Synthese (2024) 203:91
https://doi.org/10.1007/s11229-024-04500-7

ORIGINAL RESEARCH

Functionalism, interventionism, and higher-order causation

Matthew Rellihan[1]

Received: 21 November 2022 / Accepted: 20 January 2024 / Published online: 6 March 2024

Abstract
It has been argued that nonreductive physicalism's problems with mental causation disappear if we abandon the intuitive but naïve production-based conception of causation in favor of one based on counterfactual dependence and difference-making. In recent years, this response has been thoroughly developed and defended by James Woodward, who contends that Kim's causal exclusion argument, widely thought to be the most serious threat to nonreductive mental causation, cannot even be given a coherent formulation within Woodward's preferred interventionist framework. But Woodward has, even more recently, defended a pair of necessary conditions on mental causation and higher-order causation more generally, and it is here that the interventionist framework proves less friendly to nonreductive mental causation. Functionalism is arguably the most important species of nonreductive physicalism concerning specifically mental properties, but, as I argue, functional properties fail both of Woodward's tests of causal relevance, one of them in two different ways. The problem, moreover, seems unique to functionalism, for other types of higher-order properties appear to pass Woodward's tests. If functionalism faces defeat even in such friendly territory, its problems with mental causation are not an artifact of naïve metaphysics. They run deep.

Keywords Higher-order causation · Mental causation · Interventionism · Functionalism

1 Introduction

It has been argued that nonreductive physicalism's problems with mental causation disappear if we abandon the intuitive but naïve production-based conception of causation in favor of one based on counterfactual dependence and difference-making. Woodward (2008, 2015, 2017) has thoroughly developed and defended this strategy

✉ Matthew Rellihan
relliham@seattleu.edu

[1] Department of Philosophy, Seattle University, 901 12th Avenue, P.O. Box 222000, Seattle, WA 98122-1090, USA

in recent years, focusing his attention on (Kim's, 1998, 2005) causal exclusion argument.[1] Adopting an interventionist account of causation, which improves upon the standard difference-making accounts in a variety of ways,[2] Woodward argues that the question at the heart of the exclusion problem—Is it M or its physical realizer P that is responsible for some effect P*?—"has no coherent 'interventionist' interpretation" (2008, p. 255). On the assumption that a problem that cannot even be posed needn't be solved, Woodward concludes that causal exclusion is not a worry for nonreductive physicalists.

Woodward's conclusions are not uncontroversial. Baumgartner (2009, 2010), for example, has argued that a proper understanding of the interventionist criterion of causal relevance actually supports the conclusion of Kim's exclusion argument. A significant (and growing) literature has arisen on both sides of this debate.[3] However, the question of whether Woodward's interventionism, as originally articulated, strengthens or undermines the exclusion argument may be moot. This is because Woodward has, even more recently, defended a pair of necessary conditions on higher-order causation, and these pose new and, arguably, more serious threats to nonreductive mental causation. On Woodward's account, higher-level variables can be causally related to variables at the same level or at lower levels only if (1) the variables corresponding to the cause and effect can be fixed independently of each other by means of interventions and (2) the putative causal relations holding of the higher-level variables are insensitive to differences in their lower-level realizers.[4] Both conditions are violated in the case of functional properties, as I shall presently argue. But functionalism—or, more precisely, role functionalism[5]—is, arguably, the most important variety of nonreductive physicalism concerning specifically mental properties. Role

[1] See also Loewer (2002, 2007) and Crane (2001) for earlier defenses of this approach in general terms. Kim (2009, p. 44) has responded that productionist mental causation is the only form of mental causation worth having, partly because difference-making relations come too cheap. They may come cheap, but they are nevertheless too dear for functionalism, as we'll see.

[2] The interventionist account, being nonreductive, avoids many of the issues with (Lewis's, 1973, 1986) account, for example. See Paul and Hall (2013, Ch. 2) for a discussion of these problems. See Woodward (2003, Ch. 3) for a discussion of the ways in which interventionism improves upon Lewis's account.

[3] Woodward (2015, 2017) argues that Baumgartner's argument is based upon a misinterpretation of the spirit if not the letter of interventionism and proposes a condition he calls 'independent fixability' in response. Much more on this below. Further contributions to this debate are made by List and Menzies (2009), Raatikainen (2010), Weslake (2011), Shapiro (2012), Polger et al. (2018), Zhong (2020), among many others.

[4] A number of other authors have argued for a condition similar to the second of these. List and Menzies (2009, 2010) are one prominent example. One point on which Woodward's account disagrees is whether higher-level causation excludes lower-order causation. List and Menzies believe that it often does, while Woodward does not—but this is largely due to the fact that Woodward adopts a very permissive account of causation, as we'll see below. See Woodward (2021, fn. 23) for discussion. Wilson (2011, 2021) also defends a related condition within an ontology of causal powers and without presupposing a difference-making account of causation. See Antony and Levine (1997) for an early articulation of this condition.

[5] Role functionalism identifies a mental (or other higher-order) property with the second-order property of having some first-order property that occupies a certain causal role. Realizer functionalism, by contrast, identifies the mental property—e.g., pain—with the first-order property—e.g., C-fiber firing in humans—that occupies the associated role. It is for this reason that realizer functionalism is not generally regarded as a species of nonreductive physicalism. See, e.g., Kim (2011, pp. 186–189) for discussion. For ease of presentation, I shall sometimes drop the qualifier and speak of functionalism in what follows, but I should be taken to mean only the role-functionalist variety thereof.

functionalist arguments continue to be the most important for establishing the irreducibility of the mental, and role functionalism continues to be the dominant—though not, of course, unchallenged—view in the cognitive sciences.[6] If interventionism is inconsistent with functionalist mental causation, it thus poses a serious challenge to nonreductive mental causation more generally.

My arguments will not target higher-order properties or higher-order causation across the board, as is the case with the exclusion argument. Quite the contrary. We'll see that the sorts of higher-order properties that figure prominently in Woodward's discussion of the issue—properties like having such-and-such a temperature or being red—appear to meet both of the proposed conditions for causal relevance. This suggests that these conditions are not unrealistic or overly restrictive and that Woodward's framework is generally hospitable to higher-order causal relations. Functionalism's failure to meet these conditions thus suggests that its problems with mental causation are not an artifact of naïve metaphysics. They run deep. And to the extent that nonreductive physicalism about the mental rests on the tenability of functionalist versions thereof, its problems run deep as well.

I'll begin making this case in the following section by introducing the interventionist framework and Woodward's account of higher-order causation. We'll see, among other things, that the principle of independent fixability is the basis upon which Woodward argues that the exclusion argument is ill-formed and incoherent. Abandoning it thus means abandoning what some have found to be a promising response to Kim. But in section three, we'll see that the principle of independent fixability rules that functional properties cannot be causally related to their constitutive causes and constitutive effects, resulting in a *reductio* of functionalism itself. In section four, we'll see that functionalism fails to meet the second condition of realization insensitivity or conditional irrelevance in two different ways—one having to do with the phenomenon of multiple realizability, the other with the distinction between core and total realizers.[7]

2 Interventionism and higher-order causation

Let us begin with Woodward's interventionist analysis of type-level causal relations of the form X causes Y in circumstances B:

> M*: X causes Y in circumstances B if and only if there are distinct values of X, x_1 and x_2, with $x_1 \neq x_2$, and distinct values of Y, y_1 and y_2, with $y_1 \neq y_2$, and

[6] Baker (2009) notes that arguments for nonreductive physicalism in the mental domain derive either from Davidson (1980) or from Fodor's (1974) and Putnam's (1975) role functionalist arguments. And, with apologies to Davidson, I do not believe that the arguments for anomalous monism have many adherents in contemporary philosophy of mind. The importance of role functionalist arguments is also evidenced by the fact that those who argue for type identity theory—e.g., (Polger & Shapiro, 2016)—continue to direct their attacks almost exclusively at role functionalist versions of nonreductive physicalism. As Kim (2009, p. 46) observes, "[t]he [role] functionalist view of the mind is still the most widely accepted approach to the nature of mentality" and "is arguably the 'official' philosophy of cognitive science.".

[7] I've argued for claims related to these last two elsewhere (see Rellihan, 2019, 2021)), but not from within Woodward's framework. Given the importance of Woodward's account of causation and its associated response to the exclusion argument, its it is useful to see that this framework itself poses problems for mental causation.

> some intervention such if that intervention were to change X from x_1 to x_2 while holding 'off-path' variables fixed at certain values by means of interventions, then Y would change from y_1 to y_2.[8]

M* is extremely permissive on what can count as a cause of what. First, it is agnostic on the nature of causal relata. X and Y are variables; causation is a relation between whatever in the world—events, tropes, property instantiations, etc.—corresponds to these variables. But even within this metaphysically neutral terrain, M* is extremely permissive. An intervention on X with respect to Y is a way of setting the value of X so that any change to the value of Y occurs only 'through' X and not due to the effects of a common cause.[9] M* requires only that *some* of the interventions that change *some* of the values of X be associated with changes to Y. Cases we consider below will illustrate just how weak this requirement is. A path from X to Y is, roughly speaking, an independent route of influence, and one variable can influence another along multiple paths, as when birth control pills both increase and decrease the risk of thrombosis—the former directly, the latter indirectly by decreasing the likelihood of pregnancy, which itself increases the risk of thrombosis. M* requires only that there be difference-making relations along one such path.[10] Moreover, M* incorporates no assumptions antithetical to the possibility of mental causation, such as Kim's exclusion principle. The relation described in M* may hold between a pair of upper-level variables as well as between a pair of variables representing their lower-level realizers. If so, causation exists at both levels. M* even permits cases of downward causation in which variables at a higher-level cause changes to variables at a lower-level. I'll refer to both upper-upper and upper-lower causation as instances of higher-order causation.

Interventionism is extremely permissive, but not just anything goes. Consider, first, the objection that at least one species of downward causation is incoherent because the associated causal relata are not sufficiently distinct. The individual water droplet is carried along by the eddy; the rotation of the spoke follows the rotation of the

[8] See Woodward (2021, p. 242) for a formulation very close to this. That formulation omits the clause about holding off-path variables fixed but is otherwise identical. Without this clause, we have an analysis of what Woodward elsewhere calls a 'total' cause; with it, we have an analysis of what he calls a 'contributing' cause (2003, pp. 51–59). Contributing causation is the more general notion and thus the appropriate one for a general discussion of higher-order causation. (Note that a total cause can be thought of as the special case of a contributing cause in which the set of off-path variables that have to be held fixed at certain values is empty.) Woodward often frames his discussion of higher-order causation in terms of the simpler analysis and thus in terms of total causation. This is a harmless simplification because the issues surrounding higher-order causation run orthogonal to the distinction between total and contributing causes. I've nevertheless chosen to frame the discussion that follows in terms of the more general notion so as to make clear that Woodward intends his constraints on higher-order causation to apply to both types of the causal relation.

[9] A much more rigorous definition of an intervention can be found in Woodward (2003, Ch. 3).

[10] Consider an illustration. In the aggregate, birth control pills may not affect the probability of developing thrombosis, for the effects along the distinct pathways may cancel each other out. If so, birth control pills are not a total cause of thrombosis. (See fn. 8). But, according to M*, they are nevertheless a contributing cause of thrombosis because when we hold the value of the off-path variable corresponding to pregnancy constant, changes to the variable representing the ingestion of birth control pills *are* associated with changes to the variable representing thrombosis. That is, if we hold constant the fact that a woman is not pregnant, changes to the birth control variable correspond to changes to the thrombosis variable. The example is due to (Hesslow, 1976), who uses it for a different purpose. See the discussion of Hesslow's example in Woodward (2003, pp. 49–51).

wheel. These cannot be causal relations, the objection goes, because they are also mereological relations and something cannot be causally related to one of its own parts. M*, however, imposes no such restriction. Let X represent the position of the wheel, Y the position of the spoke. If I rotate the wheel in the right way, I change the value of X by means of a suitable intervention—but I also change the value of Y, for the spoke moves with the wheel. M* thus appears to permit causal relations between parts and wholes.[11] This not only offends intuition, it also trivializes downward causation by making it ubiquitous.

The correct response to this objection is complicated by the fact that not every case of causation between apparent parts and apparent wholes is objectionable. The Hodgkin-Huxley model of the action potential is, for Woodward, a case in point. According to this model, changes to the neuronal membrane potential (apparent whole) cause changes to the conductances of individual ion channels (apparent parts).[12] And yet there is nothing objectionable about causal relations of this sort.

The difference between the objectionable and unobjectionable cases is only apparent if we adopt a precise understanding of the condition that causal relata be distinct rather than simply relying on intuitions concerning what is or is not a part of what. Woodward calls this the condition of independent fixability (IF), which he defines as follows:

> IF: According to IF, variables in a set V are suitably distinct if and only if it is "possible" to set each variable to each of its values via an intervention while also setting any other variable in V to each of its values via an intervention. "Possible" here includes logical or conceptual possibility as well as causal possibility (2020, p. 861).[13]

Clearly, the variables corresponding to the positions of the spoke and the wheel do not meet this condition. The claim, recall, is that the wheel's being at x_2 rather than x_1 causes the spoke to be at y_2 rather than y_1. For IF to obtain, it would have to be possible to set the values of wheel and spoke to x_2 and y_1, respectively, via interventions, but this isn't possible (short of breaking the wheel and thereby changing the assumed background conditions). On the other hand, IF does obtain in the case of the Hodgkin-Huxley model of the action potential. Indeed, (Woodward, 2020, p. 862) points out

[11] As a reviewer points out, it is not altogether clear that this is a genuine relationship between parts and wholes—the spoke is part of the wheel, but it's less clear that the position of the former is part of the position of the latter. It is still less clear how talk of parts and wholes applies to relations between variables. My point here, however, is simply to motivate the need for IF. Woodward proposes IF as a way of precisifying the intuition that causal relata must be distinct. That it is unclear whether spokes and wheels are appropriately distinct is enough to motivate the need for IF.

[12] See Woodward (2020, pp. 860–862) for further discussion of this and similar cases. Woodward does not, of course, believe that channel conductances and membrane potentials are actually related as parts and wholes, for then they would not be sufficiently distinct. The point, again, is that we cannot rely on our intuitions concerning what is or is not a part of what. Sometimes causal relations between apparent parts and apparent wholes are objectionable, sometimes they are not. We therefore require a clearer account of distinctness than intuition affords. This is what IF is intended to provide. I am thankful to a reviewer for pointing out the need for this clarification.

[13] Woodward develops IF in more detail and argues that it is consistent with the interventionist approach in Woodward (2015).

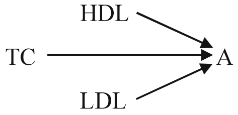

Fig. 1 Causal graph of the relations between HDL, TC, LDL, and A

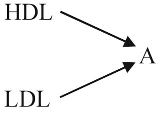

Fig. 2 Causal graph of the relations between HDL, LDL, and A

that interventions that alter ion conductances independently of membrane potential and vice versa are now technologically feasible.

Note that IF is intended as a general constraint on causal models and their associated causal graphs and not merely as a constraint on putative causes and their putative effects. Assume that low-density cholesterol (LDL) causes plaque to build up in the arteries (A), that high-density cholesterol (HDL) causes it to diminish, and that total cholesterol (TC) is defined as the arithmetic sum of HDL and LDL.[14] Consider now the causal graph depicted in Fig. 1, which is intended to model these relationships.

There is no suggestion in the graph (nor will there be in the structural equations of its associated causal model) that there are any causal relations between LDL, HDL, and TC themselves. And yet, the graph represents a violation of IF because not every assignment of values to these three variables is possible. One cannot, for example, manipulate LDL while holding both HDL and TC constant. One could, of course, devise a manner of representing hybrid causal/conceptual relations of this sort—(Woodward, 2015) suggests using double-tailed arrows to represent relations of definitional or metaphysical dependence—but even so the resulting graph will not be able to be deployed in the usual manner to draw conclusions about causal relations between its constituent variables.

Consider now the graph depicted in Fig. 2 below. Figure 2 was generated from Fig. 1 by removing the redundant variable TC and thus restoring adherence to IF. It is important to note that because of the counteracting effects of HDL and LDL on A, neither Fig. 2 nor its associated structural equations suggest that increasing LDL in and of itself results in an increase in arterial blockage. The claim, rather, is that increasing LDL results in an increase in arterial blockage when *off-path* variables like HDL are held constant. This is what is required for LDL to be causally related to A.[15] Recall that M* requires only that changes to X be associated with changes to Y when off-path variables are set to certain values by means of interventions.

It is now apparent why Fig. 1 misleads, for even though it does not suggest that HDL, LDL, and TC are themselves causally related, it does suggest that they are competing causes of A. To demonstrate the contribution of any one of them, it would

[14] This example comes originally from Spirtes and Scheines (2004). Woodward discusses it in his (2015). I follow Woodward's presentation even though it is medically inaccurate in way that makes no difference to the underlying moral—total cholesterol also measures the level of triglycerides in the blood.

[15] More precisely, this is what is required for LDL to be a contributing rather than total cause of A. (See fn. 8). It is the former that is the central notion. In what follows it can be assumed that I am speaking of contributing causes.

$$\begin{tikzcd}
M & & M^{*} \\
P \arrow[u] \arrow[rr] & & P^{*} \arrow[u]
\end{tikzcd}$$

Fig. 3 Kim's causal exclusion diagram

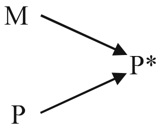

Fig. 4 Causal graph of the exclusion scenario

be necessary—as in the case of the situation depicted in Fig. 2—to manipulate that variable while holding the others fixed by means of interventions. But there is no intervention on LDL, for example, that results in a change to A when both TC and HDL are held fixed, such interventions being definitionally impossible. Without IF as an independent constraint, M* would therefore lead us to the mistaken conclusion that LDL is not causally related to A.

But now, finally, consider Kim's causal exclusion argument and its associated diagram. Woodward (2015) argues that this diagram is misleading in the same way that Fig. 1 is, for it depicts both causal and non-causal relationships between the variables (Fig. 3).

But even if we eliminate the noncausal relations, the associated graph, shown in Fig. 4 below, is still misleading, for it suggests that M and P are competing causes of P*. To establish within the interventionist framework that M is indeed a cause of P*, we would have to manipulate M while holding P fixed by means of an intervention, which is metaphysically[16] impossible.

But to conclude on this basis that M does not influence P* would be like concluding that LDL does not influence A on the basis of Fig. 1. Variable sets that violate IF do not support causal inferences, and any variable set in which higher-level variables appear alongside their lower-level realizers violates IF. We simply cannot manipulate higher-level variables while holding their lower-level realizers constant and it is therefore impossible even to apply the interventionist test of causal relevance. This being impossible, (Woodward, 2008, p. 255) concludes that the exclusion problem is simply incoherent. Woodward's response to Kim is not entirely uncontroversial, as was noted above, but any nonreductive physicalist wishing to adopt it in defense of mental causation is thereby committed to IF. This matters because, as we'll shortly see, IF

[16] I'm assuming nonreductive physicalism here, according to which mental states supervene with metaphysical necessity on physical states. It's worth noting, however, that Woodward's defense of mental causation applies to dualistic accounts holding that the mental supervenes with only nomological necessity on the physical. Chalmers (1996, famously defends such a view). This is because, as Woodward makes clear in his definition, IF is meant to cover not only metaphysical but also 'causal' possibility, which I take to mean nomological possibility. If M supervenes with nomological necessity on P, it is not causally possible to manipulate M while holding P fixed. Woodward's defense of nonreductive physicalism thus applies, *mutatis mutandis*, to at least some versions of dualism. I am thankful to reviewer for requesting clarification on this matter.

introduces new problems for nonreductive physicalism—or, rather, role functionalist varieties thereof.

The incoherence of the exclusion problem is insufficient to establish that higher-level properties like M are causally efficacious or, indeed, that they have the kind of autonomy that nonreductive physicalists have argued for. To establish this within an interventionist framework, a further condition must be met. Let X_i be a set of upper-level variables and Y_k be a set of variables corresponding to their lower-level realizers. And let us say that a set of variables is unconditionally relevant to some effect E just in case it is causally relevant in the sense spelled out in M*. We can then define a notion of conditional *ir*relevance as follows:

> CI: A set of variables Y_k is irrelevant to variable E *conditional* on additional variables X_i if the X_i are unconditionally relevant to E, the Y_k are unconditionally relevant to E, *and* conditional on the values of X_i, changes in the value of Y_k produced by interventions and consistent with these values for X_i are (unconditionally) irrelevant to E (Woodward, 2021, pp. 253–254; italics in original).[17]

It is appropriate to speak of higher-order causation and of the autonomy of higher-level explanations only when the lower-level realizers Y_k of the higher-level variables X_i are conditionally irrelevant in the sense just defined.[18] When the lower-level realizers are conditionally irrelevant in this way, we can speak of the higher-order variables themselves being realization insensitive.

Some examples will help illustrate. Consider the relation of thermodynamic variables to those of statistical mechanics. The temperature of a gas (one of the upper-level or X_i variables) is unconditionally relevant to a reading on a thermometer placed inside of it (the effect variable E), and so are the combinations of the kinetic energies of the molecules composing the gas (the lower-level or Y_k variables).[19] Moreover, the

[17] Though Woodward's use of 'if' instead of 'if and only if' suggests that these conditions are only intended to be sufficient for conditional irrelevance, this is not the case, as his clarification in an appended footnote makes clear: "conditional irrelevance is much stronger than multiple realizability. The latter requires only that some different values of the same or different micro-variables(s) realize the same value of a macro-variable. Conditional irrelevance *requires* that all variations at the micro-level consistent with the value of the macro-variable make no difference to E. As this observation suggests, multiple realizability is not sufficient for autonomy understood in terms of conditional irrelevance" (2021, p. 254; italics added and omitted). See also Woodward's (2021, pp. 258–259) discussion of the potential failure of psychological variables to meet CI with regard to neurological variables, where this is treated as a failure of autonomy. Clearly, then, in both its use and intent, CI is treated as expressing both a sufficient and a necessary condition. See also Woodward (2020, p. 866), where it is said that "for most arbitrary sets of Us, Ls, and Es, conditional independence, or even approximate conditional independence, will *fail* …. The interesting question is the extent to which there are cases in which conditional independence or something like it does hold." These remarks and the surrounding discussion also clearly imply that the stated conditions are necessary.

[18] CI is needed to establish the autonomy or *distinctive* causal efficacy of higher-order properties. It is not needed to establish the causal relevance of higher-order properties, full stop. Indeed, it presupposes this sort of causal relevance. The point is that if higher-order properties meet the constraints imposed by M* without meeting those imposed by CI, they make no independent causal contribution and are therefore explanatorily and ontologically otiose. Failure to meet CI would thus leave mental properties in a position similar to that envisioned by the exclusion argument. See Woodward (2018) for further discussion.

[19] It is true that many changes to these kinetic energies will not affect E—those changes that balance each other out and thus don't affect the average—but recall the M* requires only that *some* changes are associated with changes to E. This, again, illustrates just how weak a constraint M* is.

relationship of the latter to the former is that of a many-one surjective function. As temperature is average kinetic energy, a large number of combinations of molecular kinetic energies will realize the same temperature. And, finally, when temperature is held constant, changes to the molecular kinetic energies consistent with this temperature will not be associated with changes to E. Temperature thus represents a permissible 'coarse-graining' of molecular kinetic energy. It is a coarse-graining in the sense that it represents a reduction in the dimensionality or degrees of freedom of the lower-level variable, and it is permissible in the sense that no information concerning causation or difference-making is lost in the translation.

A second example comes from Yablo (1992). Consider a pigeon, Sophie, trained to peck at discs that are any shade of red. The upper-level variable is thus the determinable property of being red; its lower-level realizers are the determinates of this determinable. Both the determinable and its determinates are unconditionally relevant to the effect, at least given the weak standard set by M*.[20] But conditional on the selection of the upper-level variable—let it be a binary variable taking the values red and non-red—details about the lower-level realizers are irrelevant. The effect of scarlet will be no different from the effect of maroon; the effect of cyan will be no different from the effect of periwinkle. The different ways of being red and non-red make no difference, which allows us to say that CI is met and that the disc's being red is a higher-order cause of Sophie's pecking.

The question is now whether the alleged causal relations involving functional properties meet both of these proposed conditions on higher-order causation. I shall argue that they fail to meet either.

3 Functionalism and independent fixability

Functional properties are causal role properties, second-order properties that objects instantiate in virtue of their various first-order properties[21] occupying certain causal roles.[22] To be in pain, for example, is to be in some state or other that causes avoidance behavior and is caused by tissue damage. C-fiber firing realizes pain in humans, at least according to philosophical lore, because it is the first-order state that causes avoidance behavior and is caused by tissue damage in beings like us. In other organisms there could be other states occupying this role. In Lewis's (1983) Martians, for example, it is the inflation of cavities in the feet that occupies the causal role constitutive of pain.

[20] Again, the lower-level realizers—such as being maroon, for example—are unconditionally relevant by the weak standard set by M*. *Some* changes to this variable—the change from being maroon to being cyan, for example—are associated with changes to the effect variable. It is irrelevant that some other changes are not. It is only if one adopts a stronger criterion, such as the more restrictive M** discussed below, that higher-order causes exclude lower-level causes. Adopting such a standard would also require us to modify the definition of conditional irrelevance.

[21] Or whatever type of entity it is that stands in first-order causal relations. There's no need to take a stand on the nature of causal relata, as interventionism, we have seen, takes no such stand. I'll speak of properties and states—instantiations of a property by an object at a time—for ease of presentation. Everything I say can be translated into an alternative idiom.

[22] This is the standard definition, found in, e.g., Block (1978), Shoemaker (1981), and in Kim's textbook introduction to functionalism (2011, p. 183).

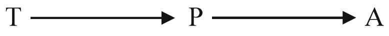

Fig. 5 Causal graph of the pain role

Implicit in this account of functional properties as causal role properties is a distinction between what (Shoemaker, 1981) calls the core and the total realizer of a functional property. C-fiber firing is the core realizer of pain in us. It is the state that occupies the causal role constitutive of pain. It is that state that comes and goes when pain comes and goes. But the core realizer of a property is not metaphysically sufficient for the realization of that property. There would be no realization of pain if C-fibers were to fire away in a petri dish or if C-fibers occupied a different causal role.[23] What is metaphysically sufficient for the realization of pain is only its total realizer—the occurrence of C-fiber firing in the appropriate causal context. It is only when C-fiber firing occurs downstream from tissue damage and upstream from avoidance behavior that it (total-) realizes pain. This distinction between core and total realizers matters because while, e.g., C-fiber firing is sufficiently distinct from avoidance behavior to be its cause, this is true of neither pain itself nor its total realizer.[24]

If pain is defined as the state that causes avoidance behavior—if C-fiber firing realizes pain only insofar as it causes avoidance behavior—then any causal model including separate variables for pain and avoidance behavior violates IF. Consider the causal graph in Fig. 5 below, which represents the causal role of pain. Let T represent tissue damage, P represent pain, and A represent avoidance behavior, and let each of these be binary variables that take the values of 1 or 0 depending upon whether or not the associated property is instantiated.

For this variable set to meet the conditions imposed by IF, it must be possible for $P = 1$ and $A = 0$—that is, it must be possible for pain to occur in the absence of avoidance behavior. But this possibility is ruled out by the definitional role of pain. It follows that the variable set incorporated into Fig. 5 violates IF and therefore does not relate variables that might enter into causal relations. There is, of course, no such difficulty if we imagine a variable C (for C-fiber firing) in the place of P. C-fibers can be stimulated independently, without having their characteristic causes or characteristic effects. C-fiber firing can cause avoidance behavior but pain cannot.

It doesn't help if we add temporal indices to our variables and make finer distinctions. Let P_t represent pain's occurrence at time t and let A_t represent a context existing at t such that if pain were to occur at t *and the context were to remain constant* avoidance behavior would occur. And let us say that it is A_t rather than A—the actual occurrence of avoidance behavior—that is necessary for the realization of pain. Then it would be possible for pain to occur in the absence of avoidance behavior, allowing IF to obtain and thereby allowing pain to cause avoidance behavior. The problem, however, is that this fails as a definition of pain, for it is consistent with pain never causing avoidance behavior. Imagine a creature in whom C-fiber firing *prevents*

[23] This manner of illustrating the distinction comes from Bennett (2003, p. 485).

[24] Failure to attend fully to the distinction between core and total realizers has perhaps obscured this problem. C-fiber firing is sufficiently distinct from pain's characteristic effects to be their cause, and it is true that C-fiber firing 'realizes' pain, but it does not follow that pain is sufficiently distinct from its characteristic effects to be their cause, for C-fiber firing is only the core and not the total realizer of pain.

avoidance behavior from occurring by altering the causal context that obtains at t. As long as A_t obtains at t, such a creature would satisfy the imagined criterion for being in pain, but no functionalist would accept such a result. Pain must *sometimes* cause avoidance behavior.

Nor does it help to appeal to tendencies and probabilities. Let us allow our variables to be assigned probability distributions and adopt as our new constitutive rule that pain causes avoidance behavior with probability p. Now pain can occur in the absence of avoidance behavior—this sometimes happens naturally and we can also bring it about by means of an intervention. But this is irrelevant, for we are now adopting a variable set in which avoidance behavior is assigned a probability distribution rather than a binary value. What's required by IF is therefore that there be pain occurring without avoidance behavior occurring *with probability p*, and this is precisely what is ruled out by our new constitutive rule.

The problem, of course, is not simply that pain cannot cause avoidance behavior. As counterintuitive as that conclusion is, it is but a faint echo of the real problem. On the functionalist account, pain is defined as the state that causes avoidance behavior. But if pain is defined as the state that causes avoidance behavior, pain and avoidance behavior violate IF, a necessary condition for entering into a causal relation, in which case pain cannot cause avoidance behavior. We have arrived at a contradiction. The real problem is thus that we have a *reductio ad absurdum* of the proposition that mental states are functional states.

It might be thought that this is simply a reformulation of what (Rupert, 2006) calls the problem of metaphysically necessary effects: if pain is defined as the state that causes avoidance behavior, then it is metaphysically necessary that pain causes avoidance behavior, and this conflicts with the widespread assumption that causal relations are metaphysically contingent.[25] I believe that these problems are distinct and I shall shortly argue that this is so, but suppose that they are not. Even if IF is just a Humean metaphysic in disguise, the preceding argument nevertheless establishes something new, which is that IF is inconsistent with functionalism. This, of course, means that functionalism cannot appeal to IF in its response to the exclusion argument and thus that an important line of defense has been lost. One cannot object to variable sets containing metaphysically necessary relations and then allow that causal relations themselves may be metaphysically necessary.

There's reason to think that these problems are distinct, however, for IF is concerned with the distinctness of the causal relata rather than with the strength of the causal relation. An intervention on a variable overrides its existing causal influences—it 'breaks' the arrows directed into the variable, as Woodward sometimes says—by causing the variable to adopt a new value by means of the intervention. This situation is depicted graphically in Fig. 6. If the values of X and Y are to be set by means of interventions, as IF proposes, the value of Y is *not* being set by the value of X and the

[25] A reviewer points out that (Block's, 1989) 'dormitive virtue' objection to functionalism is also similar to the one being discussed here. For Block, the crux of the problem is that functional properties are defined in terms of their effects. Interestingly, Block believes that functionalists can avoid his objection by adopting a counterfactual-based theory of causation: "Here … the lesson is that if you want to avoid epiphenomenalism, go for a counterfactual theory of causal relevance, not a nomological theory" (p. 159). But this is only true if counterfactual-based accounts do not adopt IF, and, as Woodward argues, such accounts are not viable.

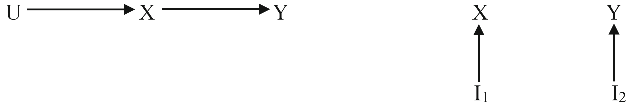

Fig. 6 The causal influences on X and Y before (left) and after (right) interventions

question of how the two variables relate when they are causally related is irrelevant. What IF outlaws are metaphysically necessary relations between variables *when the causal relations between these variables are broken.* It places no such restrictions on the causal relation itself.

This reply presupposes that the relation between X and Y can be overridden or broken, which might seem to contradict any assumption of a metaphysically necessary causal connection. But this is not so. Causal models are invariably partial, abstracting away from certain bits of information that might be causally relevant. This is why M* defines the casual relation between X and Y as being relative to some background circumstances B. But it is perfectly consistent to say that X = x necessitates Y = y in circumstances B and not in circumstances B′, the latter being circumstances in which the values of X and Y are set by means of interventions. It is nomically (and let us even suppose metaphysically) necessary that water boils at 100 °C in circumstances B. It is nevertheless also true that we can arrange by means of an intervention—by increasing atmospheric pressure, for example—a situation in which the water is at 100 °C and does not boil. The intervention changes the background circumstances, and there is no contradiction.

Causal relations are context dependent. Relations of indistinctness, by contrast, are not. If X is a proper part of Y in context B, it continues to be a proper part of Y however we may vary B. Consider the contrast between the following two statements:

(1) C-fiber firing causes avoidance behavior
(2) Pain causes avoidance behavior.

(1) may be true and even metaphysically necessary relative to some set of background conditions, but it is nevertheless easy to imagine interventions that allow C-fiber firing to occur in the absence of avoidance behavior. We might, for example, sever the neural pathways between the C-fibers and the neurons responsible for initiating avoidance behavior. The values of these variables can thus be fixed independently by varying the background conditions in which they occur. But this is not possible in the case of (2). There are no interventions, no background conditions, that allow pain to occur in the absence of avoidance behavior. This is not simply because the relation of pain to avoidance behavior is metaphysically necessary, but because it is one of indistinctness.

There is, of course, one way of conceiving of the causal relation that does not make it relative to background conditions. What (Mill, 1843) calls the ‘real cause’ does not depend on context—for the simple reason that it includes the complete set of conditions, both positive and negative, that together are sufficient for the effect. If a causal model were, *per impossible*, to contain a separate variable for each of these conditions, it would thereby violate IF. This shows that IF is not a perfect test of indistinctness. But there are two things to note. First, the issue here has nothing to do with the strength of the causal connection. Imagine a causal model containing a separate variable for each component of the real cause and suppose that the real cause only nomically necessitates its effect. It is nevertheless true that IF does not obtain in the model. Woodward explicitly states that IF is meant to include 'causal' possibility, which I take to mean nomic possibility. But if the real cause nomically necessitates its effect—if, that is, setting each element of the real cause to 1 ($R_1 = 1$; $R_2 = 1$, and so on) nomically necessitates that the value of E is 1—it is not causally (or nomically) possible for an intervention to set the various Rs to 1 and E to 0, in violation of IF. For recall that the real cause is complete, comprising all conditions, both positive and negative, that together suffice for the effect. As such, there are no 'external' conditions one can manipulate in order to prevent the real cause from having its effect. Any such external conditions would already be internal, by the definition of the real cause. Second, this is not a loophole that the functionalist can exploit, for total realizers are not real causes. Simply consider all of the conditions that must not obtain if avoidance behavior is to occur. To single out just one of an indefinite number of such conditions, there must not be a meteor about to strike the unfortunate person whose C-fibers are firing just before avoidance behavior is to be initiated. This is part of the real cause of avoidance behavior, but if it is also thereby part of the total realizer of pain, the latter will have an intolerably large supervenience base. It will include vast regions of the surrounding space, for these must be appropriately meteor-free, and this brings with it other absurdities that there is no independent reason to embrace—that our pains overlap, for example.

4 Functionalism and conditional irrelevance

Let us turn now to the second of the two conditions on higher-order causation and consider whether functional variables meet the conditions imposed by CI. One of the widely advertised features of functional properties is that they are multiply realizable by their lower-level realizers. Indeed, because functional properties are second-order properties realizable by any first-order properties occupying the appropriate causal role, and because a wide variety of first-order properties can occupy any given functional role, the set of possible realizers is often thought to be large and heterogenous. Humans, octopuses, and cyborgs can all experience pain because, despite their very different physical constitutions, each has a state that plays the pain role. The various first-order properties that occupy this role need have nothing in common save the fact that they realize the second-order property. It is this widely acknowledged feature of functionalism that comes into conflict with CI.

Let P be the higher-level variable representing pain and let us assume, with no loss of generality, that it is a binary variable taking the values of 1 or 0 depending upon whether a given organism is in pain. Let R be the lower-level variable representing the various realizers of pain. Clearly, R is not a binary variable, for R must have at least as many values as there are possible realizers of pain. And, of course, R must have a value corresponding to the situation in which no realizer of pain is instantiated. For simplicity, and with no loss of generality, let us simply consider the realizers of pain in humans, octopuses, and cyborgs. R will then adopt one of four possible values: 1, for humans, 2, for octopuses, 3, for cyborgs, and 0 for the case in which no realizer is instantiated. And, finally, let us assume that there is some binary variable E, which adopts the values of 1 or 0 depending upon whether some property distinct from pain (and its realizers) is instantiated.

Now suppose, for *reductio*, that pain causes some state E in humans. It follows, from M* that P must be unconditionally relevant to E, which means that an intervention changing the value of P from 1 to 0 must change the value of E from 1 to 0. Given our assumptions, it also follows that R is unconditionally relevant to E, for $P = 1$ entails that $R \neq 0$ and $P = 0$ entails $R = 0$. But it is not true that the value of R is conditionally irrelevant to the value of E. The set of realizers consistent with $P = 1$ is $\{1, 2, 3\}$ corresponding to the different possible realizers of pain in humans, octopuses, and cyborgs, respectively. Thus for it to be true that the value of R is conditionally irrelevant to the value of E, it would have to be the case that it doesn't matter to the value of E whether $R = 1$ or 2 or 3. But this is just ludicrously implausible. If we were to replace the realizer of pain in humans with the corresponding state of a cyborg it is highly unlikely that avoidance behavior would still occur. It is far more likely that nothing would occur save a seizure or death or a general malfunctioning of neural circuitry. Conditional irrelevance implies that the different realizers of psychological states are *interchangeable*, and this just isn't true. It follows that pain does not, in fact, cause E, and we have arrived at our contradiction.[26]

The point is not simply that the different core realizers of a mental state are not interchangeable with each other. This is undoubtedly true. Among the many reasons C-fibers will not cause avoidance behavior in cyborgs is that human neurons transmit signals by means of various neurotransmitters and integrated circuits do not. But the functionalist will justifiably complain that CI is meant to apply to total realizers rather than to core realizers, for it is only the former that are truly realizers in the intended sense of the term.[27] The point is well taken, but it doesn't affect the critical premise of the *reductio* because total realizers are no more interchangeable than core realizers are—at least, there's no reason to believe that they are. The total realizer of pain in humans, whatever precisely it may be, transmits and receives neurochemical signals and the corresponding state of the cyborg does not. Indeed, this failure of intersubstititivity is a general feature of functional states and functional systems. One cannot replace the heart of a mouse with that of an elephant; nor can one replace the entire circulatory system of the mouse with that of an elephant. The same holds for functionally identical parts of different watches or functionally identical parts of different cars. If functionalists wish to maintain that cognitive states and cognitive systems are an exception to this general rule, we are owed some reason as to why.

That we arrive at a contradiction does not, of course, tell us which of the premises leading to it must be rejected. If we maintain our commitment to CI and the assumption that each of the various Rs are realizers of pain, we must reject the assumption that

[26] I argue for a similar conclusion in Rellihan (2021), but from within the powers subset framework of Shoemaker (2007) and Wilson (2011). That framework is usually interpreted as presupposing a productionist account of causation, so it is useful to see that a similar conclusion follows from Woodward's interventionist account.

[27] Whatever else is true of the relation of a realizer to the property it realizes, it must be the case that the instantiation of the former is sufficient for the instantiation of the latter, and this is true of total rather than of core realizers.

pain causes E. E, however, is given no prior identity—it is simply a state caused by pain—so this is tantamount to a rejection of pain's causal efficacy across the board. But, as a reviewer proposes, we could also maintain CI and reject the assumption that the various Rs are all realizers of pain. Some are realizers of pain in humans, others are realizers of pain in other creatures. This, however, is tantamount to rejecting pain itself as a distinct and causally autonomous property, something functionalists will not allow. If we add a commitment to (Armstrong's, 1978) eminently reasonable Eleatic Principle, according to which to be is to have causal powers, we are also thereby forced to reject the reality of pain in favor of pain-in-human, pain-in-cyborgs, etc.[28] And this, again, is something functionalists will not allow. Neither implication is consistent with the claim that pain is multiply realizable, at least as that claim has traditionally been interpreted. The point therefore remains that CI is inconsistent with some of functionalism's deepest commitments.

Because the idea of interchangeable realizers is so ludicrous in the psychological case, it might be thought that we are misinterpreting or misapplying Woodward's condition. But consider again the paradigm of thermodynamics. Let T be the temperature of a gas, K the combination of all of the kinetic energies of each of the molecules constituting the gas. Because temperature is average kinetic molecular energy, any given temperature $T = t_1$ is realized by a large number of different molecular combinations—namely, all of those that have $T = t_1$ as their average. Suppose that the various realizers of $T = t_1$ are $\{k_1, k_2, \ldots, k_{n-1}, k_n\}$, and suppose that there is some reading on a mercury thermometer $M = m_1$ such that an intervention setting T to t_1 sets M to m_1. For conditional irrelevance, as we are interpreting it, to obtain, it would have to be the case that each of the various $k_1 \ldots k_n$ are interchangeable with each other and that no substitution of one for another would result in a difference to M. But this is exactly what occurs when a gas at constant temperature is measured over some interval of time, as kinetic combinations change and average kinetic energy does not. Conditional on the value $T = t_1$, no variation in K results in a variation to M, which is just what CI requires.

Nor must the various realizers of T be of the same physical type. If different systems consisting of different gases at the same temperature are allowed to intermingle, then, assuming they do not react, temperature will remain constant, and the gases will have the same effect on the thermometer before and after their combination. Thus it's not simply that the specific combinations of kinetic energies are conditionally irrelevant, the chemical identities of their constituent molecules are as well. We are not therefore interpreting CI in a bizarre or an unfairly restrictive way in our objection. Thermodynamic properties meet the standard. Functional properties do not.

The same point can be made in the case of other determinables. Consider again the case of the pigeon trained to peck at red discs. Let R take the value of 1 or 0 depending

[28] Kim (1992) is well known for introducing this principle—which he calls 'Alexander's dictum'—into the mental causation debate. Kim (1993) elsewhere considers the sort of 'local reductionism' discussed above and argues that it is inconsistent with the view that multiply realized properties are genuine or scientific kinds: "It must be admitted," he says, "that pain as a kind does not survive multiple local reduction" (p. 333). Genuine kinds, he argues, are individuated by their causal powers, and local reductionism denies that multiply realized properties are causally homogenous. CI is one way of precisifying the relevant sort of causal homogeneity, so the arguments I make in the body of the paper can be used to reinforce Kim's conclusion.

upon whether the disc is red, and let P take the values of 1 or 0 depending upon whether the pigeon pecks. R is realized by the various determinate shades of red. Let D represent these determinate shades, and let it take the values 1, 2, 3, and 0 for scarlet, crimson, maroon, and nonred, respectively. It's easy to see that, once again, the value of D is irrelevant conditional on the value of R. If R is 1, and the pigeon is trained to peck at red, it doesn't matter how red is realized. Each determinate shade will have the same effect on the pigeon. Thus the problem is not that CI is unrealistically restrictive or that we have been misapplying it. The problem is that this condition is simply not satisfied by functional properties as they are standardly interpreted to be.[29]

Consider, finally, (Woodward's, 2008, pp. 238–240) own illustration of higher-order causation in the psychological realm. Just after laying out the paradigmatic cases of thermodynamics and Yablo's pigeons, he reports that research conducted by his colleagues at Caltech[30] shows that a similar phenomenon occurs in the neural implementations of higher-order intentions in the macaque monkey.[31] Variations in intention (and the behavioral effects that derive therefrom) are systematically related to variations in *aggregate* firing rates of the individual neurons realizing the intention. The specific firing profiles of the specific neurons composing the group are conditionally irrelevant—as long as the aggregate profile is held constant, variations in the realizers of this aggregate are irrelevant to the effect. This example clearly supports the idea that different realizers have to be interchangeable for CI to obtain. If this is the standard, functionalism fails to meet it.

It might seem that there's a way out of this difficulty if we match the coarse-graining of our cause variable with a suitably coarse-grained effect variable. C-fiber firing causes human avoidance behavior, e_1; the inflation of pedal cavities causes Martian avoidance behavior, e_2. Pain, as such, causes neither of these. What it causes is simply avoidance behavior, E—where this is understood as a realized property having e_1 and e_2 among its many realizers.

But this response is a nonstarter for a number of reasons. First, it does little to assuage worries about mental causation. We want it to be the case that our individual behaviors are caused by our individual mental states—that, e.g., my reaching for an aspirin is caused by my headache—but the former are precisely the sort of fine-grained effect variables that are being dispensed with on the current suggestion. Moreover, such a defense of higher-order causation amounts to little more than a bait-and-switch. Recall that properties like being such-and-such a temperature or being red were supposed to be the paradigm—uncontroversial cases in which higher-order properties figure into causal relations. We've seen, though, that these paradigm cases meet the more exacting standard. The claim is not that the property of being red earns its explanatory keep

[29] It's also worth noting that in both the thermodynamic case and in the case of determinables like being red the relevant causal models meet the conditions imposed by IF.

[30] The research is described in (Musallam et al., 2004).

[31] Woodward is explicit in making the comparison between the cases: "The preference for micro or fine-grained causation that we are considering recommends that we should regard [the neural realizer] as the real cause of [the effect] on occasion t. But this seems wrong for the same reason it seems wrong to say that it is the scarlet color of the target that causes the pigeon to peck in circumstances in which the pigeon will peck at any red target and wrong to say that it is the specific molecular configuration G1 rather than the fact that the temperature of the gas has been increased to T2 which is responsible for the new pressure P2" (2008, pp. 239).

because one shade of red causes Sophie's pecking while another shade causes the pecking of another pigeon in another lab. The effect variable is fine-grained. Any shade of red causes *Sophie's* pecking, and it is for this reason that we say that her pecking is caused by the determinable rather than its determinates.

The imagined response, moreover, makes higher-order causation come too cheap. Suppose I claim that there is some causally relevant higher-order property the realizers of which are hurricanes and fires and provide as evidence that they have the common effect of causing either floods or smoke. One cannot respond that here the coarse-grained effect is spurious whereas in the case of abstract avoidance behavior it is not, for this is the very question at issue. We are trying to determine which higher-order properties are genuine and which are not, and it was supposed to be causation that was our guide. Avoidance behavior is a genuine higher-order property only if it enters into genuine higher-order causal relations of the sort that CI was intended to describe, and this, of course, is the very question at issue.

Finally, there's a case to be made for consistently adopting CI as a test of the genuineness of both higher-order causes and higher-order effects. Doing so requires us to say that X causes Y only if any realizer of X causes any realizer of Y—that pain, regardless of its specific realization, causes avoidance behavior, regardless of its specific realization. But this, again, is what is being denied. C-fiber firing causes human avoidance behavior. It doesn't, and wouldn't, cause Martian avoidance behavior. Thus, CI is inconsistent with coarse-grained effect variables of the sort being proposed. A consistent application of CI counsels us to reject as spurious the pseudo-properties corresponding to such variables.

The preceding argument focuses on the multiple realizability of functional properties and shows that this leads to a conflict with CI. We'll now see that an independent route leads to this conclusion—one that rests not on multiple realizability but on the distinction between core and total realizers.

It will be useful to begin with a variation on a previous example. Suppose, as before, that increases in LDL cause increases in arterial blockage and that TC is the sum of LDL and HDL. But now suppose that HDL has no effect whatsoever on levels of arterial blockage. It seems incorrect to say in such circumstances that TC is causally relevant to A, for TC contains an idle component that does no causal work. And this is indeed what follows from the more restrictive account of causal relevance that Woodward sometimes considers (but does not ultimately endorse).

> M**: X causes Y in *B* if and only if there are distinct values of X, x_1 and x_2, with $x_1 \neq x_2$, and distinct values of Y, y_1 and y_2 with $y_1 \neq y_2$*such that under all interventions* in *B* which change the value of X from x_1 to x_2, then Y would change from y_1 to y_2 (Woodward, 2021, p. 242, fn. 7; italics added).

Suppose that an intervention changing the value of LDL from x_1 to x_2 results in the value of A changing from y_1 to y_2. And assume, for simplicity, that this is the only intervention on LDL that alters the value of A. It follows that *some* intervention that changes the value of TC also changes the value of A—namely, the intervention that changes TC by changing LDL. But, assuming the idleness of HDL, it does not follow that *all* such interventions result in a change to A, for if we change the value of TC only by changing the value of HDL there will be no associated change in A. This means that TC fails to meet the M** condition of causal relevance. But it nevertheless continues to meet the M* condition, for that condition, recall, requires only that *some* of the interventions changing X from x_1 to x_2 be associated with changes in Y from y_1 to y_2.

It is debatable whether it is M* or M** that gives the correct condition. Woodward himself is not always clear on this issue.[32] But it doesn't matter for present purposes because even if we accept that TC is causally relevant under the less demanding M* standard, it nevertheless fails the test of conditional irrelevance and thereby the test for higher-order causation. This is because the values of the lower-level realizers of TC—the various LDL/HDL pairs that sum to the value of TC—are not conditionally irrelevant to the value of A. If we change TC by changing HDL and holding LDL constant, we can expect no change to A, but if we produce the same change to TC by holding HDL constant and changing LDL, we can expect A to change. The effect of TC on A depends on how it is realized, and for this reason TC does not enter into higher-order causal relations with A.

Total realizers are like total cholesterol in this regard. Total realizers, recall, are conjunctive properties, consisting of a core realizer and the context in which the instantiation of the core realizer is able to play the causal role constitutive of the functional property. Consider the case of pain. Restricting ourselves to pain in humans, we find that the total realizer of pain is (i) C-fiber firing in (ii) circumstances in which C-fiber firing is caused by tissue damage and causes avoidance behavior. Pain will cease to occur if either C-fiber firing ceases or C-fiber firing ceases to play the pain role. But whether C-fiber firing causes avoidance behavior is independent of whether C-fiber firing plays the pain role. To see this, simply imagine, with the aid of Fig. 5, a situation in which tissue damage causes C-fiber firing, which causes avoidance behavior (so imagine a *C* in the place of Fig. 5's *P*). And imagine an intervention that eliminates the connection between tissue damage and C-fiber firing just as the C-fibers are about to fire.[33] Because C-fiber firing no longer occurs in a context in which it is caused by avoidance behavior, it no longer realizes pain. But because the connection between C-fiber firing and avoidance behavior remains untouched, C-fiber firing will continue to cause avoidance behavior. We thus find that total realizers, like total cholesterol in the foregoing example, consist of a causally relevant and a causal irrelevant element. Interventions on the former will result in changes to the presence of avoidance behavior, but interventions on the latter will not.[34]

Let's now make this more explicit. Our upper-level variable is P, which takes the values 1 or 0 depending upon whether pain is instantiated. Its lower-level realizer is

[32] See, for example, (Woodward 2021, p. 242) where he says that neither is clearly more correct than the other.

[33] I give more elaborate examples of such procedures and discuss their causal (but not necessarily interventionist) consequences in Rellihan (2019).

[34] Bennett (2003, 2008) recognizes the problem pointed out here but argues that the background context is necessary for the core realizer to have its effect. Keaton and Polger (2014) show that this is not generally the case by constructing a counterexample. I have argued elsewhere (Rellihan, 2019) that it is, in fact, very rarely the case. Simply consider the neuron diagrams for causal modeling that Lewis made famous—which, of course, are modeled on actual neurons. It is rarely the case that disrupting any one causal connection disrupts other, disparate regions of the network. The onus is at the very least on functionalists to give some reason to believe that Bennett's happy coincidence is the norm. See Rellihan (2019) for further discussion.

what we've been calling its total realizer, which itself consists of the conjunction of a core realizer and a context. Because we are restricting ourselves to the human case, we will assume that the core-realizer is C-fiber firing and the context is the one in which C-fiber firing is caused by tissue damage and causes avoidance behavior. Let T take the value of 1 or 0 depending upon whether this total realizer is instantiated, C take the value of 1 or 0 depending upon whether C-fiber firing occurs, and B take the value of 1 or 0 depending upon whether the necessary causal context is in place for C-fiber firing to realize pain. And, finally, let A take the value of 1 or 0 depending upon whether avoidance behavior occurs.[35] We find that by the generous standard of M*—but not by the more miserly standard of M**—P causes A, for there are some interventions on P—those corresponding to interventions on its core realizer—that result in changes to A. P, however, fails the test of conditional irrelevance, for some ways of changing the value of P from 1 to 0 result in changes to the value of A and some do not. When the value of P is held constant at 0, changes to the values of C and B are not irrelevant to the value of A. When P = 0, C = 0, and B = 1, the value of A is 0, but *sometimes* when P = 0, C = 1, and B = 0, the value of A is 1. If we alter B *without* disrupting the connection between C-fiber firing and avoidance behavior, avoidance behavior will still occur even though pain does not. This violates CI, and pain cannot therefore be said to be a higher-order cause of avoidance behavior.

This argument is, admittedly, hampered by the simplicity of the example. We're imagining that pain has one constitutive cause and one constitutive effect, which makes it more difficult to see that total realizers can be modified without affecting the ability of core realizers to bring about their effects. The remedy is to imagine only a slightly more complex case. Suppose that pain can be caused either by tissue damage or by existential angst, and suppose that at just this moment it is being caused by the former and not the latter. We could vary the causal context and thus the total realizer by eliminating the inactive causal connection from angst to C-fiber firing without at all affecting the actual causal process in which tissue damage causes pain to cause avoidance behavior. The inactive causal connection is causally inert and its presence therefore makes no causal difference. But, by hypothesis, it makes a difference to whether or not pain is realized. Thus, even if the simple case fails as a counterexample—something I don't believe to be the case but will concede for the sake of argument—only slightly more complex cases succeed. The functionalist would then be put in the uncomfortable position of having to choose between (nearly) maximally simple functional roles for mental states and epiphenomenalism.

We find, then, that functional properties like pain violate CI in two ways. First it matters how they are realized. Human pain has very different effects from octopus pain, Martian pain, or cyborg pain. Second, it matters how they aren't realized. If pain fails to be realized in the human case because a given person's C-fibers are not firing, avoidance behavior will not occur. But if that very same person fails to experience pain *not* because her C-fibers are not firing but because they are not firing in the

[35] Note that this variable set does not meet IF, for P cannot be set independently of T and T cannot be set independently of C and B. This is irrelevant for present purposes, however, for we can simply test each of the potential causes of A independently. In a variable set that includes P but not T, C, or B, P causes A. In one that includes T but not P, C, or B, T causes A. And so on. This procedure suffices to show that both P and its various realizers are unconditionally relevant to A. We can then apply the CI test as in the text.

appropriate context—some causally inert but necessary part of this context having been removed—avoidance behavior likely will still occur. Either way, CI rules that pain does not cause avoidance behavior.

5 Conclusion

Interventionism is an attractive account of causation for anyone who believes that causal relations can sometimes hold between higher-level entities. It, or something like it, is implicit and sometimes even explicit in the practice of many of the special sciences, ranging from economics to biology and medicine. And, as we've seen, it also possesses the resources to rebut certain theoretical objections to higher-order causation of the sort found in recent philosophy of mind. But when the interventionist account of higher-order causation is more fully developed—when its account of higher-order causation moves beyond the merely programmatic to the nitty–gritty details—problems emerge for an important class of these higher-order entities. Functional properties, of the sort made most familiar in contemporary philosophy of mind and cognitive science but which are also to be found in various other domains, don't appear to meet the interventionist standard. I leave it as an open question whether it is functionalism or interventionism that should be rejected, but if even the seemingly friendly terrain of interventionism proves hostile to functionalism, one can't help but wonder if it isn't functionalism itself that is at fault.

Declarations

Conflict of interest The author certifies that he/she has no conflict of interest regarding the submission of this manuscript. The manuscript is the author's original work. It has not been submitted or published in whole or in part elsewhere.

References

Antony, L., & Levine, J. (1997). Reduction with autonomy. *Philosophical Perspectives, 11*, 83–105.

Armstrong, D. (1978). *A theory of universals: Universals and scientific realism* (Vol. II). Cambridge University Press.

Baker, L. R. (2009). Non-reductive materialism. In B. McLaughlin & A. Beckermann (Eds.), *The Oxford handbook of philosophy of mind* (pp. 109–127). Oxford University Press.

Baumgartner, M. (2009). Interventionist causal exclusion and non-reductive physicalism. *International Studies in the Philosophy of Science, 23*(2), 161–178.

Baumgartner, M. (2010). Interventionism and epiphenomenalism. *Canadian Journal of Philosophy, 40*, 359–383.

Bennett, K. (2003). Why the exclusion problem seems intractable and how, just maybe, to tract it. *Nous, 37*(3), 471–497.

Bennett, K. (2008). Exclusion again. In J. Hohwy & J. Kallestrup (Eds.), *Being reduced* (pp. 280–307). Oxford University Press.

Block, N. (1978). Troubles with functionalism. *Minnesota Studies in the Philosophy of Science, 9*, 261–325.

Block, N. (1989). Can the mind change the world? In G. Boolos (Ed.), *Meaning and Method: Essays in Honor of Hilary Putnam* (pp. 137–170). Cambridge University Press.

Chalmer, D. (1996). *The conscious mind*. Oxford University Press.

Crane, T. (2001). *The elements of mind*. Oxford University Press.

Davidson, D. (1980). Mental events. *Essays on actions and events* (pp. 207–224). Oxford University Press.

Fodor, J. (1974). Special sciences, or the disunity of science as a working hypothesis. *Synthese, 28*, 97–115.

Hesslow, G. (1976). Two notes on the probabilistic approach to causality. *Philosophy of Science, 43*(2), 290–292.

Keaton, D., & Polger, T. (2014). Exclusion, still not tracted. *Philosophical Studies, 171*(1), 135–148.

Kim, J. (1992). 'Downward causation' in Emergentism and Nonreductive Physicalism. In A. Beckermann, H. Flohr, & J. Kim (Eds.), *Emergence or Reduction?* (pp. 119–138). de Gruyter.

Kim, J. (1993). Multiple realization and the metaphysics of reduction. In J. Kim (Ed.), *Supervenience and mind* (pp. 309–335). Cambridge University Press.

Kim, J. (1998). *Mind in a physical world*. MIT.

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

Kim, J. (2009). Mental causation. In B. McLaughlin, A. Beckermann, & S. Walter (Eds.), *The Oxford handbook of philosophy of mind* (pp. 29–52). Oxford University Press.

Kim, J. (2011). *Philosophy of mind*. Westview.

Lewis, D. (1973). Causation. *Journal of Philosophy, 70*(17), 556–567.

Lewis, D. (1983). Mad pain and Martian pain. *Philosophical papers* (Vol. I, pp. 122–130). Oxford University Press.

Lewis, D. (1986). Postscript to 'Causation.' *Philosophical papers* (Vol. II, pp. 172–213). Oxford University Press.

List, C., & Menzies, P. (2009). Nonreductive physicalism and the limits of the exclusion principle. *Journal of Philosophy, 1006*, 475–502.

List, C., & Menzies, P. (2010). The causal autonomy of the special sciences. In Ed. McDonald & Ed. McDonald (Eds.), *Emergence in mind* (pp. 108–129). Oxford University Press.

Loewer, B. (2002). Review of mind in a physical world. *Philosophy and Phenomenological Research, 65*(3), 655–662.

Loewer, B. (2007). Mental causation or something near enough. In B. P. McLaughlin & J. Cohen (Eds.), *Contemporary debates in philosophy of mind* (pp. 243–264). Blackwell.

Mill, J. S. (1843). *A system of logic, ratiocinative and inductive*. University of Toronto Press.

Musallam, S., Corneil, B., Greger, B., Scherberger, H., & Andersen, R. (2004). Cognitive control signals for neural prostethics. *Science, 305*, 258–262.

Paul, L. A., & Hall, N. (2013). *Causation: A user's guide*. Oxford University Press.

Polger, T., & Shapiro, L. (2016). *The multiple realization book*. Oxford 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). The nature of mental states. *Mind, language, and reality* (pp. 429–440). Cambridge University Press.

Raatikainen, P. (2010). Causation, exclusion, and the special sciences. *Erkenntnis, 73*(3), 349–363.

Rellihan, M. (2019). Strengthening the exclusion argument. *Synthese, 198*(7), 6631–6659.

Rellihan, M. (2021). A familiar dilemma for the subset theory of realization. *Analytic Philosophy, 64*(1), 68–90.

Rupert, R. (2006). Functionalism, mental causation, and the problem of metaphysically necessary effects. *Noûs, 40*(2), 256–283.

Shapiro, L. (2012). Mental manipulations and the problem of causal exclusion. *Australasian Journal of Philosophy, 90*(3), 507–524.

Shoemaker, S. (1981). Some varieties of functionalism. *Philosophical Topics, 12*(1), 93–119.

Shoemaker, S. (2007). *Physical realization*. Oxford University Press.

Sprites, P., & Scheines, R. (2004). Causal inference of ambiguous manipulations. *Philosophy of Science, 71*(5), 833–845.

Wilson, J. (2011). Non-reductive physicalism and the powers-based subset strategy. *The Monist, 94*(1), 121–154.

Wilson, J. (2021). *Metaphysical emergence*. Oxford University Press.

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

Woodward, J. (2008). Mental causation and neural mechanisms. In J. Hohwy & J. Kallestrup (Eds.), *Being reduced: new essays on reduction, explanation, and causation* (pp. 218–262). Oxford University Press.

Woodward, J. (2015). Interventionism and causal exclusion. *Philosophy and Phenomenological Research, 91*(2), 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 University Press.

Woodward, J. (2018). Explanatory autonomy: The role of proportionality, stability, and conditional irrelevance. *Synthese, 198*(1), 1–29.

Woodward, J. (2020). Causal complexity, conditional independence, and downward causation. *Philosophy of Science, 87*(5), 857–867.

Woodward, J. (2021). Explanatory autonomy: The role of proportionality, stability, and conditional irrelevance. *Synthese, 198*, 237–265.

Yablo, S. (1992). Mental causation. *Philosophical Review, 101*(2), 245–280.

Zhong, L. (2020). Intervention, fixation, and supervenient causation. *Journal of Philosophy, 117*(6), 293–314.

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