Philosophical Studies (2023) 180:2685–2706
https://doi.org/10.1007/s11098-023-02001-w

The causal efficacy of composites: a dilemma for interventionism

Thomas Blanchard[1]

Accepted: 2 June 2023 / Published online: 7 July 2023

Abstract

This paper argues that the interventionist account of causation faces a dilemma concerning macroscopic causation – i.e., causation by composite objects. Interventionism must either require interventions on a composite object to hold the behavior of its parts fixed, or allow such interventions to vary the behavior of those parts. The first option runs the risk of making wholes causally excluded by their parts, while the second runs the risk of mistakenly ascribing to wholes causal abilities that belong to their parts only. Using as starting point Baumgartner's well-known argument that interventionism leads to causal exclusion of multiply realized properties, I first show that a similar interventionist exclusion argument can be mounted against the causal efficacy of composite objects. I then show that Woodward's (2015) updated interventionist account (explicitly designed to address exclusion worries) avoids this problem, but runs into an opposite issue of over-inclusion: it grants to composites causal abilities that belong to their parts only. Finally, I examine two other interventionist accounts designed to address Baumgartner's argument, and show that when applied to composites, they too fall on one horn (exclusion) or the other (over-inclusion) of the dilemma. I conclude that the dilemma constitutes an open and difficult issue for interventionism.

Keywords Causal exclusion · Causal models · Interventionism · Parts and wholes

In recent years, a lively debate has unfolded concerning the implications of the interventionist account of causation for Kim's exclusion argument. While many authors have appealed to interventionism to try to defuse Kim's argument, others (most

✉ Thomas Blanchard
tblancha@uni-koeln.de

[1] University of Cologne, Cologne, Germany

notably Michael Baumgartner[1]) have insisted that interventionism in fact supports Kim's contention that mental and other special-scientific properties must (if distinct from physical properties) be causally impotent. Baumgartner's argument is that if high-level properties are distinct from their physical realizers, an intervention on the former would have to hold the latter fixed, which is impossible. Thus, on interventionism, high-level properties cannot be wiggled by interventions and hence are causally inefficacious. In response, a number of interventionists have sought to clarify or update the interventionist framework so as to avoid the charge of exclusion. The most influential proposal, due to Woodward (2015), salvages the causal efficacy of mental properties by allowing interventions to vary the supervenience base of their targets.

In this paper, I argue that the exclusion worry uncovered by Baumgartner generalizes to threaten the causal efficacy of all macroscopic properties – i.e., properties of composite objects. This is because reasoning similar to Baumgartner's suggests that a proper intervention on a composite object would have to hold fixed the microscopic properties of its parts and hence is impossible. Furthermore, I show that Woodward's (2015) account avoids this generalized problem of exclusion only by falling into the opposite problem of granting to certain composite objects causal abilities that belong to their parts only. The upshot is that when it comes to macro-causation, interventionism faces a dilemma between *exclusion* and *over-inclusion.*[2] I also consider two other interventionist lines of response to Baumgartner's argument besides Woodward's, and show that they also fall prey to one horn or the other of the dilemma. I conclude that the dilemma constitutes an open and difficult issue for interventionism.

1 Interventionism, non-reductive physicalism, and causal exclusion of mental properties

Let me start with a brief summary of Kim's causal exclusion argument against non-reductive physicalism (see Kim, 1998, 2005). Suppose that M_1 and M_2 are mental properties such that M_2 is a putative effect of M_1. P_1 and P_2 are the physical supervenience bases of M_1 and M_2 respectively. (See Fig. 1, where black arrows represent supervenience relations, while blue arrows represent causal relations.) In line with non-reductive physicalism, we assume that M_1 and M_2 are multiply realizable at the

$$\begin{tikzcd}
M_1 \arrow[rr, dashed, "?"] & & M_2 \\
P_1 \arrow[u] \arrow[rr] & & P_2 \arrow[u]
\end{tikzcd}$$

Fig. 1 The setup of Kim's exclusion argument

[1] (Baumgartner, 2009, 2013, 2018). See also (Hoffman-Kolss, 2014; Gebharter, 2017).

[2] I owe the term "over-inclusion" to Toby Friend.

physical level and hence distinct from their physical supervenience bases. We also assume that P_1 is causally sufficient for P_2, in accordance with the principle of the causal closure of the physical. Kim argues that given these assumptions, M_1 cannot in fact cause M_2. Because M_2 supervenes on P_2, the fact that P_1 is causally sufficient for P_2 means that it is causally sufficient for M_2 as well. Hence, there is no causal work left for the distinct property M_1 to do in bringing about M_2. Recognizing M_1 as an additional cause would be to admit that M_2 - and more generally every effect of a mental property – is overdetermined, a view that Kim regards as too implausible to be believed. Kim concludes that unless mental properties can be reduced to physical properties, they must be causally impotent.

While several authors (e.g. Shapiro and Sober, 2007; Raatikainen, 2010) have appealed to Woodward's (2003) interventionist account of causation to try to defuse Kim's argument, Baumgartner has argued that Woodward's account actually supports Kim's view (Baumgartner, 2009, 2013, 2018). The intuitive idea behind interventionism is that a cause makes a difference to its effect insofar as intervening on the cause would change the effect. Woodward (2003) defines several causal notions based on this idea, including that of *direct cause* and *contributing cause*:

> "(M) A necessary and sufficient condition for X to be a (type-level) direct cause of Y with respect to a variable set **V** is that there be a possible intervention on X that will change Y or the probability distribution of Y when one holds fixed at some value all other variables Z_i in **V**. A necessary and sufficient condition for X to be a (type-level) *contributing cause* of Y with respect to variable set **V** is that (i) there be a directed path from X to Y such that each link in this path is a direct causal relationship; that is, a set of variables $Z_1 \ldots Z_n$ such that X is a direct cause of Z_1 which is in turn a direct cause of Z_2, which is a direct cause of $\ldots Z_n$, which is a direct cause of Y, and that (ii) there be some intervention on X that will change Y when all other variables in **V** that are not on this path are fixed at some value." (Woodward 2003: 59)

While direct and contributing causation so-defined are relative to a variable set, Woodward (2008, 209) also provides a de-relativized notion of contributing causation on which X is a contributing cause of Y *simpliciter* iff there exists a variable set in which X counts as a direct or contributing cause of Y according to (M). Woodward (2003) also offers the following definition of an intervention variable:

> "(IV) I is an intervention variable on X with respect to Y iff
> I.1. I causes X.
> I.2. I acts as a switch for all the other variables that cause X. (…)
> I.3. Any directed path from I to Y goes through X.
> I.4. I is (statistically) independent of any variable Z that causes Y and that is on a directed path that does not go through X." (2003: 98)

An intervention can then be defined as a value of I that causes X to take a specific value. For our purposes, the key parts of (IV) are I.3 and I.4. To see the motivation for them, suppose that a manipulation I of X causes Y in a way that bypasses X, or is correlated with a common cause of X and Y. In that setup, we would find that I is correlated with a change in Y even if X doesn't cause Y. To be plausible, the interventionist account must therefore prevent confounded manipulations of this kind from counting as legitimate intervention variables. This is what conditions I.3 and I.4 do. According to Baumgartner, however, these conditions also provide the basis of a sui generis interventionist exclusion argument against non-reductive physicalism.

To see why, let's return to Fig. 1, and let's now read the capital letters in it as variables. Specifically, let us suppose that M_1 and M_2 are binary variables that take value 1 if the relevant mental properties are instantiated and 0 otherwise. And let us suppose that P_1 is a many-valued variable representing possible physical realizers of M_1. Each possible physical realizer of $M_1=1$ and $M_1=0$ is represented by a specific value of P_1. Likewise, P_2 represents possible physical realizers of M_2. We assume that M_1 supervenes on P_1 and M_2 on P_2, as represented by the black arrows in Fig. 1. (I follow Hoffman-Kolss (2021) in defining supervenience between variables as follows: a variable X supervenes on a set of variables $\mathbf{Z}$ just in case for every value of X, there is a possible assignment of values to members of $\mathbf{Z}$ that necessitates the value of X.)

Now, note that according to (M), M_1 causes M_2 in Fig. 1 just in case M_2 would change under some intervention on M_1 (with respect to M_2). But since there is a directed path from P_1 to M_2 (via P_2) that does not go through M_1, I.4 entails that a proper intervention on M_1 should be statistically independent of P_1. But since M_1 supervenes on P_1, *any* process changing M_1's value must also change P_1's value, and hence must also violate I.4. In addition, any such process plausibly counts as a cause of P_1 and is thus connected to M_2 via a directed path that doesn't go through M_1, so that I.3 is also violated. Thus, a proper intervention on M_1 with respect to M_2 is impossible. Moreover, because (IV) is not relativized to a variable set, it entails that an intervention on M_1 should be independent of P_1 even if the latter is omitted from the variable set. So there is no variable set in which M_1 and M_2 satisfy (M), and hence M_1 cannot be a contributing cause of M_2*simpliciter*. In short, M_1 is causally excluded by its supervenience base, whose presence makes it impossible to properly intervene on M_1 with respect to M_2 in a non-confounded manner.

This reasoning, note, works only if M_1 is distinct from its supervenience base. (If M_1 were reducible to P_1, then they would be the same variable, and a proper intervention on the former would not have to hold the latter fixed.) It thus exposes a tension between interventionism and the non-reductive physicalist view that mental properties are both causally efficacious and irreducible to their physical supervenience bases– a tension that should worry virtually all interventionists, who are either committed to non-reductive physicalism or have no wish to rule out its truth *a priori*. To solve this tension, the key task for interventionists is to show how to spell out the key ideas of their framework in a way that avoids exclusion, and makes it possible for mental properties to remain efficacious even if distinct from their realizers. I will discuss several existing attempts to meet this challenge later on. Beforehand, however, I want to show that the interventionist exclusion worry identified by Baumgartner generalizes to threaten all macrocausation.

2 The interventionist exclusion problem generalized

To clarify this claim, it is helpful to distinguish two ways in which exclusion worries may generalize beyond mental properties. One issue is whether those worries also apply to other special-scientific properties that non-reductive physicalists regard as supervenient on but distinct from physical properties. It is clear that Kim's and Baumgartner's arguments generalize in this way, since nothing in them depends on M_1 being a mental rather than (say) a biological or social property – all that matters is that it be distinct from its physical supervenience base P_1. Another issue is whether exclusion worries generalize to threaten all *macroscopic* properties – i.e., properties of composite objects. The setup of both Kim's and Baumgartner's arguments in fact presupposes the existence of macroscopic causation, as both assume causal efficacy of P_1, which is meant to be or represent a physical property of a composite object (the brain instantiating it). But a number of authors have argued that Kim's argument in fact eliminates macrocausation by making all properties of composite objects (including physical ones such as P_1) causally excluded by their microscopic parts (see e.g. Bontly, 2002 and Block, 2003). Here I want to show that the interventionist exclusion worry identified by Baumgartner also generalizes in this way.

To see why, consider the (macrophysical) causal relationship between firing a gun at close range and the target of the shooting dying. Let F take value 1 if the gun is fired (0 otherwise), and V take value 1 if the victim dies (0 otherwise). In addition, let S be a many-valued variable representing the various possible microphysical states of the victim's body at time of death (one value for each possible state). Finally, let us assume the gun is composed of n elementary particles at the time of firing, and let $A_1 \ldots A_n$ represent possible physical states of those particles: that is, each A_i is a many-valued variable whose values represent possible fundamental physical states of the ith particle composing the gun at the time of firing. I leave it open which physical properties exactly the A_is represent, as this does not matter for the argument – think of them as whatever properties our best physical theory posits as fundamental properties of elementary particles.

I assume that the following four conditions hold in this situation (see Fig. 2). First, the microphysical state of the victim's body fixes whether she lives or dies, so that V supervenes on S (hence the black arrow $V \rightarrow S$ in Fig. 2). Second, the states of the particles taken collectively fix whether the gun fires – that is, F supervenes on the set of the A_i variables, so that F cannot change in value without at least some of the A_is

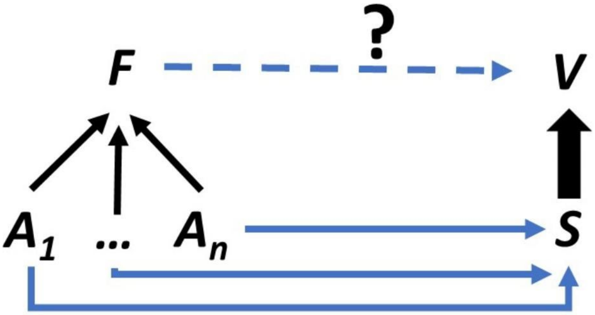

Fig. 2 The gun example

changing value as well. (The black arrows from each A_i to F in Fig. 2 represents the fact that each A_i is in the supervenience base of F. These arrows are thinner than the one from S to V, to indicate that F does not supervene on any individual A_i considered in isolation, but only collectively.) Third, I assume that the relevant supervenience relationships are of a non-reductive kind, i.e. that macroscopic variables are distinct from their microscopic supervenience bases (more on this assumption below). Fourth, I assume (though nothing crucial hangs on this) that each A_i is a direct cause of S. This seems reasonable, since the laws of physics dictate that the physical state of each particle composing the gun influences the exact microstate of the victim's body at time of death[3], and that influence does not seem mediated by any other variable in the graph.

But now, using Baumgartner's logic, we can show that read literally (M) and (IV) together entail that F cannot cause V. Given that each A_i lies on a directed path to V (via S) that does not go through F, I.4 entails that a proper intervention on F with respect to V should be statistically independent of each A_i. But since F supervenes on the A_i variables, any process changing F's value must also change the value of at least some of the A_is. In addition, each such process plausibly counts as a cause of at least some of the A_is and is thus connected to V via directed paths that bypass F, so that I.3 is also violated. Thus, a proper intervention on F with respect to V is impossible. Moreover, because (IV) is not relativized to a variable set, it entails that an intervention on F with respect to V must be independent of each A_i even if the latter are omitted from the variable set. So there is no variable set with respect to which interventions on F with respect to V are possible, hence no variable set in which F and V satisfy (M). So the former cannot be a contributing cause of the latter *simpliciter*. In short, F is causally excluded by the A_is, whose presence makes it impossible to properly intervene on F in a non-confounded manner. Note that the fact that *all* A_is are direct causes of S is not crucial to the argument. For suppose only some of them are. Then any change in F associated with a change in V would have to change the state of at least some of the particles causally relevant to S. So any such change would fail to count as an intervention by the lights of (IV), and F would still come out as causally irrelevant to S.

As should be clear, there is nothing special about the gun example: the same reasoning can be used to threaten the causal efficacy of any property of a composite object. For any such property X, at least some of the object's parts will have properties that are directly causally relevant to X's putative effect Y (or the microphysical states on which Y supervenes). Moreover, those microscopic properties will be included in the supervenience base for X, so that any intervention on X that changes Y would have to change at least some of those microscopic properties, which is prohibited by I.4. The upshot is that all properties of composite objects are causally excluded, even macrophysical properties. (Consider for instance P in Fig. 1. Although the setup of Baumgartner's original argument presupposes that P is causally efficacious, the logic

[3] Even atoms in (e.g.) the gun's handle, which may at first glance not seem connected to S, do exert a minute gravitational influence on the atoms in the victims' body and hence on the value taken by S.

of the argument implies that in reality P is made causally impotent by the properties of the particles that compose my brain.[4])

A number of remarks are in order here. First, as noted above, several authors have argued that Kim's argument also generalizes to threaten all macrocausation (see e.g. Bontly, 2002 and Block, 2003). This is generally taken as an objection against Kim's argument – an indication that it proves too much.[5] (Kim himself takes it that way, and has thus sought to show that his argument does not in fact generalize: see (Kim, 2003).) Here the dialectical upshot is different. Baumgartner's original argument raises a worrisome prospect for interventionists – the prospect that their preferred theory of causation may turn out to be incompatible with non-reductive physicalism (a view that many interventionists regard as plausible, or at least not to be ruled out on conceptual grounds). Far from defusing this worry, the generalization just presented only deepens it, by raising the threat that interventionism may be unable to make sense of macrocausation generally.

Second, the problem under consideration is similar to one that has been discussed by several authors in the literature on mechanistic explanation, including by Baumgartner himself (Baumgartner & Gebharter, 2016; Baumgartner & Casini, 2017; see also Eronen and Brooks, 2014 and Romero, 2015). An important task in this literature is to explain what it is for a part (or component) to be constitutively relevant to the mechanism for a whole's behavior. What is it, for instance, for sodium channels to be constitutively relevant to the neuron's action potential? According to Craver's (2007) influential mutual manipulability account, a part is relevant to the behavior of a whole when an intervention on the part would change the whole, and an intervention on the whole would change the part. In response, the authors just cited have argued that this account is in fact incompatible with the Woodwardian framework on which it relies. In particular, they have shown that any process that changes both the behavior of a whole X and the behavior of a part Y must by the lights of (M) and (IV) count as a direct cause of *both* X and Y. Given I.3, this means that interventions on wholes with respect to their parts are impossible. The argument just presented uncovers a related but distinct aspect in which interventions on wholes are conceptually problematic: viz., (IV)-interventions on wholes *with respect to their putative effects* are impossible. And it shows that the inapplicability of (IV) in contexts involving part-whole relationships raises deep issues not only for Craver's account of constitutive relevance (which interventionists themselves need

[4] Note that to get the result that *every* property of composite objects is causally excluded, it is best to represent the state of each part individually and separately. In the gun example, suppose that instead of using a separate variable for the state of each particle, we were to collapse them into a single many-valued variable A, each value of which represents a possible assignment of physical states to all of the particles composing the gun. One could then still run the argument that F is causally excluded by postulating A as a direct cause of V. Yet it could be argued that A still represents a physical property of the gun itself (one that describes its exact physical state), and hence that this way of setting things up still grants causal efficacy to some property of the gun. Representing the properties of each particle with a separate variable avoids this objection.

[5] One exception is Merricks (2001), who endorses the view that (non-living) composite objects are indeed causally excluded by their parts. See Yang (2013) for an interventionist response to Merricks. Yang's paper is an instance of a general interventionist strategy for addressing exclusion worries that I discuss in Sect. 4 below.

not endorse[6]), but for the interventionist account of causation itself, by threatening to make the account unable to uphold the causal efficacy of wholes generally.

Third, as noted above, the interventionist exclusion problem presented by Baumgartner arises only on the assumption that mental variables are irreducible to their physical supervenience bases. Similarly, the generalization just presented works only on the assumption that macroscopic variables are distinct from their microscopic supervenience base. If, in the gun example, F were identical to the set of A_i variables, interventions on the former would obviously not be required to hold the latter fixed. But the standard multiple realizability considerations that support the autonomy of mental variables also support the autonomy of macroscopic variables. Just like the mental property represented by M_1 can be realized by several values of P_1, the gun firing ($F=1$) can be realized (or constituted[7]) by several combinations of values of $A_1,\ldots, A_n$. And plausibly the same is true of every other macroscopic property.

The argument just presented raises a challenge for interventionists similar to the one raised by Baumgartner's original argument – the challenge of spelling out the interventionist framework in a way that avoids the charge of exclusion and vindicates the causal efficacy of composite entities. In what follows, I will consider various updated interventionist accounts that have been formulated in response to Baumgartner's argument, and examine whether they avoid the charge of exclusion for composite objects. I will start by looking at the most influential such account, which is due to Woodward (2015).

3 Woodward's (2015) updated interventionist account

In response to Baumgartner, Woodward (2015) offers updated versions of (IV) and (M) that contain explicit exception clauses for supervenience bases. More precisely, he proposes a new definition of an intervention variable on X with respect to Y (IV*) on which I.3 and I.4 in (IV) are replaced by the following conditions:

> I.3*. Any directed path from I to Y goes through X, *or some variable in the supervenience base of* X.
> I.4*. I is (statistically) independent of any variable Z that causes Y and that is on a directed path that does not go through X, *unless* Z *is in the supervenience base of* X.

Similarly, his updated version of (M) makes it explicit that off-path variables in the supervenience base of X should not be held fixed when assessing contributing causation:

[6] Woodward (2015), for instance, seems skeptical of the project of spelling out constitutive relevance and other non-causal dependence relations in terms of interventions.

[7] The exact nature of the relationship is a matter of controversy. Some authors (e.g. Craver, 2007) think of realization as a relation between properties of the same object, and prefer to speak of constitution to designate the relationship between a whole's properties and properties of its parts, while others (e.g. Gillett, 2013) conceive of constitutive relationships between parts and wholes in terms of realization.

> (M*) A necessary and sufficient condition for X to be a (type-level) direct cause of Y with respect to a variable set **V** is that there be a possible intervention on X that will change Y or the probability distribution of Y when one holds fixed at some value all other variables Z_i in **V** *that are not in the supervenience base of* X. A necessary and sufficient condition for X to be a (type-level) contributing cause of Y with respect to variable set **V** is that (i) there be a directed path from X to Y and that (ii) there be some intervention on X that will change Y when all other variables in **V** that are not on this path *and are not in the supervenience base of* X are fixed at some value.[8]

The notion of `supervenience base' can be understood as follows: Z is in the supervenience base of X just in case Z is part of a set of variables **V** such that (a) X supervenes on **V** and (b) no proper subset of **V** satisfies (a). Woodward argues that the exception clauses built into these definitions are well-motivated, since it is not part of scientific practice to hold supervenience bases when assessing causation.

Clearly, on this updated interventionist account, the causal exclusion worries discussed previously disappear. Returning to Kim's scenario in Fig. 1, (IV*) makes interventions on M_1 possible again by allowing them to vary the value of P_1. Because such interventions would also change the value of M_2, (M*) counts M_1 as a cause of M_2 in Fig. 1, and hence also *simpliciter*. So mental and other non-physical properties are not causally excluded by their realizers anymore. Likewise, the threat of causal exclusion of wholes by their parts disappear. Because properties of wholes supervene on the properties of their parts, (IV*) does not subject interventions to the impossible demand of holding the parts fixed while varying the whole, so that causation by composites is possible after all. As an illustration, in the gun example, the A_i variables collectively form a supervenience base for F. Accordingly, (IV*) allows an intervention on F (with respect to V) to also change the value of some or all of the A_is. (M*) thus validates the intuitive claim that F causes V, as V would change in value under some such interventions.

However, Woodward's updated account only avoids exclusion of composites by their parts by running into an opposite problem of *over-inclusion*. Because it allows interventions on a whole to vary the behavior of its parts, the account sometimes mistakenly counts a whole as the cause of an effect that is really due to one of its parts only. As an illustration, consider the following case:

> *Judge*. A court of law is composed of three judges, all of whom must vote to convict for a defendant to be officially declared guilty by the court. (The court's verdict thus non-causally depends on each of the three votes.) When the first judge votes to convict, she gets nervous as a result, out of fear of having possibly contributed to convicting an innocent. On a given occasion, all three judges vote to convict the defendant.

[8] Woodward does not give explicit statements of IV* and M* but merely indicates their main characteristics. My formulation follows Baumgartner's (2010: 378; 2014: 13 – 4) reconstruction of these definitions, with slightly amended wording.

Intuitively, while the vote of the first judge is a cause of whether she gets nervous, the verdict as a whole isn't. (To bolster that intuition, you may suppose that the first judge does not even know whether the defendant has been found guilty. Perhaps each judge is being kept unaware of the verdict reached by the court as a whole.) The causal structure of the case is represented in Fig. 3, where C represents the verdict of the court (1 for guilt, 0 for innocence), V_1, V_2 and V_3 represent respectively how the first, second and third judges vote (1 to convict, 0 to acquit), and N represents whether the first judge gets nervous (1 if she does, 0 otherwise).

Note that C supervenes on the set $\{V_1, V_2, V_3\}$, as indicated by the black arrows in Fig. 3, so that any process changing C's value must change the value of one or more of these three variables. Of course, on (IV*), this does not make interventions on C impossible. But now, note that some (IV*)-interventions on C are associated with changes in N – namely, all those interventions that change the court's verdict partly or solely by changing V_1. And this is enough for (M*) to count C as a cause of N in Fig. 3, and hence also *simpliciter*. Thus Woodward's updated account mistakenly counts a whole (the court) as the cause of an effect (the nervousness) that is really due to one of its parts only (the first judge). Or, to put it in terms of properties: it makes a property of the whole causally responsible for an effect that is properly due to a property of one of its parts.

Thus, the dilemma faced by interventionism now comes into view. On an interventionist account, one must either require an intervention on a composite to hold its parts fixed, or allow interventions on a composite to vary its parts. The first option runs the risk of making composites causally inefficient, while the second is in danger of ascribing to wholes causal abilities that belong to their parts only. Without an exception clause for supervenience bases in the definition of interventions, interventionism falls on the first horn; but building such an exception clause into the definition, as Woodward (2015) does, yields an account that falls on the second. In the remainder of the paper, I will examine two other updated versions of interventionism that have been offered in reaction to Baumgartner's argument, and show that they also fall on one horn or the other of this dilemma.

Beforehand, however, let me briefly mention one possible way for interventionists to deal with cases such as *Judge*. This would be to supplement Woodward's (2015) account with a *proportionality* requirement on causation. Roughly, proportionality says that a cause should be commensurate to its effect, i.e. just specific enough to account for the effect, but no more specific than that. Proportionality implies in par-

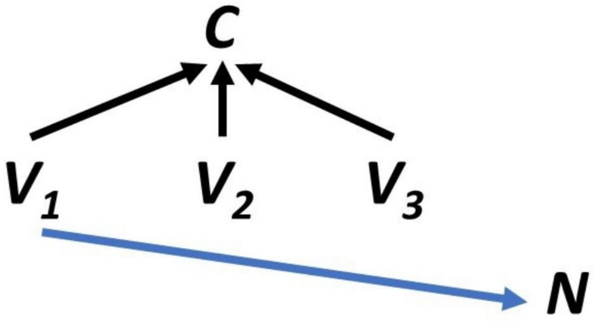

Fig. 3 The causal structure of *Judge*

ticular that the cause should not be "characterized in such a way that alternative states of it *fail* to be associated with changes in the effect" (Woodward, 2010: 298). The *C*-*N* relationship in *Judge* does not satisfy that condition, as there are some changes in *C* that are not associated with changes in *N* (namely changes in *C* due solely to the second and/or third judges changing their votes). Thus if we posit proportionality as a necessary condition on causation, as some authors (Yablo, 1992; List & Menzies, 2009) propose, we can avoid the result that *C* causes *N*. However, this solution would not satisfy Woodward himself, who (like many others) explicitly rejects proportionality conceived as a necessary condition on causation.[9] And indeed there are strong reasons to do so. For one thing, a proportionality requirement on causation seems to count several perfectly appropriate causal claims as false (Bontly, 2005; McDonnell, 2017; Weslake, 2017; Woodward, 2017). (For example, proportionality arguably gives the wrong verdict that in my gun example the firing is not a cause of the death, as it is not specific enough: gun firings, even at close range, do not always result in death.[10]) It also appears to grant causal status to overly disjunctive factors (Shapiro and Sober, 2012), and has trouble dealing with contrastive causation in contexts involving many-valued variables (Weslake, 2017). In my view these are sufficient reasons to reject proportionality. But I will not argue this point further here. Readers better disposed towards proportionality are welcome to read my argument as offering a new consideration in its favor.

4 Independent fixability and apt causal models

Besides Woodward's (2015) account, another interventionist line of response to Baumgartner's argument is due to Polger et al. (2018). Their account is an instance of a popular interventionist line of thinking that seeks to defuse exclusion worries by appealing to constraints on proper causal models. The idea is that the models like Fig. 1 used to motivate exclusion arguments violate a basic principle of causal modeling which is often called (after Woodward, 2015) *Independent Fixability* (IF). IF says that every variable included in a model should be individually manipulable – i.e., fixable at any of its possible values by intervention while other variables in the model are held fixed at any of their possible values by intervention. This principle seems implicitly presupposed in foundational work in the causal modeling tradition (e.g. Spirtes et al. 1993; Pearl 2000; Woodward, 2003).[11] Given the striking achievements of this tradition, it seems legitimate to uphold IF as a constraint on apt causal models. But the models used to motivate exclusion arguments violate IF, since variables and their supervenience bases cannot be manipulated independently of one another. Once we restrict our attention to models that do obey IF (and make further necessary

[9] See e.g. Woodward (2017). On Woodward's view, proportionality should instead be regarded as a desideratum on good causal explanations.

[10] Thanks to a reviewer for helping me see this point.

[11] For instance, as Eronen (2012) points out, it is assumed in these works that appropriate causal models must satisfy the causal Markov condition. But graphs that violate IF typically do not. For instance, in Fig. 1, M_1 and P_1 are correlated despite being causally unconnected.

adjustments), exclusion worries can be made to disappear. Interventionists have used this strategy to address Kim's original exclusion argument (Eronen, 2012; Weslake, forthcoming) and Merricks's extension of it to all macrocausation (Yang, 2013). Here I will consider how Polger et al. (2018) use it to address Baumgartner's argument.

As Polger et al. recognize, ruling out the model of Fig. 1 as inapt is not enough to answer Baumgartner's argument. For as noted above, (IV) still makes interventions on M_1 impossible even if P_1 is not included in the model. On their view, the right way to deal with that issue is to relativize interventions to a variable set, as in Hausman and Woodward's (1999) early version of interventionism.[12] In that framework, the key notion there is that of an intervention variable I on X with respect to *Y relative to a variable set* $\boldsymbol{V}$ (in which X and Y are included). Polger et al. suggest that the key way in which this relativized notion differs from (IV) is that it replaces I.4 with the following relativized clause:

> I.4_{REL}. I is (statistically) independent of any other variable *in* $\boldsymbol{V}$ that causes Y and is not on a path from X to Y.

We may thus tentatively define an intervention variable I on X with respect to Y in $\mathbf{V}$ as a variable that satisfies I.1, I.2, I.3 and I.4_{REL}. Call that definition (IV_{REL}). X can then be said to cause Y (*simpliciter*) just in case there exists at least one apt variable set that satisfies (M).[13] Following Hausman and Woodward (1999), Polger et al. posit two constraints on apt variable sets: IF and causal sufficiency, which says that a variable set should not omit any direct common causes of variables it contains. The latter principle (which is also standard in causal modeling) is needed to avoid some counterexamples. For instance, suppose that X and Y have a common cause Z, and that X does not cause Y. Relative to the variable set $\{X, Y\}$, there are manipulations of X that satisfy all the conditions on interventions (including I.4_{REL}) and which lead to changes in Y, for instance manipulations that are caused by the omitted common cause Z. Without the causal sufficiency requirement, the account would mistakenly entail that X causes Y.

According to Polger *et al.*, this package view is better motivated than Woodward's later (2003) framework. In particular, it is more faithful to scientific practice, in which experimental interventions are always evaluated with respect to a particular variable set. (Scientists cannot hope to control for every variable that might conceivably be included.) Moreover, they claim, the view elegantly avoids Baumgartner's exclusion problem without the need for special adjustments (e.g. exception clauses in the definitions). Consider the variable set $\{M_1, M_2\}$, which clearly satisfies IF. Because P_1 is not included in that set, a process that changes M_1 can count as an intervention even if it would also change the value of P_1, as this change is kept `off-stage' and hence not forbidden by I.4_{REL}. Polger et al. thus conclude that an intervention on M_1 with respect to M_2 is possible relative to $\{M_1, M_2\}$. Moreover, under some such interven-

[12] See also (Woodward, 1997).

[13] Polger et al. in fact use Woodward's notion of *total causation* instead of contributing causation as their preferred notion of causation. This makes no substantive difference, so I will keep using (M) as working definition of causation here.

tion the value of M_2 would change as well. Provided that the variable set is causally sufficient, this means that M_1 is a cause of M_2.

One concern with this argument is that I.3 still makes interventions on M_1 impossible, even relative to $\{M_1, M_2\}$[14]: for any such intervention I must cause P_1, and thereby be connected with M_2 via a directed path that bypasses X. This suggests that not only I.4 but I.3 needs to be relativized. Taking inspiration from Woodward (1997: S30), we may formulate that relativized clause as follows:

> I.3$_{\mathrm{REL}}$. I is (a) neither a direct cause of Y in $\mathbf{V} \cup I$ (b) nor a direct cause of any other variable in $\mathbf{V} \cup I$ that lies on a directed path to Y that does not go through X.

Now let I be a putative intervention variable on M_1 with respect to M_2 relative to $\{M_1, M_2\}$, and consider the variable set $\{I, M_1, M_2\}$. In that set, I does not count as a direct cause of M_2, provided that wiggling I doesn't change M_2 when M_1 is held fixed (which is perfectly compatible with the fact that I causes P_1).[15] Moreover I also trivially satisfies clause (b) of I.$_{\mathrm{3REL}}$. So if we replace I.3 by I.3$_{\mathrm{REL}}$ in (IV$_{\mathrm{REL}}$) the concern disappears.

A more pressing question is whether Polger et al. are right to assume that $\{M_1, M_2\}$ is an apt variable set. One issue concerns causal sufficiency. Because M_1 supervenes on P_1, any causal process that changes M_1 must also change the value of P_1. And because P_1 is itself a cause of M_2, this means that every cause of M_1 is connected to M_2 via P_1, i.e. via a path that does not go through M_1. One might take this to mean, as Baumgartner (2018) does, that on an interventionist understanding of causation every cause of M_1 is a *common* cause of M_1 and M_2. If so, $\{M_1, M_2\}$ is in fact causally insufficient, contrary to what Polger et al. assume. In reply, one may follow Woodward (2022) in taking common causes to be variables that influence their effects through *independent* paths, in such a way that one can temper with one path without affecting the other. (Standard examples of common causes satisfy this condition: for instance, one can disrupt the mechanism by which atmospheric pressure affects the reading on a barometer without affecting the pressure→storm relationship.) If one takes causal sufficiency to require only the inclusion of common causes so understood, $\{M_1, M_2\}$ is causally sufficient after all. But there is also a second issue. It is commonly agreed in causal modeling that an apt variable set should not omit any variable that might

[14] Thanks to a reviewer for drawing my attention to this concern and the next one discussed below.

[15] That last step raises an issue. According to (M), we can conclude that I is not a direct cause of M_2 only if no interventions on I with respect to M_2 lead to a change in *$M_2$when $M_1$is held fixed at some value by intervention.*But the italicized clause presupposes the notion of an intervention on M_1, whose very coherence is precisely at issue. There are two ways one might reply to this worry. First, in Woodward's (2003) framework it is left open whether to assess if X is a direct cause of Y in $\mathbf{V}$, the interventions on other variables in $\mathbf{V}$ must be interventions with respect to X or with respect to Y. Either option seems acceptable, and if one chooses the first one, the present issue disappears, as interventions on M_1 with respect to I are unproblematic. In particular, the fact that any such intervention would change P_1 is no reason for concern, as P_1 is clearly not a cause of I. Second, one may insist that to assess direct causation, it is enough that other variables in $\mathbf{V}$ be held fixed by conditionalization. As long as the variable set is causally sufficient and the change in X is due to an intervention, this proposal still yields an extensionally adequate definition of direct cause, as far as I can see. And that proposal does yield the result that I is not a direct cause of M_2.

induce spurious dependencies between variables also included in the set. (This is why apt variable sets are required to be causally sufficient.) But if so, to assume that $\{M_1, M_2\}$ is an apt variable set may seem tantamount to presupposing from the get-go that the M_1-M_2 dependency cannot be a spurious one due to P_1, thus begging the question against the proponent of the exclusion argument. That proponent might insist that the set $\{M_1, M_2\}$ is insufficient in some important sense, and that the requirement of causal sufficiency should be extended to require the inclusion of all variables on which two variables in the set depend, either causally or non-causally.[16]

The question of what sufficiency requires in multi-level settings is a complex one, but fortunately I don't need to settle it here. For whatever stance one takes on it, it can be shown that Polger et al.'s proposal falls on one of the two horns of the dilemma concerning composite causation anyway.

Suppose first that one holds $\{M_1, M_2\}$ insufficient on one of the two grounds just discussed – that is, because (a) a causally sufficient set should include all causes of M_1, or (b) because it should include P_1. Then on Polger *et al.*'s proposal mental properties remain excluded, and composite exclusion also follows. (a) Supposing that Baumgartner is right that every apt set that includes M_1 and M_2 must also contain all causes of M_1, including all interventions on M_1, (M) then requires us to hold fixed all those causes, leaving no room for a further intervention to wiggle M_1 anymore. Polger *et al.*'s proposal must then count M_1 as causally impotent. Moreover, similar reasoning as above shows that in the gun example, every set that includes F and V must include all causes of F. (For every cause of F is a cause of at least some of the A_is, and is therefore connected to V via a path that does not go through F.) (M) then requires us to hold fixed all those causes, leaving no room for some further intervention to wiggle F anymore. Polger *et al*'.s proposal thus entails that F cannot cause V. (b) Next, suppose one holds that every set including M_1 and M_2 must include P_1, on the ground that M_1 and M_2 both depend on it (either causally or non-causally). Every set that includes both M_1 and P_1 violates IF, and therefore counts as inapt on Polger *et al.*'s proposal. So on their view there is simply no apt model of the relevant situation, and a fortiori no apt model in which (M) counts M_1 as a cause of M_2. And a similar result obtains in the gun example. Every variable set apt to determine whether F causes V must also include the A_is, since V and F both depend on them (causally or non-causally). But any such variable set violates IF. So there is no apt variable set in which (M) counts F as a cause of V.[17] So mental properties and composites once again remain excluded.

[16] Stern and Eva (2021) consider and ultimately reject such a generalization of causal sufficiency. (See their discussion of "e-parent sufficiency" on p. 15.) Note that although Stern and Eva's paper is entitled "Antireductionist Interventionism", they do not directly address Baumgartner's exclusion argument – rather, their concern is to reply to a different exclusion argument due to Gebharter (2017) seeking to show that the causal Bayes net framework supports exclusion of mental properties. (Although interventionism and the causal Bayes net framework are closely related, Gebharter's argument does not rely in any way on the notion of intervention.) I lack the space to examine whether the dilemma raised for interventionism in this paper might also arise within the causal Bayes net framework.

[17] Note that giving up on IF as a constraint on apt models is of no help here. While the set $\{M_1, M_2, P_1\}$ would now count as an apt variable set, an intervention on M_1 with respect to M_2 would still be impossible (as any such intevrention would violate I.3_{REL} and I.4_{REL}). Likewise, while the set containing F, V and the

Second, suppose one understands sufficiency requirements in such a way that $\{M_1, M_2\}$ is an apt variable set after all, as Polger et al. contend. Then their proposal does vindicate mental causation, in the way indicated above. And it vindicates composite causation as well (though see below). In the gun example, the set $\{F, V\}$ counts as an apt variable set: it is sufficient, and obeys IF. And relative to that set, an intervention changing F's value is perfectly possible. (Of course, any such intervention is bound to change the value of at least some of the A_i variables, and hence also be connected to V via paths that bypass F. But since the A_is are not included in the variable set this is allowed by ($\mathrm{IV}_{\mathrm{REL}}$).) And since the value of V would change under some such intervention on F, F counts as a cause of V. However, just like Woodward's updated account, Polger et al.'s version of interventionism now runs into the problem of over-inclusion. Consider *Judge* again. On the option considered here the variable set $\{C, N\}$ satisfies sufficiency. Since it also satisfies IF, it is an apt variable set. But relative to that set, there exist interventions on C that also change the value of N, namely, interventions that change C by changing the value of the offstage variable V_1. So as in Woodward's updated account C wrongly comes out as a cause of N.[18]

Problems do not stop here. On the presently considered way of interpreting sufficiency, Polger et al.'s account not only yields over-inclusion, but also arguably still leads to exclusion of composites in certain cases. Consider for instance the following variation on *Judge*:

> *Judge 2*. The scenario is the same as in *Judge*, with two modifications. First, the first judge's nervosity is now an effect not only of her own vote, but also of the court's verdict as a whole. Not only does voting to convict make the first judge nervous; in addition, when the court as a whole finds the defendant guilty, this makes the judge even more nervous (by increasing her fear of having possibly contributed to punishing an innocent). Second, the first judge's vote causally influences the second judge's vote, so that the latter votes for conviction only if the former does.

Here, by contrast to *Judge*, C (the court's verdict) *is* a cause of N (whether the first judge gets nervous). But it is not clear that Polger et al.'s account can get this result. For note that in this situation C depends on V_1 (the first judge's vote) in two ways. First, C non-causally depends on V_1, insofar as the first vote is a constituent of the verdict. But one may also claim that V_1 also *causally* influences C, by causally influencing the second vote (on which the verdict also non-causally depends). Admittedly, this claim is controversial, as it goes against Lewis's dictum that causes and effects must be logically and metaphysically independent of each other (Lewis, 1986). But let me offer two considerations in its favor. First, in *Judge 2*, there are clearly two ways in which C depends on V_1. The dependence due to the fact that the first vote

A_is would not count as apt, an intervention on F with respect to V relative to that set would violate $\mathrm{I.3}_{\mathrm{REL}}$ and $\mathrm{I.4}_{\mathrm{REL}}$ and therefore remain impossible.

[18] As should be clear, issues of over-inclusion do not arise if one posits an extended sufficiency requirement on apt models. Because V_1 is a common parent of C and N, every apt model of the situation has to include V_1. ($\mathrm{IV}_{\mathrm{REL}}$) then rules that an intervention on C with respect to N must leave V_1 unchanged, and under any such intervention the value of N remains unchanged as well.

is a constituent of the verdict looks like a straightforward instance of metaphysical dependence. But the second form of dependence (which goes by way of V_1 causing V_2) does not. (For one thing, this second route of influence of V_1 on C goes entirely by way of V_1's influence on V_2, which is metaphysically distinct from V_1. It is hard to see how metaphysical dependence could be mediated by metaphysically distinct entities.) But if this dependence isn't metaphysical, what else can it be but a form of *causal* dependence? Second, there are other, similar cases in which it makes sense to say that non-independent events are causally related. Consider Kim's (1973) example of the event of writing "Larry", which contains as part the event of writing "rr". Lewis uses this example to motivate his dictum, but as Friend (2019) rightly objects, we can imagine contexts in which it is natural to regard writing "rr" as a cause of writing "Larry": imagine for instance that the writer suffers from an obsessive-compulsive disorder that forces them, whenever they write "rr", to preface it with "La" and suffix them with "y". This case reinforces the suspicion that Lewis's dictum is not quite correct.

But now, note that if we accept the claim that V_1 is a cause of C, we get the result that V_1 is also a *common* cause of C and N. So even if one rejects the extended sufficiency requirement discussed above, the *original* causal sufficiency requirement nevertheless entails that every model that includes C and N must include V_1. (This is so even if one understands common causes as variables that operate via independent paths: that condition appears satisfied in the present case.) But any such model violates IF. For instance, holding C fixed at value 1, V_1 cannot be set at any value other than 1. So *Judge 2* is a case where the causal sufficiency and independent fixability constraints conflict: every model that includes C and N must violate one or the other constraint. Polger et al. must then say that there is simply no appropriate model of the relevant causal structure, and hence that the court's verdict is not a cause of the judge's nervosity. In effect, the verdict is causally excluded by one of its parts (the first judge's vote), because the manner in which the verdict depends on this part makes it impossible to model the situation in a way that simultaneously respects the two constraints on apt models posited by Polger et al.

In short, depending on how one understands the sufficiency requirement in contexts involving non-causal dependencies, Polger et al.'s proposal runs either into the problem of exclusion of composites, or into the problem of over-inclusion (and perhaps even into both problems at once).

5 Zhong (2020) on interventionism and exclusion

The two interventionist accounts considered in the last two sections both attempt to solve Baumgartner's problem by allowing interventions on a mental property to also change its physical realizer. By contrast, the third and last revision of interventionism I will consider, Zhong's (2020) account, aims to solve exclusion worries for interventionism while retaining the idea that a proper intervention on a mental property should hold the realizer variable fixed. Let us consider what this account implies for the causal competition between wholes and their parts.

The key claim on which Zhong's response to Baumgartner hinges is that the standard interpretation of P_1 in Fig. 1 is wrong. On that interpretation (which I have adopted so far in this paper), P_1 is a many-valued variable representing the full supervenience base of M_1 (with one value for each possible realizer of M_1). According to Zhong, this reading is problematic, the reason being that on non-reductive physicalism, the supervenience base of mental properties is "an indefinite or even infinite disjunction of diverse subvenient properties" (2020: 298). A many-valued reading of P_1 thus makes that variable too disjunctive to be a real cause. Instead, for Zhong, we should read P_1 as a binary variable such that $P_1 = 1$ represents the specific realizer property instantiated in Kim's scenario, and $P_1 = 0$ the absence of that property.[19] Crucially, on Zhong's construal of P_1, M_1 does *not* supervene on P_1, as $P_1 = 0$ is compatible with both $M_1 = 1$ and $M_1 = 0$ depending on whether some other realizer property is instantiated. If so, and *pace* Baumgartner, it is possible to change the value of M_1 in a way that does not change the value of P_1, provided that P_1 is first held fixed at its (non-actual) value 0 by another intervention. Moreover, because holding P_1 fixed at 0 and changing M_1 would induce a change in M_2, (M) counts M_1 as a cause of M_2 after all. The threat of exclusion is defused, in a way that appears to retain the key idea that a proper intervention should not directly affect variables other than its target.

As Zhong notes, this reasoning faces an objection (raised by McDonnell's (2017: 1467) in a discussion of an earlier paper by Zhong (2014). Though on Zhong's reading M_1 does not supervene on P_1, it *does* supervene on the set of all of its realizer variables. Thus any process that changes M_1 while P_1 is held fixed at 0 must also change the value of another such realizer variable. That is, there must be some other realizer variable Q_1 (with value 1 representing the presence of the relevant realizer, and 0 its absence) that changes value as M_1 is wiggled. Moreover, Q_1 lies on a directed path to M_2 (via P_2 or some other realizer variable for M_2). Hence, any such change in M_1 still violates (IV). As I understand it, Zhong's response to this worry is as follows.[20] In Kim's scenario, there are two competing causal candidates: the mental property $M_1 = 1$, and the physical property that realizes it on the relevant occasion ($P_1 = 1$). To determine which is the cause, we need to wiggle one property while keeping the other fixed at some (possibly non-actual) value, and check whether we observe a change in the effect. In doing so, we are allowed to vary other, off-stage variables that represent properties not instantiated in the actual scenario. Given that the property represented by Q_1 is not instantiated in Kim's scenario, it is therefore acceptable to wiggle that variable when examining whether the mental property was causally efficacious.

Clearly, this response requires adjustments to (IV), though Zhong does not make it precise what they are. One way to read his proposal is as endorsing the same relativized version of (IV) that Polger *et al.* endorse, where the relevant variable set is fixed by the context of inquiry and should include only variables that we regard as serious causal candidates.[21] (In our case that variable set should include M_1 and P_1, and omit

[19] Presumably we should read P_2 in a similar way, though Zhong works with a model that omits P_2 and therefore does not discuss this point.

[20] See (2020: 307-8).

[21] One piece of evidence for that reading is Zhong's claim that interventionism not only makes M_1 a cause of M_2, but also entails that P_1 does not cause M_2. (That is, interventionism yields `downward exclusion' of

variables representing properties not instantiated in the relevant situation, such as Q_1.) This is the reading of Zhong's proposal with which I will work in what follows, though I believe my points would still stand on other possible readings.

Besides this interpretative question, Zhong's account raises a number of issues. The fact that the account must allow interventions to vary some other variables than their targets after all somewhat undercuts its primary motivation. And the prohibition on disjunctive causal relata seems at odds with the spirit of interventionism (which includes a distrust of such metaphysical constraints on causation). But let's leave those issues aside, and consider how Zhong's account fares with respect to the issue of composite causation.

One virtue of Zhong's proposal is that it avoids the problem of over-inclusion that affects the two other interventionist accounts examined in Sects. 3 and 4. As an illustration, consider *Judge* again. Here the two candidates for causing *N* are *C* (the court's verdict) and V_1 (the first judge's vote), so the appropriate variable set is $\{C, V_1, N\}$. Applied to this set, (M) rightly entails that the first judge's vote caused her nervosity but the court's verdict did not. For on the one hand, holding *C* fixed at value 0 (i.e. verdict of innocence) while changing the value of V_1 would change *N*, so that V_1 causes *N*. But on the other hand, there is no setting of V_1 under which it is possible to change *N* by intervening on *C*. If V_1 is held fixed at its value 1 (vote to convict), such an intervention is possible, but leaves *N* unchanged. And if V_1 is held fixed at value 0 (vote to acquit), *C* is bound to take value 0, and intervening on it is impossible.

But unfortunately, when it comes to composite causation Zhong's account avoids the second horn of the dilemma (over-inclusion) only at the price of falling into the first one (exclusion). To see this, return to the gun example of Fig. 2, where the causal competition here is between the gun (whose behavior is represented by *F*) and its microscopic parts (whose behaviors are represented by A_1, …A_n). For Zhong's proposal to count *F* as a cause of *V*, there would need to be a possible setting of A_1, …, A_n that does not fix the value of *F* and hence leaves room for an intervention that changes *F*'s value. But there is no such setting: fixing how each of the particles composing the gun behaves also fixes whether the gun fires, leaving no room to wiggle the value of *F*. Here there is a crucial asymmetry between the M_1-P_1 relationship and part-whole relationships. In the former case, setting P_1 to its `absent' value still leaves room to wiggle M_1. But in the case of parts and wholes, there is simply no way to set the behavior of the parts that does not also fix the behavior of the whole. Consequently, Zhong's account has to yield the result that composites are causally excluded by their parts.[22] (This means that while Zhong's account entails that M_1 is

subvenient by supervenient properties.) His argument for that claim is that there is no way to wiggle M_2 by wiggling P_1 while holding M_1 fixed. (Either we hold M_1 fixed at value 0 and P_1 cannot be wiggled, or M_1 is held fixed at value 1 and M_2 takes value 1 whatever value P_1 takes.) This presupposes that every model relevant for evaluating P_1's causal status must include M_1. The principle that a model should include all serious causal candidates provides a rationale for this presupposition.

[22] Of course, if we consider a model that includes only a few of the A_i variables, then it may still be possible to wiggle the behavior of the gun while holding those variables fixed. However, such a model is inappropriate to settle the question "was it the behavior of the gun or the behavior of the atoms that caused the effect?" To address that question, we need a model that includes enough variables to represent the behaviors of all of the atoms that compose the gun. This makes for a crucial difference with the case of mental causation: for in Kim's scenario, a model that contains M_1 and P_1 (construed along Zhong's lines) *is*

not causally excluded by P_1, it rules that both *are* causally excluded by the properties of the particles that compose my brain.)

In fact, Zhong's account has even more unwelcome implications: it entails that parts are *also* causally excluded by their wholes. For consider Fig. 2 again. For $A_1 \ldots A_n$ to cause V on Zhong's account, there would have to be a possible setting of F under which interventions on $A_1 \ldots A_n$ would change the value of V. Now, holding F fixed at any value still makes it possible to wiggle the values of $A_1 \ldots A_n$, as there are many possible settings of those variables that can realize both the gun firing and its not-firing. However, holding the value of F fixed plausibly fixes the value of V: whether the gun fires fixes whether the victim dies, no matter how the firing or non-firing is realized microscopically. This suggests that Zhong's account yields *both* upwards and downwards exclusion in the context of part-whole relationships.

Admittedly, Zhong's account yields downward exclusion in the gun example only if we read the A_is as many-valued variables that can represent all possible physical states of the relevant particles. Only on that interpretation is it plausible to claim that F supervenes on the set of A_is and hence that every possible value setting of the A_is fixes the value of F. As noted, Zhong rejects a similar reading of P_1 in Fig. 1, and may likewise reject it here. And if each A_i is read in a similar manner as Zhong's preferred interpretation of P_1, i.e. as a binary variable with one value representing the actual physical state of the relevant particle on a certain occasion and the other value the absence of that state, it now becomes possible to wiggle the value of F while holding the A_is fixed at their non-actual values.

However, there are two points to make in response. First, Zhong's reasons for rejecting the many-valued interpretation of P_1 (viz. that it makes P_1 too `disjunctive' to be a cause) do not apply the A_i variables. After all, the possible values of those variables are meant to represent possible physical states of an elementary particle, which intuitively form a far more unified and natural set than the collection of possible physical realizers of M_1 (which may well have nothing theoretically interesting in common except that they realize the same property). Second, even if a good reason could be found to reject a many-valued interpretation of the A_is, Zhong's account would still lead to exclusion of composites by their parts (though perhaps not their *microscopic* parts) in certain cases. To see this, consider *Judge* again, and let J represent whether the defendant is jailed. One might wonder whether it is the verdict (the whole) or the votes (the parts) that cause J. In the present case the parts are represented by *binary* variables, not many-valued ones: no disjunctiveness worry, then. But now, applying (M) to the set $\{C, V_1, V_2, V_3, J\}$ yields the result that the verdict is not a cause of the jailing, since every possible combination of values of V_1, V_2 and V_3 fully fixes whether the court finds the defendant guilty, leaving no room for an intervention on C. (By the same token, we get the result that the votes are not causally relevant either, as varying them while holding C fixed is either impossible or fails to wiggle J.) The case thus shows that a prohibition on many-valued variables

appropriate to settle the question "was it the mental property or its specific realizer that caused the effect?" And in that model, it is possible to wiggle M_1 while holding P_1 fixed (at value 0). Thanks to a reviewer for pressing me to address this point.

is not enough to help Zhong's account avoid exclusion issues in contexts involving part-whole relationships.

6 Conclusion

Interventionism has two options when it comes to properties of composite objects: either require interventions on those properties to hold the behavior of the object's parts fixed, or allow those interventions to change how the parts behave. The first strategy runs the risk of making wholes causally excluded by their parts, whereas the second strategy is in danger of mistakenly ascribing to composite objects causal abilities that properly belong to their parts only. I have argued that all major versions of interventionism fall on one horn or the other of this dilemma, including all the versions of interventionism that have been proposed to address the exclusion worry concerning mental and other non-physical properties identified by Baumgartner. I do not mean to claim, however, that the dilemma is an insuperable problem for interventionism. What might a proper solution look like? Outside of interventionism, a number of authors have sought to solve Kim's exclusion problem by carving a middle way between reductionism and traditional anti-reductionism – one that recognizes multiple realizability while denying that high-level properties are distinct enough from low-level ones to causally compete with them.[23] (See e.g. Jessica Wilson's (2011) view that high-level properties are subsets of causal powers of their realizers, and Piccinini's (2021) recent `aspect view' of realization.) It is doubtful, however, whether interventionism could solve the problem by denying distinctness. If high-level variables are not distinct from lower-level ones, then it seems clear that one should not hold the latter fixed while wiggling the latter. But as we have seen, it is difficult to implement this line of thought without running into the problem of over-inclusion. Another line of response would be to appeal to metaphysically impossible interventions that can vary the behavior of a whole while keeping its parts fixed, thereby allowing us to vindicate the causal efficacy of composites without ascribing them causal powers that belong to their parts. This move would not be entirely without precedent, as metaphysically impossible interventions have already been recruited to do theoretical work in other contexts (see e.g. Wilson (2018). But many interventionists will likely find such interventions overly mysterious. In my view, the most promising strategy would be to find a principled way for the interventionist framework to distinguish between cases in which interventions on composites can vary the behavior of their parts and cases where they cannot. But how exactly to draw and justify the relevant distinction is a question beyond the scope of this paper.

Acknowledgements The general framing of the paper owes much to Toby Friend, who provided very useful comments on an early version of it. Andreas Hüttemann and Thomas Krödel also read and made helpful suggestions on the paper. Finally, I would like to thank two anonymous reviewers as well as audiences at the University of Cologne, Oxford University and the 2021 Conference of the Society for the Metaphysics of Science for helpful comments.

[23] Thanks to a reviewer for drawing my attention to this point.

Funding Open Access funding enabled and organized by Projekt DEAL. No funding was received for conducting this study.

Declarations

Competing interests The author has no relevant financial or non-financial interests to disclose.

References

Baumgartner, M. (2009). Interventionist causal exclusion and non-reductive physicalism. *International Studies in the Philosophy of Science*, *23*, 161–178.

Baumgartner, M. (2013). Rendering interventionism and non-reductive physicalism compatible. *Dialectica*, *67*, 1–27.

Baumgartner, M. (2018). The inherent empirical underdetermination of mental causation. *Australasian Journal of Philosophy*, *96*, 335–350.

Baumgartner, M., & Casini, L. (2017). An abductive theory of constitution. *Philosophy of Science*, *84*, 214–233.

Baumgartner, M., & Gebharter, A. (2016). Constitutive relevance, mutual manipulability, and fat-handedness. *The British Journal for the Philosophy of Science*, *67*, 731–756.

Block, N. (2003). Do causal powers drain away? *Philosophy and Phenomenological Research*, *67*, 133–150.

Bontly, T. D. (2002). The supervenience argument generalizes. *Philosophical Studies*, *109*, 75–96.

Bontly, T. D. (2005). Proportionality, causation, and exclusion. *Philosophia*, *32*, 331–348.

Craver, C. F. (2007). *Explaining the brain*. Oxford: Oxford University Press.

Eronen, M. I. (2012). Pluralistic physicalism and the causal exclusion argument. *European Journal for Philosophy of Science*, *2*, 219–232.

Eronen, M. I., & Brooks, D. S. (2014). Interventionism and supervenience: A new problem and provisional solution. *International Studies in the Philosophy of Science*, *28*, 185–202.

Friend, T. (2019). Can parts cause their wholes? *Synthese*, *196*, 5061–5082.

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

Gillett, C. (2013). Constitution, and multiple constitution, in the sciences: Using the neuron to construct a starting framework. *Minds and Machines*, *23*, 309–337.

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

Hoffman-Kolss, V. (2014). Interventionism and higher-level causation. *International Studies in the Philosophy of Science*, *28*, 49–64.

Kim, J. (1973). Causes and counterfactuals. *Journal of Philosophy*, *70*, 570–572.

Kim, J. (1998). *Mind in a physical world: An essay on the mind-body problem and mental causation*. Cambridge, MA: MIT Press.

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

Kim, J. (2005). *Physicalism, or something near enough*. Princeton: Princeton University Press.

Lewis, D. K. (1986). Events. *Philosophical Papers* (II vol., pp. 241–269). Oxford: Oxford University Press.

List, C., & Menzies, P. (2009). Nonreductive physicalism and the limits of the exclusion principle. *The Journal of Philosophy*, *106*, 475–502.

McDonnell, N. (2017). Causal exclusion and the limits of proportionality. *Philosophical Studies*, *174*, 1459–1474.

Merricks, T. (2001). *Objects and persons*. Oxford: Oxford University Press.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Piccinini, G. (2021). *Neurocognitive mechanisms: Explaining biological cognition*. Oxford, New York: Oxford University Press.

Polger, T. W., Shapiro, L. A., & Stern, R. (2018). In defense of interventionist solutions to exclusion. *Studies in History and Philosophy of Science*, *68*, 51–57.

Raatikainen, P. (2010). Causation, exclusion, and the special sciences. *Erkenntnis*, *73*, 349–363.

Romero, F. (2015). Why there isn't inter-level causation in mechanisms. *Synthese*, *192*, 3731–3755.

Shapiro, L., & Sober, E. (2007). Epiphenomenalism—The Do's and the Don'ts. In P. Machamer, & G. Wolters (Eds.), *Thinking about causes: From Greek philosophy to modern physics*. Pittsburgh: University of Pittsburgh Press.

Shapiro, L., & Sober, E. (2012). Against proportionality. *Analysis, 72*(1), 89–93.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction and search*. Cambridge, MA: MIT Press.

Stern, R., & Eva, B. (2021). Anti-reductionist interventionism. *The British Journal for the Philosophy of Science*. https://doi.org/10.1086/714792.

Weslake, B. (2017). Difference-making, closure, and exclusion. In H. Beebee, C. Hitchcock, & H. Price (Eds.), *Making a difference* (pp. 215–231). Oxford: Oxford University Press.

Weslake, B. (Forthcoming). Exclusion excluded. *International Studies in the Philosophy of Science*.

Wilson, J. (2011). Non-reductive realization and the powers-based subset strategy. *The Monist*, *94*, 121–154.

Wilson, A. (2018). Metaphysical causation. *Noûs, 52*, 723–751.

Woodward, J. (1997). Explanation, invariance, and intervention. *Philosophy of Science*, *64*, S26–S41.

Woodward, J. (2003). *Making things happen*. Oxford: Oxford University Press.

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

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

Woodward, J. (2015). Interventionism and causal exclusion. *Philosophy and Phenomenological Research*, *91*, 303–347.

Woodward, J. (2017). Intervening in the exclusion argument. In H. Beebee, C. Hitchcock, & H. Price (Eds.), *Making a difference* (pp. 251–268). Oxford: Oxford University Press.

Woodward, J. (2022). Modeling interventions in multi-level causal systems: Supervenience, exclusion and underdetermination. *European Journal for Philosophy of Science*, *12*, 1–34.

Yablo, S. (1992). Mental causation. *The Philosophical Review*, *101*, 245–280.

Yang, E. (2013). Eliminativism, interventionism and the overdetermination argument. *Philosophical Studies*, *164*, https://doi.org/10.1007/s11098-012-9856-0.

Zhong, L. (2014). Sophisticated exclusion and sophisticated causation. *The Journal of Philosophy*, *111*, 341–360.

Zhong, L. (2020). Intervention, fixation, and supervenient causation. *The Journal of Philosophy*, *117*, 293–314.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.