, we need to introduce the notion of a *proper subgraph*. A DAG G' is a *proper subgraph* of G if and only if (i) G and G' are defined over the same variable set, and (ii) the set of parent-child relationships that obtain in G' is a proper subset of the set of parent-child relations that obtain in G.

Causal Minimality Condition (CMIN): A graph G and a probability distribution Pr satisfy the Causal Minimality Condition if and only if there exists no proper subgraph G' of G such that G' and Pr jointly satisfy the CMC.

To illustrate the inferential power of the CMIN, suppose that X and Y are probabilistically independent given any value of Z, but are unconditionally correlated. Then both of the DAGs in Figure 3 satisfy the CMC. However, the DAG on the right is not minimal. By deleting the edge from X to Y, we obtain the proper subgraph on the left, which still satisfies the CMC since X and Y are by hypothesis probabilistically independent given any value of Z. Thus the CMIN provides advice insofar as it tells us that we would be mistaken to treat X as a direct cause of Y in this case since there is a more economical representation of the causal structure that is compatible with what we know about the probabilistic relations between variables. Intuitively, the CMIN can be interpreted as telling us to include only those causal relationships that are necessary to ensure that the CMC is satisfied, or, alternatively, as requiring that each directed edge is encoding some actual dependence. Moreover, like the CMC, the CMIN has been shown by Zhang and Spirtes (2011) to be presupposed by interventionists in nearly every single case of causal inference that we ever actually confront.[12] How do these conditions underlie interventionism? Very roughly, the CMC is what ensures that the intervention on X is correlated *only* with its effects, and the CMIN is what

[11] See Forster et al. (2018) for recent discussion of the CMIN and its stronger counterparts.

[12] Zhang and Spirtes' (2011) point applies whenever the probability distribution over $\mathbf{V}$ is *positive* — i.e., when every possible assignment of values over $\mathbf{V}$ is assigned positive probability. We will see that when we consider variables that enter into non-causal dependence relations, the probability distribution over the variable set at hand is often not positive. But when the variable set is restricted to variables that are distinct in the sense required to qualify as the *relata* of causal relations, the distribution is very often (and perhaps always) positive. See Stern (forthcoming) for more discussion of this issue.

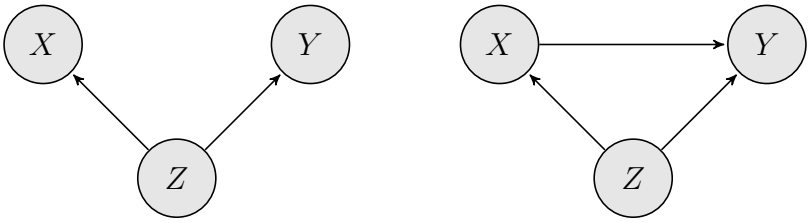

Figure 3: Minimal and Non-Minimal DAGs

ensures that the intervention on X is correlated with *all* of its effects.[13] Thus the CMC and CMIN jointly provide the axiomatic foundations for the theory of CBNs and the interventionist account of causation.[14]

There is just one more assumption of the CBN framework that we must introduce here, largely because it will play a crucial role in our analysis of Gebharter's (2017) CBN-theoretic parsing of the exclusion argument. The assumption concerns the kinds of variable sets to which the CBN axioms can be legitimately applied. Consider the variable set $\mathbf{V} = \{IC, SL\}$, where IC and SL represent daily ice cream sales and suntan lotion sales. Plausibly, these two variables are highly correlated (high suntan lotion sales are strongly indicative of high ice cream sales). Thus, when applied to $\mathbf{V}$, the CMC implies that there is some causal relationship between IC and SL. But this, of course, is implausible. The problem is that we've neglected the fact that the correlation between IC and SL is causally explained by some latent common cause — e.g., the weather (W). Omitting this common cause from the variable set $\mathbf{V}$ leads us to make a spurious causal inference, but when we consider the extended variable set $\mathbf{V}^+ = \{IC, SL, W\}$, we see that the structure in which W is represented as a common cause of SL and IC (and no other causal relationships obtain) satisfies both of the axioms (provided that SL and IC are screened off by W). Thus, the problem can be remedied by stipulating that we can only legitimately apply the CBN axioms to variable sets which are *causally sufficient*, where $\mathbf{V}$ is said to be causally sufficient if and only if for any $X, Y \in \mathbf{V}$, if Z is a common cause of X and Y, then $Z \in \mathbf{V}$.[15] The restriction to causally sufficient variable sets is a common background assumption in the theory of CBNs.

[13]We will see later that there are cases where the intervention on X is not unconditionally correlated with some effect of X (because of path cancellation), but even in these cases, the intervention on X is correlated with the relevant effect when one of the canceling paths is blocked by conditioning on an intermediate variable. The CMIN is what entails this conditional dependence.

[14]Strictly speaking, what Zhang and Spirtes (2011) show is that *if* the CMC is satisfied and one interprets direct causation in an interventionist manner, then the CMIN holds. Thus this is the precise sense in which their result bears on the foundations of interventionism. But there are other results in the offing. Gebharter and Schurz (2014) show how that the CMC and CMIN can be used to derive the interventionist treatment of direct causation, and Gebharter (2017c) and Stern (forthcoming) show that the CMC and the CMIN can be used to underwrite Woodward's non-direct notions of causal relevance in many contexts.

[15]There are numerous ways to narrow the set of common causes that must be included in a variable set. We opt for this stronger constraint in order to simplify things.

3 Generalizing Causal Graphs

Because the CMC is stated in terms of causal dependencies (insofar as *parents* and *non-descendants* are defined in terms of direct and indirect *causal* relations), the CMC and the CMIN cannot be justifiably assumed in multi-level settings — i.e., settings in which $\mathbf{V}$ is permitted to contain variables X and Y such that Y (asymmetrically) supervenes on X. The basic problem is that the CMC accounts for correlations by positing causal dependencies, but in multi-level settings, these correlations can be due to non-causal asymmetric supervenience dependencies. For example, if we assume the axioms over a variable set that includes one variable representing your current psychology and another representing your current neurophysiology, then, because the state of your psychology is evidentially relevant to the state of your brain, the axioms entail that either your current brain state causes your current psychology, or that your current psychology causes your current brain state. But this seems unreasonable since your current brain state and your current psychology are neither spatiotemporally distinct nor individually manipulable, and the *relata* of causal relations *are* spatiotemporally distinct and individually manipulable. So assuming the CBN axioms in a multi-level setting typically leads to spurious causal inferences.

Might there be a way of revising the CMC and CMIN in order to incorporate non-causal asymmetric supervenience dependencies? Perhaps it *prima facie* seems that we cannot because some asymmetric supervenience relations are very clearly not causal relations. But as Schaffer (2015) notes,[16] there are many structural similarities between the two notions. First, just as causes explain their effects, but not vice versa, it seems that the subvenient explains the supervenient, but not vice versa. To use Schaffer's example — just as one can explain Koko the gorilla's current psychological state with a causal story about previous events in her life, but not vice versa, one can explain Koko's current psychological state in terms of her current neurophysiology, but not her neurophysiology in terms of her psychological state. Second, and especially important here, it seems that supervenience relations undergird probabilistic screening-off relations in much the same way that causal relations do (see Schaffer (2015: 56-57)). If the well-being facts supervene on the psychological facts and the psychological facts supervene on the physical facts, then the physical facts are probabilistically independent of the well being facts conditional on the psychological facts. Similarly, if two aspects of a system supervene on its more fundamental properties, then the supervenience base screens off the two aspects. For example, if the color and electrical conductivity of a surface both supervene on that surface's subatomic structure, then color and conductivity are clearly probabilistically independent when we specify the surface's subatomic structure to a sufficient degree of precision.[17]

This suggests that we may be able to generalize the CBN framework to allow for the consideration of variable sets whose members supervene on one another. There is a sense in which Gebharter (2017a) has already accomplished this feat (since he deploys the framework in the current setting), but we prefer to reformulate the CBN axioms in our own way by introducing new terminology that will earn its keep in subsequent applications. However, insofar as it relates to the analysis of the exclusion argument, our generalization is materially equivalent to Gebharter's.

[16] Schaffer is primarily interested in grounding rather than supervenience, but the structural analogies he observes all carry over to the supervenience case.

[17] This follows from the fact that the subatomic structure fixes both.

The easiest way to revise the CBN framework so that it accommodates asymmetric supervenience dependencies is to reinterpret the significance of a directed edge in a DAG disjunctively — i.e., so that the presence of a directed edge from X to Y no longer signifies that X is a direct cause of Y, but rather signifies that *either* X is a direct cause of Y, *or* Y directly (asymmetrically) supervenes on X. Let us say that if either (i) X is a direct cause of Y, or (ii) the value of Y directly (asymmetrically) supervenes on the value of X, then X is an *e-parent* of Y and that Y is an *e-child* of X.[18] An *e-directed path* between two variables X and Y is an ordered sequence of variables $D = \langle X, ..., Y \rangle$ such that each variable in the sequence is an e-child of the variable that comes before it. If there exists an e-directed path from X to Y, we say that Y is an *e-descendant* of X and that X is an *e-ancestor* of Y. If no e-directed path from X to Y exists, then we say that Y is a *non-e-descendant* of X.[19] With this terminology in hand, we can slightly modify the causal modeling axioms in order to incorporate non-causal supervenience dependencies as follows.

Multi-Level Markov Condition (MMC): A graph G and a probability distribution Pr satisfy the Multi-Level Markov Condition if and only if every variable X in $\mathbf{V}$ is probabilistically independent of its non-e-descendants conditional on its parents according to Pr.

Multi-Level Minimality Condition (MMIN): A graph G and a probability distribution Pr satisfy the Multi-Level Minimality Condition if and only if there exists no proper subgraph G' of G such that G' and Pr jointly satisfy the MMC.

The basic motivation behind the generalized axioms is simple. In the standard setting, the CBN axioms can be interpreted as specifying which causal relations we need to posit in order to adequately account for all the observed correlations. When we generalize the setting to allow for variables which supervene on one another, we introduce the possibility that the observed correlations are indicative of supervenience relations, rather than causal relations. The MMC takes this possibility into account by generalizing the CMC condition in the obvious way—i.e., by accounting for observed correlations via dependence relations which could be either causal relations or supervenience relations.

Gebharter (2017a) provides a list of three independent justifications for a multi-level generalization of the CBN framework, but we think the most compelling motivations are (1) that causal and supervenience relations seem to ground screening off relations in much the same way, (2) that there are concrete realistic causal inference tasks in which it is intuitively desirable to consider multi-level variable sets (and it is *prima facie* desirable that we be allowed to continue using the theory of CBNs in these cases), and (3) that the MMC and MMIN axioms are intuitively plausible generalizations of the CMC and CMIN that provide plausible solutions to difficult causal inference problems (as we will see in §6). Furthermore, we should note that readers who remain unconvinced by these considerations still have something to gain from the present analysis. Specifically, in section §4, we show that *even if* one accepts Gebharter's generalization of CBNs to the multi-level setting, it is still possible to reject his purported vindication of the causal exclusion argument *on his own terms*

[18] We prefix these notions with 'e-' because both causal relationships and asymmetric supervenience relationships capably support *e*xplanations.

[19] As before, the one exception is the case where X and Y denote the same variable. Just as we treated X as a descendant of itself, we treat X as an e-descendant of itself.

— i.e., within the generalized framework.

Now, one concern that one might have about the generalized axioms is that although they may be useful for identifying when two variables are related by some dependence relation, they do not specify the nature of the dependence. For example, imagine that the axioms identified some DAG as compatible with Pr in which X is an e-parent of Y. The axioms don't tell us anything about whether we should regard X as a direct cause of Y or, alternatively, as a supervenience base of Y. Clearly, if we want to get an accurate picture of the causal structure of $\mathbf{V}$, we need a way to distinguish between the edges that represent causal dependencies and those that represent supervenience relations.

Towards this end, it is helpful to characterize exactly what it means for the value of one variable to asymmetrically supervene on the value of another variable in a probabilistic setting.[20] Since it is often said that Y asymmetrically supervenes on X when (i) changes to Y necessitate changes to X but not vice versa, and (ii) X determines Y, it is natural to say that Y supervenes on X when (i) multiple values of X are compatible with some value of Y but not vice versa, and (ii) any value of X fixes the value of Y. Here, a value x of X can be understood as "compatible" with a value y of Y when $Pr(x|y) > 0$, and x can be said to "fix" the value of Y when $Pr(y|x) = 1$ for all values of X and Y.[21,22] For example, when we say that your psychological state asymmetrically supervenes on your physical state, it is implied that there are multiple distinct physical realizations compatible with your psychological state, but that there are not multiple distinct psychological realizations compatible with your physical state (since your physical state fixes or determines your psychological state).

With this rough characterization of asymmetric supervenience dependencies in place, we can begin to investigate what distinguishes supervenience relations from causal relations in multi-level settings in which MMC and MMIN are assumed. At first pass, it is attractive to claim that Y supervenes on X exactly when (i) Y is an e-descendant of X, (ii) multiple values of X are compatible with some value of Y but not vice versa, and (iii) any value of X fixes the value of Y. It is clear, we think, that these three conditions are necessary for Y to asymmetrically supervene on X, but it is less clear that they are jointly sufficient because this rules out the possibility of some

[20] As we emphasize below, there may be reason to think that metaphysical asymmetric supervenience relations cannot be fully characterized in terms of probability theory. Still, it's important to get the probabilistic signature right — especially since this is where the action lies in the context of incorporating asymmetric supervenience relations into the CBN framework.

[21] When supervenience is understood in this way, it is clear that the operative probability distributions will not generally be positive in the sense that every assignment of values over $\mathbf{V}$ is assigned positive probability. This means that the MMIN cannot be justified in multi-level contexts on the grounds that Zhang and Spirtes (2011) show that interventionist treatments of causation presume the CMIN. But this shouldn't surprise or worry us too much. First, we share this assumption with Gebharter (2017a), so it is clearly appropriate to assume as we engage with his version of the exclusion argument. Second, since interventionist treatments of causation traditionally say nothing about when supervenience relations should be posited, it is obvious from the get-go that no such formal result or justification is in the offing. The assumption of the MMIN thus rests simply on its plausibility as a generalization of the CMIN to multi-level contexts. If the CMIN is plausible when no supervenience relations are present, then, by our lights, the MMIN is plausible when they are.

[22] Because our probabilistic characterization of supervenience specifies nothing in addition to these determination relations, it is neutral between every theory of supervenience that says that when Y supervenes on X, there is no change in Y without a change in X.

kinds of deterministic causation (if no dependency can be both a supervenience dependency and a causal dependency). Consider a system defined over a variable L that encodes whether a light is on or off, and another variable S that encodes whether a switch is engaged in one of three positions: off, dim, or bright. If putting the switch into any position other than 'off' is sufficient for the light's being on, then a directed edge from S to L meets the conditions for supervenience. So despite the common intuition that the dependence between the S and L is causal, it appears to count as an asymmetric supervenience relation, rather than a causal relation.

There are a several ways to respond to this. First, since supervenience dependencies are typically regarded as synchronic, we can add a fourth condition to the analysis requiring that X and Y must not be spatiotemporally distinct in order for Y to supervene on X. With this condition in place, all four conditions can be treated as individually necessary and jointly sufficient without yielding the verdict that L supervenes on S since L and S describe spatiotemporally distinct states of affairs. Second, one might acknowledge that L supervenes on S, but contend that L can supervene on S *and* be caused by S when L and S describe spatiotemporally distinct states of affairs. This response has something going for it since the value of S determines the value of L in exactly the same way that is characteristic of supervenience, but it requires a rather drastic revision to the philosophical lexicon since we do not typically characterize one and the same dependency as both causal and supervenient.[23] Finally, one can respond that S does not cause L despite appearances because, for one reason or another, the dependence between S and L does not meet the conditions required for causation.

This last response may strike some as the least intuitive since philosophers sometimes write as though deterministic causation is the norm. But as it turns out, through the lens of MMC and MMIN, there is at least some reason to treat a dependency as non-causal when it meets the first three conditions. Hausman (1998) convincingly argues that many plausible analyses of causal relevance are committed to the claim that if X causes Y, then Y is also caused (causally influenced) by some means that are independent from X.[24] As things turn out, it is true (given MMIN and MMC) that Y cannot be caused by means independent from X if X and Y jointly meet the first three necessary conditions for supervenience provided above (because no minimal graph contains an edge that can represent such a dependence). So there is at least some good reason to think of these relations as non-causal.

At any rate, we assume in what follows that we have access to a principled method for distinguishing edges that represent causal dependencies from edges that represent supervenience dependencies. We don't stick our necks out regarding which method is best, and the reader is welcome to fill in the gaps however they see fit.

We now turn to our reconstruction of Gebharter's (2017a) CBN-theoretic formulation of the exclusion argument in our multi-level framework.

[23]For those who regard supervenience as a completely formal dependence, this may not require any revision to the concept (since cases like the light switch still exemplify scenarios in which there is no change in one variable without change in another). We take no stand on whether a purely formal characterization of supervenience is more useful for philosophical purposes than others that are additionally intended to capture some metaphysical dependence between higher and lower level properties.

[24]A cause of Y qualifies as independent from X if and only if it is neither causally downstream nor upstream from X and X and Y are not (indirect or direct) effects of some common cause.

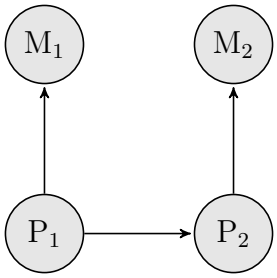

Figure 4: Assumed Dependencies in Exclusion Argument

4 CBNs and Causal Exclusion

The table is now set to consider the exclusion argument against the backdrop of our generalized axiomatic framework.[25] Recall the simple example where P_1 and P_2 are variables whose values are given by the possible physical states of an agent at times t_1 and t_2, respectively. Similarly, let M_1 and M_2 be variables whose values are given by the agent's possible mental states at those times. Suppose that the agent's mental state at a time supervenes on their physical state at that time, i.e., the value of P_1/P_2 fixes the value of M_1/M_2. Suppose further that the agent's physical state at time t_1 causally influences (and is correlated with) their physical state at time t_2. These assumptions straightforwardly require that we posit each of the directed edges that are depicted in Figure 4.

This is in line with the usual formulation of the exclusion argument, where it is assumed that M_1 and M_2 supervene on P_1 and P_2, and that P_1 directly causally influences P_2. The question now is whether the generalized CBN axioms require/allow us to add additional edges from M_1 to P_2 and/or M_2. First, we need to check whether the DAG in Figure 4 (call it 'G') satisfies the MMC. According to the MMC, G requires only that M_1 and M_2 are independent conditional on any value of P_1 or P_2. And indeed, it is easy to see that both of these conditional independencies are guaranteed to hold by the supervenience assumptions outlined above. Specifically, since M_1 is assumed to supervene on P_1, conditioning on any value of P_1 will uniquely fix the value of M_1. And once the value of M_1 is fixed with probability 1, it can no longer be correlated with M_2 (or with any variable whatsoever). Symmetric reasoning shows that the independence of M_1 and M_2 given P_2 is guaranteed by the assumption that M_2 supervenes on P_2. Thus, the two conditional independencies which are required for G to satisfy the MMC are indeed guaranteed to hold by the assumptions of the exclusion argument. This immediately entails that any DAG G^* which — (i) respects the assumptions of the exclusion argument, and (ii) includes an edge from M_1 to either P_2 or M_2 — is bound to violate the generalized minimality condition MMIN. Why? Since G itself satisfies MMC, any supergraph of G will not qualify as minimal (including those in which there are edges protruding from M_1).

At first blush, this looks like an elegant formal vindication of the causal exclusion argument. Assuming only the multi-level generalizations of the basic axioms of the theory of CBNs (and the supervenience of the mental on the physical), we seem to have demonstrated the causal inefficacy of mental phenomena. Furthermore, there is a clear sense in which the argument given here justifies

[25]The formalization of the argument given here is somewhat different from Gebharter's formalization, but not in any way that affects the philosophical analysis that follows.

the intuition behind Kim's original exclusion argument. The standard informal version of the exclusion argument is based on the idea that mental causation is in some sense redundant (since all the real causal work is being done at the physical level), and that positing mental causation leads to overdetermination. Similarly, the argument from MMIN shows that it is possible to have an adequate representation of causal structure (in the sense of satisfying the MMC) without including any mental causation. So mental causation is theoretically superfluous, and positing it means positing redundant causal relations. As Gebharter puts it,

> Our result may be interpreted as empirically informed support for epiphenomenalism or as evidence against non-reductive physicalism: If causation is characterized by means of the causal Markov condition and the causal minimality condition, we assume that mental properties are non-identical to their physical supervenience bases, and that every physical property has sufficient physical cause, then mental properties cannot act as causes for physical properties or as causes for other mental properties – they possess no causal power. (Gebharter 2017a: 364)

At this stage, it's worth pausing to note that in our reconstruction of the exclusion argument, the 'causal closure of the physical' assumption plays no role. In particular, unlike Gebharter (2017a), we do not assume that every physical property has sufficient physical cause.[26] Indeed, we don't assume that the relationship between P_1 and P_2 is deterministic in any way. We assume only that P_1 has some causal influence on P_2, and leave the nature of that causal relationship completely unspecified. This is significant, since the causal closure assumption has been the source of much controversy (regarding both its plausibility and its proper formulation) in the literature (see e.g. Baker (1993), Hendry (2006) and Stapp (2005)), and dispensing with the assumption seems to significantly strengthen Kim's argument.[27]

Overall, then, the generalized CBN axioms ground a formally rigorous statement of the causal exclusion argument that seems to entail, in full generality, the causal inefficacy of mental phenomena (given nonreductive physicalism).[28] Moreover, the resulting formulation of the argument dispenses with one of the strongest and most controversial premises of the standard formulation of the exclusion argument (the causal closure of the physical). So as far as CBNs are concerned, things may seem to look good for the exclusion argument. But not so fast. It's time to put the champagne back on ice.

[26] In later work, Gebharter (2017b) himself acknowledges that this assumption need not play any role and produces his own alternative formalization of the argument that dispenses with the causal closure condition.

[27] It is worth noting that Kim provided an alternative version of his own argument that relied on the causal closure assumption rather than any claim about supervenience. This version is neither strictly stronger nor strictly weaker than our own.

[28] It is worth noting that, unlike previous interventionist discussions of the exclusion argument, this reconstruction focuses primarily on the implications of the multi-level CBN axioms, rather than on the formally related problem of representing macro-level interventions in multi-level settings. We intend to return to this latter issue in future work that generalizes Eva and Stern's (2019) interventionist treatment of causal explanatory power to multi-level settings.

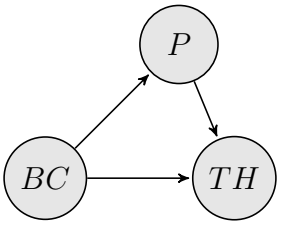

Figure 5: Birth Control, Pregnancy, Thrombosis

5 Antireductionist CBNs

In the previous section, we showed that any DAG which represents the causal structure of the variable set $\mathbf{V} = \{P_1, P_2, M_1, M_2\}$ in a way that satisfies some antireductionist commitments and MMC and MMIN will depict M_1 as causally inert. But this observation alone is not sufficient to warrant the conclusion that M_1 is causally inert.

To illustrate, consider Hesslow's (1976) example (depicted in Figure 5), according to which taking birth control (BC) probabilistically promotes thrombosis (TH) if you are pregnant (or if you aren't), but also reduces the risk of thrombosis by reducing the risk of pregnancy (P). Now, further suppose that the probabilistic effect of BC that is mediated by P cancels out the probabilistic effect that is due to BC's direct effect on TH.

The graph in Figure 5 accurately captures the true causal structure over the variable set $\mathbf{V} = \{TH, BC, P\}$. But this is not the only variable set that is deemed appropriate for consideration in this context. Since the variable P is not a common cause of BC and TH, there is nothing wrong with omitting it from the variable set under consideration (at least according to standard practice). Thus the variable set consisting only of BC and TH is considered fair game for causal inference. But when we apply the CMC and the CMIN to the reduced variable set $\mathbf{V}^- = \{BC, TH\}$, we obtain a DAG that does not include any causal arrows despite our background knowledge that BC is causally relevant to TH. For, by stipulation, BC and TH are probabilistically independent, which means that no directed edges are required in order to satisfy the CMC, and hence that the completely unconnected graph is minimal.

The lesson of this is of course not that BC is causally irrelevant to TH. Rather, it's that we cannot conclude that there is no causal relationship between two variables when we find that those variables are not linked by a directed edge in some admissible DAG over some causally sufficient variable set. That's just not the way that the epistemology of causation is related to its metaphysics — i.e., the absence of an arrow in a DAG does not always mean the absence of a causal relationship in the world. For, it could be that there exists another causally sufficient variable set relative to which there *does* exist an admissible DAG in which X is depicted as a cause of Y. And in some cases, this is enough to indicate that X really is a cause of Y (e.g., when X and Y are BC and TH).

To make things a bit more precise, consider the following two principles concerning the relationship between the existence of admissible DAGs over causally sufficient variable sets and the causal structure of the world.

Strict Causation Principle (SCP): In order for a variable X to count as causally relevant to a variable Y, it must be the case that for *every* causally sufficient variable set $\mathbf{V}$ containing X and Y, there exists a directed path from X to Y in some admissible graph over $\mathbf{V}$.

Weak Causation Principle (WCP): In order for a variable X to count as causally relevant to a variable Y, there must be *some* causally sufficient variable set $\mathbf{V}$ containing X and Y such that there exists a directed path from X to Y in some admissible graph over $\mathbf{V}$.

If we employ the SCP, then we are forced to conclude that in cases like the one describe above, BC is not a cause of TH, which looks like the wrong verdict. More generally, it seems that the SCP sets the bar too high for identifying the presence of causal relations in the world. In contrast, the WCP gets the case just right. Since BC is depicted as a cause of TH in an admissible DAG over *some* causally sufficient variable set, it's consistent with the WCP that BC really is a cause of TH. Of course, since the WCP and the SCP only specify candidate necessary conditions for causal relevance, one must identify some other truths about causal relevance in order to identify conditions that are jointly sufficient. But if, for example, causes temporally precede their effects, then as Stern (forthcoming) argues, we can justifiably infer that X is causally relevant to Y whenever X is a direct cause of Y relative to some causally sufficient variable set, where the operative notion of 'direct cause' takes stock of the fact that causes must temporally precede their effects.[29] Similarly, according to Woodward (2008b), if we use his (2003) interventionist treatment of causation to find that X is a "contributing cause" of Y relative to *some* variable set, then we can safely infer that X is causally relevant to Y, *simpliciter*.

We return now to the CBN-theoretic reformulation of the exclusion argument. In this context, in order to infer a DAG from a probability distribution, we must assume more than just that causes temporally precede their effects since there are now non-causal relations at play. But if we additionally assume that non-causal supervenience edges must go from the more fundamental (micro-level) to the less fundamental (macro-level),[30] then, as we've implicitly shown above, any DAG over the variable set $\mathbf{V} = \{P_1, P_2, M_1, M_2\}$ that satisfies MMC and MMIN will depict M_1 as causally inert. Since there's no reason to think that $\mathbf{V}$ omits any common causes, we know that there is an admissible DAG containing M_1 and M_2 relative to which M_1 is not depicted as a cause M_2. And the same goes for M_1 and P_2. Thus if we assume the SCP, then M_1 turns out to be causally inert. But as we've just argued, we don't even need to consider the multi-level setting in order to see that the SCP condition is too strong — i.e., even in the single-level setting, there are cases where two variables are causally related despite the fact that the relation is not apparent relative to some appropriate variable set. Hence our advocacy of the WCP over the SCP. And once we replace the SCP with the WCP, the observation that M_1 is depicted as causally inert relative to some causally sufficient variable set is not enough to warrant the conclusion that $M1$ is causally inert in the world. It could be that, like BC, the causal efficacy of M_1 is revealed in some variable

[29] To be clear, Stern's treatment of causal relevance is in keeping with the WCP insofar as he argues that X is a cause of Y exactly when X is a direct cause of Y relative to *some* causally sufficient variable set. The temporal aspect of Stern's treatment of causal relevance comes in at the level of his understanding of direct causal relevance, and thereby provides a necessary condition for causal relevance that works in tandem with the WCP and the axioms of the CBN framework.

[30] In this paper, we always make this assumption in multi-level contexts.

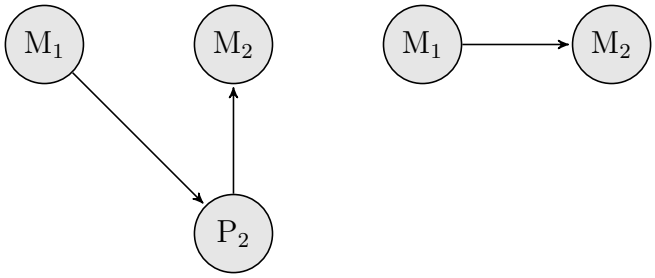

Figure 6: Minimal DAG's over $\mathbf{V_1}$ and $\mathbf{V_2}$

sets, and masked in others.

So the antireductionist has an escape route. They can adopt the WCP and attempt to identify some other causally sufficient variable set relative to which there exists some admissible DAG that depicts M_1 as causally efficacious, thereby vindicating the antireductionist's conviction that mental events need not be causally inert. Two natural variable sets to consider here are $\mathbf{V_1} = \{P_2, M_1, M_2\}$ and $\mathbf{V_2} = \{M_1, M_2\}$. The DAGs in Figure 6 satisfy both MMC and MMIN for these variable sets.[31] The set $\mathbf{V_1}$ is obtained by omitting M_1's supervenience base P_1, and relative to the minimal DAG over $\mathbf{V_1}$ depicted in Figure 6, M_1 is represented as a direct cause of P_2. The set $\mathbf{V_2}$ is obtained by omitting both of the physical variables, P_1 and P_2, and relative to the minimal DAG over $\mathbf{V_2}$ depicted in Figure 6, M_1 is represented as a cause of M_2. Thus, if we are happy to accept the variable sets $\mathbf{V_1}$ and $\mathbf{V_2}$ as causally sufficient, then the WCP allows M_1 to qualify as a cause of both P_2 and M_2.

The antireductionist can make a simple argument to get across the finishing line here. Neither P_1 nor P_2 are common *causes* of any pairs of variables under consideration. The only causal relation that either P_1 or P_2 stand in with respect to these variables is that P_1 causally influences P_2. By stipulation, the dependence of M_1/M_2 on P_1/P_2 is a supervenience relation, not a causal relation. So neither $\mathbf{V_1}$ nor $\mathbf{V_2}$ omit any common causes. Ergo, they are both causally sufficient variable sets. So the antireductionist can conclude that regarding M_1 as a cause of both P_2 and M_2 is perfectly consistent with the WCP.

Fans of Kim's exclusion argument may be unimpressed by this response. They can counter that like the CMC and the CMIN, the definition of what counts as an appropriate variable set needs to be generalized upon moving to the multi-level setting. In particular, they can propose that we replace the normal definition of causal sufficiency with the following natural multi-level generalization.

E-Parent Sufficiency: A variable set $\mathbf{V}$ is *e-parent sufficient* if and only there do not exist any variables L, X, Y such that (i) $L \notin \mathbf{V}$, (ii) $X, Y \in \mathbf{V}$, and (iii) L is a common e-parent of X and Y in an admissible graph over the extended variable set $\mathbf{V} \cup \{L\}$.

This is just the obvious generalization of the causal sufficiency condition to the multi-level setting. To get a feeling for how it works, recall the variable sets $\mathbf{V_1} = \{P_2, M_1, M_2\}$ and $\mathbf{V_2} =$

[31] Here, we assume only that M_1 is correlated with P_2 and M_2. This is a very weak and compelling assumption, given the premises of the exclusion argument (that M_1 and M_2 supervene on P_1 and P_2 and that P_2 is causally influenced by and correlated with P_1). In fact, if we suppose that all values of P_1 and P_2 have positive prior probability, then the assumption follows from the premises. It is also assumed by Gebharter (2017a) in his formulation of the exclusion argument.

$\{M_1, M_2\}$. Neither set is e-parent sufficient since they both omit the variable P_1, which is an e-parent of M_1 and P_2 when added to $\mathbf{V_1}$, and an e-parent of M_1 and M_2 when added to $\mathbf{V_2}$. So neither of the variable sets that the antireductionist uses to resist the exclusion argument satisfies the natural multi-level generalization of causal sufficiency. This means that fans of Kim's argument can resist the antireductionist's rebuttal by arguing that we should adopt a version of the WCP that is articulated in terms of e-parent sufficiency rather than causal sufficiency, and, in the process, block the conclusions that the antireductionist draws from $\mathbf{V_1}$ and $\mathbf{V_2}$.

According to this line of thought, the dispute between antireductionists and their opposition rests on a disagreement about what variable sets can be legitimately considered when investigating the causal structure of multi-level systems.[32] If, on the one hand, we require only that variable sets be causally sufficient, then the antireductionist is apparently able to demonstrate the causal efficacy of mental properties. However, if, on the other hand, we require satisfaction of the generalized e-parent sufficiency condition, then the CBN-theoretic vindication of the exclusion argument appears to go through successfully. Thus the crucial question is whether there is any principled reason to replace the standard causal sufficiency condition with the much stricter e-parent sufficiency condition.

To the extent that there is independent reason for non-reductive physicalists to countenance the existence of macro-level causal relations, there appears to be reason to side with causal sufficiency. Remember that Kim's argument is supposed to show that non-reductive physicalism has the *untoward* consequence of implying that macro-level properties are causally inert. If antireductionists can consistently avoid this untoward consequence by siding with causal sufficiency over e-parent sufficiency, then even Kim should agree that there is reason to side with causal sufficiency.

But the issue is complicated by the fact that not every non-reductive physicalist regards the consequence of macro-level epiphenomenalism as untoward. These philosophers' motivations differ from one to the next,[33] but they are typically in agreement that scientific practice should be revised so that higher-level causes are not countenanced. This dispute between the epiphenomenalist and the antireductionist is unfortunately not one that we can settle here. In order to do so, we would have to take stock of the many arguments offered on both sides, and there simply isn't enough space in this paper to do so effectively. What we *can* show, however, is that Kim's conditional conclusion — i.e., that if non-reductive physicalism is true, then macro-level properties are causally inert — is false when viewed through the lens of CBNs. This follows from what we've already demonstrated — i.e., that CBNs can be used to develop an account of multi-level causal inference that is consistent with both non-reductive physicalism and the causal efficacy of macro-level properties simply by siding with causal sufficiency over e-parent sufficiency. Thus by demonstrating the existence of

[32]Polger et al. (2018) make a similar point. Their response to Kim's argument is that mental variables and physical variables should not be in competition, and their argument turns on their observation that it's inappropriate for interventionists to consider variable sets that include P_1 and M_1 (or P_2 and M_2) under the assumption of the CMC. Here, we are more ecumenical about what variables *can* be considered alongside each other — indeed, we generalize the usual CBN framework partially to allow for consideration of variable sets that Polger et al. rightly deem problematic in the standard CBN setting — but we agree with Polger et al. that at least as far as CBNs are concerned, the debate turns on what variable sets should be considered appropriate for causal inference.

[33]For example, Gebharter (2017a) can be interpreted as championing epiphenomenalism on the grounds that it follows from his preferred view of causation in conjunction with the the thesis that mental properties supervene on micro-level physical properties, while Segal (2009) champions epiphenomenalism on the grounds that mental properties are dispositional, and dispositions are not causally efficacious.

this choice point in the development of interventionism, we've shown that not all non-reductive physicalist interventionist roads lead to epiphenomenalism, and have thereby revealed new paths to antireductionist interventionism.

Still, there are some extant arguments in the literature that *prima facie* seem to provide positive reason to side with e-parent sufficiency over causal sufficiency on the grounds that e-parent sufficiency correctly falsifies some causal generalizations that are in terms of non-projectible predicates (e.g., 'jade' in Kim's famous (1992) example),[34] while causal sufficiency problematically countenances these causal generalizations as true. In the next section, we show that there is yet another constraint on variable sets that shares e-parent sufficiency's ability to rule out these problematic causal generalizations without sharing e-parent sufficiency's vindication of full-blown macro-level epiphenomenalism.

6 Difference Maker Sufficiency

We expect that some philosophers will contend that there are real scientific contexts where it is of utmost importance that we attend to the way in which some macro-level property is realized, and that in these contexts, it is e-parent sufficiency that gets things right. For example, in Kim's (1992) example, chemists err if they study minerals in terms of what we once called 'jade' rather than in the more fine-grained terms of 'jadeite' and 'nephrite' (where 'jade' designates the disjunction of 'jadeite' and 'nephrite') because it matters whether a given instance of jade is actually jadeite or nephrite when it comes to predicting its chemical behavior.

In the causal inference literature, this same problem has reared its head in the so-called "cholesterol problem." Following Spirtes and Scheines (2004), suppose that high density cholesterol (HDC) causally inhibits (and is negatively correlated with) heart disease (D) and that low density cholesterol (LDC) causally promotes (and is positively correlated with) heart disease. Furthermore, let the variable TC ('total cholesterol') denote the sum of high and low density cholesterol, i.e. $TC = HDC + LDC$. Suppose also that in actual fact, TC is positively correlated with D. People with high total cholesterol have, on average, a significantly higher risk of heart disease. However, the specific make up of an individual's total cholesterol has a major impact on their risk of heart disease. Two individuals with the same total cholesterol can have very different risks of heart disease. For example, an individual with $HDC = 80$ and $LDC = 120$ will have a much higher risk than an individual with $HDC = 120$ and $LDC = 80$, even though they both have $TC = 200$.

In this context, the question of what happens when we intervene to set the value of total cholesterol to some particular value seems to have no determinate answer. If we increase an individual's total cholesterol by feeding them something that increases their high density cholesterol, then the intervention may reduce the risk of heart disease. But if we increase total cholesterol the same amount by feeding them something that increases their low density cholesterol, then their risk of heart disease will be significantly increased. The problem is that interventions on TC are too "fat-handed" for determinate inference. In this case, the realizers make a difference. And e-parent

[34]Kim (1992) famously argues that special science kinds are not autonomous from lower-level kinds on the grounds that generalizations about jade cannot be included in good science, and must instead be replaced with generalizations about jadeite and nephrite.

sufficiency redirects our attention towards these realizers. To see this, note that $\mathbf{V} = \{TC, D\}$ is not e-parent sufficient because its omits the common e-parent, $LDC \times HDC$.[35] So e-parent sufficiency forces us to consider the variable whose values are realize the values of TC. In contrast, causal sufficiency permits us to neglect TC's realizers (since $\mathbf{V} = \{TC, D\}$ *is* causally sufficient).[36] By the same token, e-parent sufficiency also forces us to consider whether a given instance of jade is jadeite or nephrite when considering its bearing, e.g., on whether it scratches a steel nail (because the variable that denotes whether something is jadeite, nephrite, or neither is a common e-parent of the variable that expresses whether something is jade and the variable that expresses whether something scratches a steel nail), while causal sufficiency allows one to ignore whether a given instance of jade is jadeite or nephrite when considering the same query (because the variable that denotes whether something is jadeite, nephrite, or neither is *not* a common cause of whether something is jade and whether something scratches a steel nail). Thus the cholesterol problem *prima facie* seems like grist for Kim's mill.

We submit that it's too early to claim victory for e-parent sufficiency because there is yet another constraint on variable sets that solves the cholesterol problem without implying full-blown macro-level epiphenomenalism. Recall that the motivation for regarding the cholesterol case as supporting e-parent sufficiency is that e-parent sufficiency forces us to attend to $LDC \times HDC$. This is *prima facie* desirable because the effect of TC on D varies with different realizations of TC. In probabilistic terms, this means that TC does not screen off its realizers from D.[37] An individual's total cholesterol may tell us something about their risk of heart disease, but we would gain significantly more information about that risk if we additionally knew the specific way in which their total cholesterol is realized in terms of low density cholesterol and high density cholesterol. If it were the case that TC screened off $LDC \times HDC$ from D, then the effect of interventions on TC would be perfectly well specified, since we could just choose an arbitrary realizing value of $LDC \times HDC$ for the value to which TC was set, and this arbitrary choice would make no difference to the probability distribution over D. So the intuition that we must attend to $LDC \times HDC$ over and above TC is reliant on the fact that TC does not screen off its supervenience base $LDC \times HDC$ from D. Similarly, the reason that 'jade' strikes us as a non-projectible chemical predicate is that it does not screen off its realizers from its chemical behav-

[35] Here, $LDC \times HDC$ denotes the Cartesian product of HDC and LDC. By definition, TC supervenes on this variable, and by stipulation this variable is correlated with D. Given MMC and the fact that TC doesn't screen off $LDC \times HDC$ from D, we can infer that $LDC \times HDC$ is (relative to this variable set) a direct cause of D.

[36] This variable set is typically treated as causally sufficient in the literature, but there could be reason to believe that it's not since, e.g., whether someone has a fatty diet is causally relevant to their total cholesterol and whether they get heart disease. Moreover, it may be that the cholesterol problem isn't so problematic when we attend to these common causes of TC and D since their presence in $\mathbf{V}$ implies constraints on what counts as a *bona fide* intervention on TC relative to $\mathbf{V}$, and thereby rules out some problematic "interventions". (For example, we cannot intervene to increase someone's total cholesterol by giving them bacon cheeseburgers because this is not independent from whether they have a fatty diet.) We are interested in pursuing this line of reasoning further, but abstain from doing so here for reasons of brevity. It's worth noting, though, that if we can make good on this line of reasoning — i.e., on using considerations of causal sufficiency to solve the cholesterol problem — then this only helps the antireductionist's case (since siding with causal sufficiency over e-parent sufficiency wins the debate for the antireductionist). Thus there is no reason to worry that we are pulling the wool over the reader's eyes by not pursuing this line of reasoning here.

[37] See Chalupka et al. (2017) and Woodward (2010) for a related diagnosis cholesterol-like problems.

ior — e.g., from whether the the steel nail gets scratched. This all motivates the following definition

Difference Maker (DM) Sufficiency: $\mathbf{V}$ is *DM sufficient* if and only if there do not exist any variables L, X, Y such that (i) $L \notin \mathbf{V}$, (ii) $X, Y \in \mathbf{V}$, (iii) L is a common e-parent of X and Y in an admissible graph over the extended variable set $\mathbf{V} \cup \{L\}$, and (iv) it is not the case that either X screens off L from every variable in $\mathbf{V} \setminus \{X, L\}$ or that Y screens off L from every variable in $\mathbf{V} \setminus \{Y, L\}$.

It is easily observed that DM sufficiency is a strictly weaker requirement than e-parent sufficiency. While e-parent sufficiency requires the inclusion of all common e-parents (since it just says that the first three sub-conditions must be satisfied), DM sufficiency requires only the inclusion of those common e-parents that satisfy the fourth sub-condition — roughly, those e-parents that are not screened off from everything in the variable set by one of their e-children.

To illustrate, let's apply DM sufficiency to the cholesterol problem. Consider again the set $\mathbf{V} = \{TC, D\}$ under the assumption that TC does not screen off D from $LDC \times HDC$. Since $LDC \times HDC$ is a common e-parent of TC and D (as observed above) and TC does not screen it off from D (or vice versa, by symmetry of independence), it follows that $LDC \times HDC$, TC, and D jointly fulfill the four conditions on L, X and Y in the definition of DM sufficiency. Thus, as desired, $\mathbf{V}$ is not DM sufficient in this case. Like e-parent sufficiency, DM sufficiency forces us to attend to TC's supervenience base, $LDC \times HDC$.

This means that DM sufficiency provides a natural weakening of e-parent sufficiency that is still capable of solving the cholesterol problem. Crucially, DM sufficiency does not require that we always include all common e-parents of the variables being considered, and thereby manages to vindicate some macro-level causal claims. It rather requires that we include just those common e-parents that *make a difference* in the sense that the effects of interventions on variables that supervene on those e-parents are underspecified.[38,39]

Now, there are two obvious criticisms that one could level at the DM sufficiency condition, which we reply to in turn. First, fans of causal sufficiency may worry that DM sufficiency is not *significantly* weaker than e-parent sufficiency, and that, like e-parent sufficiency, it still invalidates most of the causal claims of the special sciences. It is rarely the case that a variable completely screens off its supervenience bases from the variables to which it is (or seems to be) causally related in the real world. For example, it is economic orthodoxy that productivity causally influences

[38]Note that there is a strong affinity between the motivating intuitions behind our definition of DM sufficiency and the treatment of causal proportionality given by List and Menzies (2009). List and Menzies argue that higher level causes trump lower level causes whenever the effect is insensitive to variations in the lower level realizers of the prospective higher level cause. This condition is not fulfilled by the total cholesterol variable in the cholesterol problem, so advocates of List and Menzies' view would presumably acknowledge the need for a definition of sufficiency that does not entail the causal efficacy of total cholesterol. And like List and Menzies, our definition of DM sufficiency focuses on whether the effect is sensitive to the higher level cause's lower level realisers when the cause is held fixed.

[39]Observe that DM sufficiency renders the question of whether a variable's supervenience base must be included as dependent on what other variables are included in the variable set. For example, while we need to include the jadeite/nephrite variable when considering the effect of jade on whether the nail is scratched (because this makes a difference to whether the nail is scratched), we don't need to include it when we're exclusively concerned with the effect of jade on the price of jewelery in an economy whose participants are unaware of the difference between jadeite and nephrite.

economic growth. But productivity supervenes on both the number of hours worked and the value of the output of that work. And it seems implausible to claim that productivity completely screens off economic growth from these two factors. So when assessing the causal influence of productivity on growth, DM sufficiency will always require us to include an extra variable that corresponds to hours worked and the value of the output, and relative to this larger variable set, productivity will typically be represented as causally inert. Thus, like e-parent sufficiency, DM sufficiency *prima facie* seems to invalidate most macro-level causal claims. If this is right, then DM sufficiency is of limited use to enemies of epiphenomenalism.

We are sympathetic to this challenge, but it's important to note that it does not bear on the central argument of our paper. These considerations may push us towards causal sufficiency over DM sufficiency, but, again, the antireductionist interventionist's case is resolutely made if causal sufficiency is the appropriate constraint in multi-level contexts.[40] What matters for our argument is that it undercuts the alleged advantage that e-parent sufficiency has over its competitors by representing an option that aptly diagnoses problematic macro-level causal claims (e.g., involving jade and total cholesterol) as false without implying *full-blown* epiphenomenalism. Here, we submit that DM sufficiency *improves* upon e-parent sufficiency by providing us with a tool that distinguishes those cases where macro-level properties are not autonomous from those cases where they are.

Another prospective criticism of DM sufficiency is that it can lead one to omit common causes in the single variable setting and, in so doing, yield spurious causal inferences. To illustrate, consider three pairwise correlated variables, C, E_1, E_2 and suppose that (i) C is a common cause of E_1 and E_2, and (ii) C is not a difference maker for E_1/E_2, i.e. C is screened off from E_2 by E_1. Then the DM sufficiency condition will allow us to make causal inferences over the variable set $\mathbf{V}^- = \{E_1, E_2\}$ — i.e. it will allow us to omit the variable C from our consideration of the relationship between E_1 and E_2. And since E_1 and E_2 are correlated, we will be forced to infer a causal relationship between them. But it's perfectly possible that the correlation between E_1 and E_2 is due entirely to their sharing a common cause C, in which case the inferred relationship is spurious. So DM sufficiency is too weak insofar as it allows for the consideration of too many variable sets, and thereby licenses some of the spurious inferences that were prohibited by causal sufficiency.

In order to respond to this criticism, it will be instructive to briefly consider exactly what it means for C to count as a (direct) common cause of E_1 and E_2. A popular and intuitive formalization of this claim is that C counts as a direct cause of both E_1 and E_2 relative to the variable set $\mathbf{V} = \{C, E_1, E_2\}$, obtained by adding C to $\mathbf{V}^- = \{E_1, E_2\}$. But we've still not specified precisely what it means for one variable to count as direct cause relative to $\mathbf{V}$. There are two plausible options. On the first option, X is a direct cause of Y relative to $\mathbf{V}$ if X is a direct cause of Y relative to *every* admissible minimal graph over $\mathbf{V}$. On the second option, X is a direct cause of Y relative to $\mathbf{V}$ if X is a direct cause of Y relative to *some* admissible minimal graph over $\mathbf{V}$. Call the first condition the 'stringent' condition and the second condition the 'lax' condition.

Suppose that we adopt the stringent condition. Then C's being a common cause of E_1 and E_2 means that every admissible minimal graph over $\mathbf{V} = \{C, E_1, E_2\}$ represents C as a direct cause of E_1 and E_2. But we've stipulated that C is not a difference maker for E_1/E_2, meaning that E_1

[40] At this juncture, it's worth stressing that some violations of DM sufficiency may be worse than others. It's plausible that a given violation is problematic to the extent that the omitted supervenience base makes a difference. This scalar property can perhaps be analyzed in terms of degrees of correlation, but we leave this for a later date.

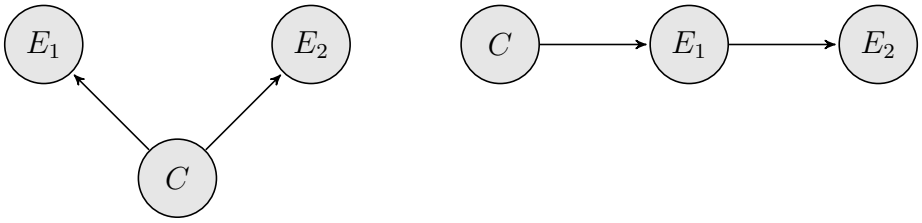

Figure 7: Minimal DAGs over $\{C, E_1, E_2\}$ when C is not a difference maker for E_1/E_2.

screens off C from E_2. This in turn entails that the right most DAG in Figure 7 is minimal, and hence that there exists an admissible minimal graph over $\mathbf{V} = \{C, E_1, E_2\}$ in which C is not a direct cause of E_2. Since we're adopting the stringent condition, this implies that C is not really a common cause of E_1 and E_2. So when we adopt the stringent condition, we automatically rule out the possibility of there existing variable sets which satisfy the DM sufficieny condition while leaving out common causes, and the criticism simply dissolves.

Now let's consider the lax condition. Given this understanding of 'direct cause', C's being a common cause of E_1 and E_2 means that there exists *some* admissible minimal graph over $\mathbf{V}$ in which C is represented as a common direct cause of E_1 and E_2. Thus the lax condition says that it is indeed the case that C counts as a common cause of E_1 and E_2, since the left DAG in Figure 7 represent C as a direct cause of both E_1 and E_2. But note also that the right most DAG in Figure 7 represents E_1 as a direct cause of E_2. The lax condition also implies that E_1 counts as a direct cause of E_2 relative to $\mathbf{V}$. This then suggests that the the inference that E_1 is causally related to E_2 is not spurious at all. So either one adopts the stringent condition, in which case DM sufficiency never leads one to leave out common causes in the first place, or one adopts the lax condition, in which case the causal verdicts that one obtains by leaving out common causes are all non-spurious. Either way, the problem is defused.

How does this all bear on the status of the exclusion argument? As before, we adopt the WCP in favor of SCP. This means that in order to establish the possible causal efficacy of M_1, we only need to find one DM sufficient variable set $\mathbf{V}$ such that M_1 is represented as causally efficacious in some minimal graph over $\mathbf{V}$. As before, a salient option here is $\mathbf{V} = \{M_1, M_2\}$, i.e. the variable set that omits the physical realizers of the relevant mental states. Whether this set is DM sufficient depends on whether M_1 screens off P_1 from M_2. And whether that is the case, or approximately the case, is a substantive empirical question that probably cannot be settled from the armchair. Either way, there doesn't seem to any *a priori* reason to rule out the *possibility* that mental variables like M_1 sometimes screen off their realizers from other mental variables like M_2, which means that we cannot rule out the possibility that $\mathbf{V}$ is DM sufficient. And since M_1 is represented as a cause of M_2 in a minimal graph over $\mathbf{V}$, advocates of DM sufficiency can't rule out the possibility of mental causation *a priori*. Thus like causal sufficiency, DM sufficiency allows for the construction of an antireductionist theory of multi-level causal inference.

7 Conclusion

It's time to take stock. In the first few sections of the paper, we considered a natural generalization of the CBN framework to the multi-level setting, and concluded that the soundness of Gebharter's CBN-theoretic formulation of the causal exclusion argument hinges on the question of which variable sets can legitimately be considered in the context of multi-level causal inference. Then, we argued that on some plausible ways of constraining variable sets in multi-level settings, the causal exclusion argument does *not* go through successfully. More specifically, opting for causal sufficiency as a constraint results in the full-blown vindication of antireductionism, while opting for DM sufficiency vindicates a more moderate brand of antireductionism.

To see the picture more clearly, it may be helpful to consider its application to Putnam's (1975) peg. In order to vindicate the autonomous explanatory power of macro-level properties, Putnam famously considers the example of a wooden board containing two holes. The first hole is circular with a diameter of 1 inch and the second hole is square with a length of 1 inch per side. A cubical peg that is 15/16 of an inch on each side will fit through the second hole, but not the first. According to Putnam, this is explained entirely by the macro-level properties of the peg and the holes, as described above. From an explanatory perspective, the micro-level properties of the peg are redundant. Once you know the macro-level properties, the micro has no further role to play.

It is exactly this screening off property (referenced in DM sufficiency) that accounts for the fact that nothing is gained from attending to the micro-details in this case. That is, because the nature of the peg's realization base makes no difference for the effect of its size on whether it will fit, DM sufficiency says that it's fine to ignore the peg's realization base. But in other examples (e.g., the cholesterol case or Kim's (1992) jade example), the macro-level property's effects vary with how it is realized, and DM sufficiency correspondingly requires that we include the realization base. If we side with DM sufficiency over causal sufficiency, then we get a nice picture of why Putnam's peg is a legitimate causal kind and why Kim's jade is not — i.e., DM sufficiency requires that you attend to whether the jade is jadeite or nephrite (and in the process repudiates the legitimacy of 'jade' as a cause), while it does not require that you attend to the realizer of Putnam's peg (and thus does not repudiate the legitimacy of Putnam's peg as a causal kind).

Of course, as we mentioned earlier, if we side with DM sufficiency, then we condemn a great many causal claims that are part and parcel of the special sciences (since the micro-details often do make *some* difference). Insofar as there is reason to square with this practice, there may be reason to side with causal sufficiency. But DM sufficiency does have the decided advantage over causal suffiency of being able to classify some genuinely problematic higher-level causal generalizations as false. Either way — i.e., no matter whether you prefer causal sufficiency or DM sufficiency — there is an open path to antireductionist interventionism. The question is just *how* antireductionist this interventionism should be.

References

Baker, L.R. (1993). *Metaphysics and Mental Causation.* In Heil and Mele eds, *Mental Causation.* Oxford: Clarendon Press. 75–95.

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

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

Eva, B. and Stern, R. (forthcoming). *Causal Explanatory Power. British Journal for the Philosophy of Science.*

Forster. M, Raskutti, G., Stern, R., and Weinberger, N. (2018). *Frugal Inference of Causal Relations. British Journal for the Philosophy of Science* 69(3): 821–848.

Gebharter, A., and Schurz, G. (2014). *How Occam's razor provides a neat definition of direct causation.* In J. M. Mooij, D. Janzing, J. Peters, T. Claassen, and A. Hyttinen (Eds.), *Proceedings of the UAI workshop Causal Inference: Learning and Prediction. Aachen.*

Gebharter, A. (2017a). *Causal exclusion and causal Bayes nets. Philosophy and Phenomenological Research*, 95(2), 353-375.

Gebharter, A. (2017b). *Causal exclusion without physical completeness and no overdetermination. Abstracta Linguagem, Mente E Acao*, 10, 3-14.

Gebharter, A. (2017c). *Causal nets, interventionism, and mechanisms.* Springer. http://doi.org/10.1007/978-3-319-49908-6

Hausman, D. (1998). *Causal Asymmetries.* Cambridge: Cambridge University Press.

Hausman, D. and Woodward, J. (1999). *Independence, Invariance and the Causal Markov Condition. British Journal for the Philosophy of Science*, 50(4): 521-583.

Hendry, R.F. (2006). *Is There Downwards Causation in Chemistry?.* In D. Baird, E. Scerri, and L. McIntyre eds, *Philosophy of Chemistry: Synthesis of a New Discipline; Boston Studies in the Philosophy and History of Science* 242: 173–89.

Hesslow, G. (1976). *Discussion: Two Notes on the Probabilistic Approach to Causality. Philosophy of Science*, 43: 290-292.

Hitchcock, C. (2012). *Theories of Causation and the Causal Exclusion Argument. Journal of Consciousness Studies*, 19 (5-6): 40-56.

Hitchcock, C and Woodward, J. (2003). *Explanatory Generalizations, Part 2: Plumbing Explanatory Depth. Nous*, 37(2): 181-199.

Kim, J. (1992): *Multiple Realizability and The Metaphysics of Reduction. Philosophy and Phenomenological Research*, 52(1): 1-26.

Kim, J. (1989): *Mechanism, purpose, and explanatory exclusion. Philosophical Perspectives*, 3: 77-108.

Kim, J. (2000): *Mind in a Physical World.* MIT Press.

Kim, J. (2003): *Blocking causal drainage and other maintenance chores with mental causation. Philosophy and Phenomenological Research*, 67(1): 151-176.

Kim, J. (2005): *Physicalism, or Something Near Enough.* Princeton University Press.

Lewis, D. (1986): *Causal Explanation.* In *Philosophical Papers: Volume 2*: 214-240. Oxford: Oxford University Press

List, C. and Menzies, P. (2009). *Nonreductive Physicalism and the Limits of the Exclusion Principle. Journal of Philosophy* 106(9): 475-502.

Pearl, J. (1988): *Probabilsitic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann.

Pearl, J. (2009): *Causality: Models, Reasoning and Inference*, 2nd edition. 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.

Putnam, H. (1975). *Philosophy and our mental life.* In *Mind, Language, and Reality*, 291?303. London: Cambridge University Press.

Raatikainen, P. (2010). *Causation, Exclusion and the Special Sciences. Erkenntnis*, 73(3): 349-363

Reichenbach, H. (1956). *The Direction of Time.* Berkeley: University of Los Angeles Press.

Schaffer, J. (2016). *Grounding in the Image of Causation. Philosophical Studies*, 173: 49-100

Schupbach, J. and Sprenger, J. (2011). *The Logic of Explanatory Power. Philosophy of Science* 78(1): 105-127

Segal, G. (2009). *The Causal Inefficacy of Content.* Mind and Language, 24: 80–102

Shapiro, L. (2010). *Lessons From Causal Exclusion. Philosophy and Phenomenological Research* 81(3): 594-604

Shapiro, L. and Sober, E. (2007). *Epiphenomenalism – The Do's and Dont's.* In G. Wolters and P. Machamer (eds.), *Thinking about Causes: From Greek Philosophy to Modern Physics.* Pittsburgh: University of Pittsburgh Press: 253-264.

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

Stapp, H. (2005). *Quantum Interactive Dualism: An Alternative to Materialism. Journal of Consciousness Studies* 12: 43–58.

Stern, R. (forthcoming). *Causal Concepts and Temporal Ordering. Synthese.*

Weslake, B. (2011). *Exclusion Excluded. International Studies in the Philosophy of Science.*

Woodward, J. and Hitchcock, C. (2003). *Explanatory Generalizations, Part 1: A Counterfactual Approach. Nous*, 37(1): 1-24.

Woodward, J. (2005). *Making Things Happen: A Theory of Causal Explanation*, Oxford Studies in the Philosophy of Science. Oxford: Oxford University Press.

Woodward, J. (2008a). *Mental Causation and Neural Mechanisms.* In H. Price and J. Kallestrup (eds.), *Causation, Physics and the Constitution of Reality.* Oxford: Oxford University Press: 66-105.

Woodward, J. (2008b). *Response to Strevens.* Philosophy and Phenomenological Research, 77(1), 193-212.

Woodward, J. (2010). *Causation in biology: stability, specificity, and the choice of levels of explanation. Biology & Philosophy*, 25(3): 287-318.

Woodward, J. (2015). *Interventionism and Causal Exclusion. Philosophy and Phenomenological Research*, 91(2): 303-347.

Zhang, J. and Spirtes, P. (2011). *Intervention, determinism, and the causal minimality condition. Synthese*, 182(3): 335-347.