Synthese (2018) 195:35–54
https://doi.org/10.1007/s11229-014-0564-5

# Mutual manipulability and causal inbetweenness

**Totte Harinen**

Received: 21 February 2014 / Accepted: 20 September 2014 / Published online: 10 October 2014


**Abstract** Carl Craver's mutual manipulability criterion aims to pick out all and only those components of a mechanism that are constitutively relevant with respect to a given phenomenon. In devising his criterion, Craver has made heavy use of the notion of an ideal intervention, which is a tool for illuminating causal concepts in causal models. The problem is that typical mechanistic models contain non-causal relations in addition to causal ones, which is why the standard concept of an ideal intervention is not appropriate in that context. In this paper, I first show how top-down interventions in mechanistic models violate the conditions for ideal interventions. Drawing from recent developments in the causal exclusion literature, I then argue for extended interventionism better suited for the purposes of the new mechanist. Finally, I show why adopting such an extended account leads to the surprising consequence that an important subset of mechanistic interlevel relations comes out as causal.

**Keywords** Mutual manipulability · Mechanisms · Supervenience · Realization · Interventionism · Causal inbetweenness

## 1 Introduction

The notion of an ideal intervention has been developed in the causality literature in an attempt to make sense of certain basic causal concepts in the context of models containing counterfactual dependency relations between variables (Woodward 2005). Craver (2007) has adopted the concept in characterizing his criterion of mutual manipulability, which aims to pick out all and only those components of a mechanism that

T. Harinen (✉)
Department of Philosophy, King's College London, The Strand, London WC2R 2LS, UK
e-mail: totte.harinen@kcl.ac.uk

are constitutively relevant for its behaviour. Mechanistic models, however, contain causal as well as *non-causal* relations, which is a problem for Craver's account of mutual manipulability. It is a problem because the presence of non-causal relations in mechanistic models renders ideal interventions, which Craver uses to define his concept of mutual manipulability, not possible or not likely in the mechanistic context. In this paper, I first pose the above problem to Craver and then propose to solve it by adopting an *extended* account of ideal interventions and by arguing that mutual manipulability relations are best understood as involving not two but *three* variables. After proposing my solution, I investigate what the consequences of its adoption are in the mechanistic context. One of those consequences turns out to be that an important subset of the interlevel relations in mechanistic models comes out as causal. Thus, the debate on the metaphysics of mutual manipulability will be advanced in three ways: (1) by discovering a new problem for Craver's account of mutual manipulability; (2) by solving that problem via (i) arguing for an account of ideal interventions suitable in the mechanistic context and (ii) unpacking mutual manipulability as a three-variable affair; and (3) by analyzing the implications of the proposed solution. The resulting picture of the metaphysics of mechanisms I call 'causal inbetweenness'. Why the name of my theory involves not only 'causal' but also 'inbetweenness' will become clear in the final section of the paper.

## 2 Mutual manipulability and the problem of constitutive relevance

It is now well established that we explain many things mechanistically (Bechtel and Richardson 1993; Machamer et al. 2000; Bechtel and Abrahamsen 2005; Craver 2007). We are not content to merely determine the regularities but we also want to know why those regularities hold (Cummins 2000). In psychology, neuroscience and elsewhere, a satisfactory explanation for the existence of some regularity is often a description of an underlying mechanism. Such descriptions specify how a mechanism exhibiting a behaviour comes to do so as the result of the behaviours of its components. This type of explanation is appropriately called *constitutive*, because is involves explaining the behaviour of a whole in terms of the behaviours of its parts.

That is the descriptive story, in very brief detail, but what about the prescriptive one? A good account of scientific explanation not only provides a faithful picture of the practice of science, but also gives some grounds for distinguishing good explanations from bad ones. In the mechanistic context, a plausible ideal for a good explanation is this: an explanation should describe *all and only* those components that are *relevant* with respect to the *explanandum* phenomenon. To get an idea of what this means, suppose I want to explain why my bicycle slows down when I squeeze the lever mounted to its handlebar; but in addition to cables, brake pads and rims, I also describe bartapes, bottle cages and dustcaps. The last three things are parts of my bike all right, but they aren't relevant parts with respect to the phenomenon to be explained, at least in the majority of circumstances. Similarly, my explanation can go wrong if it includes too little detail. For example, if I fail to mention calipers, bolts or cable guides, I have not yet provided a full account of the braking mechanism. A theory of mechanistic explanation should be able to determine the components that count as relevant and

to specify the nature of the relation that holds between those components and the *explanandum*.

Craver (2007) has recognized the importance of distinguishing genuine components from mere parts, and he has worked out an account of the norms of mechanistic explanation. His *mutual manipulability* criterion aims to pick out precisely those components that are 'constitutively relevant' in a mechanism with respect to a given phenomenon. An ideal mechanistic explanation of a phenomenon would then consist of a description of the organized behaviours of those constitutively relevant components. If successful, the mutual manipulability criterion would help to decide whether a putative mechanistic explanation describes the correct entities (constitutively relevant components) and to elucidate the nature of the relation between those entities and the *explanandum* (a relation of mutual manipulability). It would therefore be a great step forward in the process of developing a prescriptive account of mechanistic explanation.

In order to see how mutual manipulability works, let us introduce some notation. First, we have a mechanism S that 'engages in activity' $\psi$ where S's $\psi$-ing is understood as a complex input-output relationship. The inputs here include all of the conditions that are required for S to $\psi$. S has a set of components $\{X_1, X_2, \ldots, X_n\}$ that also engage in activities, $\phi_i$. ($X_i$'s $\phi_i$-ing can also be understood as a complex input-output relationship, given that the mechanistic decomposition is often thought to continue a couple of levels 'downwards' until it 'bottoms out' at some pragmatically determined level. Here, $X_i$'s $\phi_i$-ing can be intuitively thought to be at a lower mechanistic level than S itself because $X_i$ is a component in S's $\psi$-ing.)

According to Craver, the norms of constitutive relevance are in fact implicit in the experimental practice in neuroscience and elsewhere. There are two basic kinds of experiment: bottom-up and top-down. Bottom-up experiments include interference and stimulation experiments. In what Craver calls interference experiments, one prevents some suspected component X's $\phi$-ing, in some suspected mechanism S, and observes the resulting changes in S's $\psi$-ing. In what Craver calls stimulation experiments, one excites X's $\phi$-ing and again observes the changes in S's $\psi$-ing. The point in both kinds of experiment is to see whether one is able to manipulate S's $\psi$-ing by manipulating the $\phi$-ing of some of its putative components. Finally, in what Craver calls activation experiments, one varies the conditions for S's $\psi$-ing and observes whether changes in S's $\psi$-ing are accompanied with changes in some putative component X's $\phi$-ing. In this top-down intervention the point is to see whether one is able to manipulate X's $\phi$-ing by manipulating S's $\psi$-ing. Note that the similarity between the expressions 'stimulation experiment' and 'activation experiment' is potentially misleading: the two are sharply distinct because the former is a bottom-up experiment while the latter is a top-down experiment.

The idea, then, is that component X's $\phi$-ing is constitutively relevant for mechanism S's $\psi$-ing when the following conditions are satisfied:

> (i) X is part of S; (ii) in the conditions relevant to the request for explanation there is some change to X's $\phi$-ing that changes S's $\psi$-ing; and (iii) in the conditions relevant to the request for explanation there is some change to S's $\psi$-ing that changes X's $\phi$-ing (Craver 2007, p. 153).

Why is it required that the relationship be bidirectional? The argument is that neither top-down nor bottom-up experiments *alone* suffice to determine whether X's $\phi$-ing is relevant for S's $\psi$-ing. Suppose I want to know whether the mudguards in my bicycle are constitutive parts of its braking mechanism. As it happens, a simple intervention on the mudguards (loosening their fixings) changes the behaviour of the braking mechanism as a whole, by blocking the movement of the brake calipers and preventing the brake pads from reaching the rims. A change in a putative component, therefore, changes S's $\psi$-ing. Yet it would be hasty to conclude, on the grounds of this bottom-up intervention, that the mudguards are constitutive parts of the braking mechanism. For note that under normal circumstances it is *not* possible to loosen the fixings of the mudguards by squeezing the brake levers. A change in S's $\psi$-ing, in other words, doesn't result in a change in the component under investigation. This means that the manipulability relation between the mudguard and the braking mechanism is unidirectional and not bidirectional as required by Craver's conditions for constitutive relevance; the top-down intervention rules out the mudguards as constitutively relevant for the braking mechanism.[1]

Similarly, suppose I intervene on the braking mechanism as a whole, by squeezing the brake levers, in order to figure out whether my speedometer is one of its constitutive parts. This top-down intervention is not enough because there is a change in my speedometer (its reading) that systematically co-varies with changes in the braking mechanism as a whole. Thus, a change in S's $\psi$-ing *does* result in a change in the putative X's $\phi$-ing. Yet it is easy to rule out the speedometer as a part of the braking system by attempting to manipulate the behaviour of the braking mechanism as a whole by manipulating the behaviour of the speedometer. Even if I shut down the speedometer, the bike is still going fast (and the police won't have any of my excuses). This time it is the bottom-up intervention that rules out the putative component as constitutively relevant. Since mutual manipulability requires that X's $\phi$-ing passes both top-down *and* bottom-up tests, it delivers the correct result with problem cases such as these.

Finally, it is very easy to see why intuitively genuine components, such as the cables in a typical bicycle brake mechanism, successfully pass the mutual manipulability test. Squeezing the brake levers moves the cables in their housings; preventing the cables from moving makes the whole system unresponsive. So far so good.

## 3 The ideality of interventions

One of the innovations in Craver (2007) is the use of the manipulationist approach to causality to make sense of the relations holding between entities in mechanistic models. A central notion the manipulationist literature is that of an *ideal intervention,* which can be utilized to define various causal concepts (Woodward 2005). It turns out to be an important notion for Craver's purposes as well, for the changes in X's $\phi$-ing and

[1] As an anonymous reviewer points out, there is bound to be some interest-relativity as to how detailed one's mechanistic story of a phenomenon should be. The situation is analogous to causal modeling, where the appropriate grain of one's variables depends on the explanatory task at hand.

S's $\psi$-ing in conditions (ii) and (iii) for mutual manipulability are supposed to occur as follows: 'There should be some *ideal intervention* on $\phi$ under which $\psi$ changes, and there should be some *ideal intervention* on $\psi$ under which $\phi$ changes'(Craver 2007, p. 154, my emphasis). Focusing on the first case, Craver explicates how such ideal interventions are to be understood:

> (I1) the intervention I does not change $\psi$ directly;
> (I2) I does not change the value of some other variable $\phi$* that changes the value of $\psi$ except via the change introduced into $\phi$;
> (I3) I is not correlated with some other variable M that is causally independent of I and also a cause of $\psi$; and
> (I4) I fixes the value of $\phi$ in such a way as to screen off the contribution of $\phi$'s other causes to the value of $\phi$(Craver 2007, p. 154).

Why is it important that interventions satisfy the above conditions? In the causal literature, the answer is very simple. If one's interventions are ham-fisted, it will not be possible to isolate the variables[2] that are doing the causal work. Suppose my intervention on $\phi_1$ always changes the value of $\phi_2$, and that $\phi_2$ is causally related to $\psi$ via some path that is different to the one via which $\phi_1$ is causally related to $\psi$. Then it won't be possible for me to decide whether it is $\phi_1$ or $\phi_2$ that is doing the causal work with respect to $\psi$. The only way for me to do it would be to hold one of the variables fixed while wiggling the other. It is this same consideration that motivates the inclusion of the ideality conditions in the characterization of mutual manipulability. If one's intervention always changes the behaviours of several entities at a time in a way that conflicts with ideality, then it won't be possible for one to isolate the specific entity, or entities, that changed the behaviour of the mechanism as a whole.

Here I have been talking about ideal interventions on the component entities with respect to the behaviour of a mechanism as a whole. What about the opposite direction? As we recall, the interventions on the behaviour of a mechanism as a whole with respect to any of its components should also satisfy the conditions of ideality. But note two things. First, mechanistic models include multiple levels: the behaviour of a mechanism as a whole is thought to be at a higher level than the behaviour of its components. Second, the relationship between mechanistic levels is thought to be non-causal, and it is reasonable to require that the nature of this relationship is that of metaphysical supervenience or realization. As we will see in the next section, these two observations land us into trouble.

[2] Throughout this paper I will engage in variable-talk. Variables are very flexible in that they can be thought to correspond to all kinds of things. For example, if one wishes, it is possible to think of variables as corresponding to events, with the following two values: {occurs, doesn't occur} For many, this would be the appropriate interpretation for the purposes of causal analysis. However, the values of variables needn't correspond to events, or be binary, which is welcome given that there are many situations in which it is useful to have e.g. many-valued variables. Later I will also talk as if the values of variables would engage in supervenience or realization relations, in which case the variables are best interpreted as corresponding to properties.

## 4 The interlevel relations in mechanistic models

Figure 1 is a typical representation of a mechanistic model. The behaviour of the mechanism as a whole is at a higher level and the behaviours of its components are at a lower level. The relations *within* the levels, represented by the arrows, are causal. But the relations *between* the levels are generally thought to be non-causal (Craver and Bechtel 2007; Craver 2007, p. 153–154). It will be useful to divide the latter case, i.e. the *inter*level relations, into two types. First, we have the relation between a mechanism and its individual components, which is the type of relation picked out by the mutual manipulability criterion, and is not represented in Fig. 1 (but will be represented in Fig. 3). I will return to this relation in Sect. 6 where I argue, against Craver and others, that it is causal. Second, we have the relation between the mechanism and its *components-taken-together,* which is the type of relation on which I will focus in this section, and to which I refer when I talk about '$X_{1,\ldots,n}$'s $\phi_{1,\ldots n}$-ing'. This is the relationship that is represented by the dotted lines in Fig. 1 and with respect to which I agree with Craver and others that it is non-causal.

Is S's $\psi$-ing something *over and above* of the organized $\phi$-ings of all of the Xs passing the mutual manipulability test, that is, $X_{1,\ldots,n}$'s $\phi_{1,\ldots n}$-ing? Most philosophers and scientist would probably agree that there is some sense in which S's $\psi$-ing is indeed more than just the sum of the $\phi$-ings of its Xs, but that the relation between the two should not be that of spooky, materialistically inexplicable *emergence*. At the same time, many would not want to *identify* S's $\psi$-ing with $X_{1,\ldots,n}$'s $\phi_{1,\ldots n}$-ing, and so there is a market for an intermediate type of interlevel relation. Below I will mostly consider the consequences of requiring that S's $\psi$-ing *metaphysically supervenes* on $X_{1,\ldots,n}$'s $\phi_{1,\ldots n}$-ing. I will also briefly discuss an approach in which the relation between the two is understood in terms of *realization*. These are two common ways of providing non-spooky metaphysics for mechanistic models. My argument is that both spell trouble for the possibility of ideal interventions.

Supervenience may appear like a promising articulation of an ontologically non-spooky interlevel relation. In its broad-brush formulation, supervenience is the claim

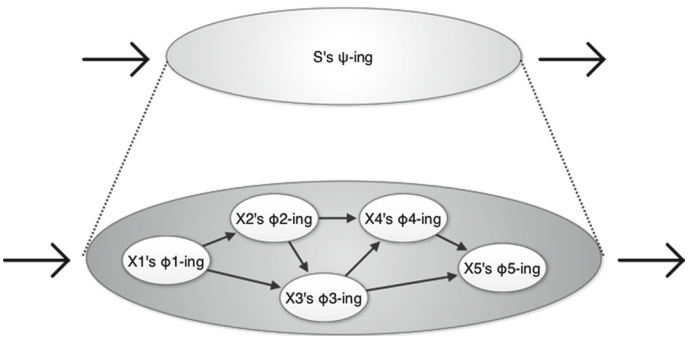


**Fig. 1** A typical representation of a mechanistic model, where S's $\psi$-ing is at the higher level and $X_{1,\ldots,n}$'s $\phi_{1,\ldots n}$-ing is at the lower level. The relations within levels are causal; the relation between S's $\psi$-ing and $X_{1,\ldots,n}$'s $\phi_{1,\ldots n}$-ing is non-causal

that, if a set of properties A supervenes on a set of properties B, then there cannot be any difference in the properties in A without some difference in the properties in B. Applied to mechanisms, the claim is that all changes in S's $\psi$-ing (the supervenient set) must be accompanied with some changes in $X_{1,\ldots,n}$'s $\phi_{1,\ldots,n}$-ing (the subvenient set). The subvenient set is formed by what I've called components-taken-together because changes in a mechanism as a whole generally do not metaphysically necessitate changes in a particular component of that mechanism. There are various ways to make the supervenience claim more specific, depending, among other things, on one's preferences regarding the relation's modal force and the conditions imposed on the relevant property sets (McLaughlin and Bennett 2014).[3] But rather than dwelling on the interesting and energy-consuming differences between the various definitions of supervenience, it suffices for my purposes to focus on the very basic idea.[4]

As we recall, X's $\phi$-ing stands in the relation of mutual manipulability with S's $\psi$-ing just in the case there is an ideal intervention on X's $\phi$-ing (with respect to S's $\psi$-ing) that results in a change in S's $\psi$-ing; *and* there is an ideal intervention on S's $\psi$-ing (with respect to X's $\phi$-ing) that results in a change in X's $\phi$-ing. The problem is that the presence of supervenience in mechanistic models threatens to render the latter kinds of intervention non-ideal. If S's $\psi$-ing supervenes on $X_{1,\ldots,n}$'s $\phi_{1,\ldots n}$-ing, a change in S's $\psi$-ing is necessarily accompanied by a change in $X_{1,\ldots,n}$'s $\phi_{1,\ldots n}$-ing, and the resulting change in $X_{1,\ldots,n}$'s $\phi_{1,\ldots n}$-ing is plausibly one that conflicts with conditions (I1) or (I2) on ideal interventions. The reason is that the target variable is likely to be among those variables that change in the intervention directly due to the presence of supervenience, or is likely to be causally related with some variables that do. Condition (I1) is violated if the target variable is in the supervenience base for S's $\psi$-ing; condition (I2) is violated if the target variable is causally related with the variables in the supervenience base for S's $\psi$-ing. To see why one of these is likely to be the case, suppose that interventions on S's $\psi$-ing regularly change some X's $\phi$-ing, but that the $\phi$-ing in question does *not* change directly in the intervention as the result of being in the supervenience base for $\psi$-ing, and is *not* causally related with any such directly changing variables. In that case, there would appear to be a primitive covariance relation between S's $\psi$-ing and that X's $\phi$-ing; I contend that the model would be regarded as incomplete.[5]

[3] In the mechanistic context, Soom (2012) has argued that the relation between S's $\psi$-ing and $X_{1,\ldots,n}$'s $\phi_{1,\ldots n}$-ing should be understood in terms of 'strong supervenience', while Harbecke (2013) has put forward a mechanistic modification of what is known as 'coordinated multiple-domain supervenience'. Both types of supervenience were originally articulated by Kim (cf. 1984; 1988).

[4] Note that the overall behaviour of a mechanism is usually characterized in extrinsic terms, as a kind of input-output regularity. The advocates of mechanistic supervenience, then, argue for a very similar view as those who hold that dispositions supervene on their 'bases'. This is worth mentioning because the supervenience thesis about dispositions has been challenged (cf. Mumford 1994, McKitrick 2003).

[5] Here I am following Woodward (2014, p. 29–31) in making the ontologically and empirically plausible assumption that an intervention on a supervening variable $\psi$ is a *direct and simultaneous* intervention on its supervenience base SB($\psi$): there is just one intervention that is changing both. For those who do not wish to grant this assumption, it is worth noting that the relationship between $\psi$ and SB($\psi$) also violates the interventionist requirement that one should be able to vary the value of each variable in one's model via an intervention while holding the values of any of the other variables in the model fixed at any of the values within their normal range via independent interventions (Hausman and Woodward 1999; Woodward 2014,

Before moving on, let me briefly address one issue. Some philosophers have argued for realization-based accounts of mechanistic interlevel relations (Polger 2010; Gillett 2013; see also Craver 2007, p. 212). The idea in such accounts is that $X_{1,\ldots,n}$'s $\phi_{1,\ldots n}$-ing together realize S's $\psi$-ing, where the precise definition of the realization relation varies from author to author. Consider Gillett's (cf. 2013) 'dimensioned' account of realization, which is expressly tailored for the purposes of constitutive explanation. The standard example of that type of realization is the relationship between a diamond's hardness and the bonding and alignment relations of its constituent carbon atoms. The bonding and alignment relations of those atoms 'together non-causally result in' the hardness of the diamond, just as the $\phi$-ings of Xs might be thought to 'together non-causally result in' the $\psi$-ing of S. To see why this doesn't help with the problem raised above, just consider what happens if I want to 'intervene' on the hardness of a diamond whilst holding fixed the bonding and alignment relations of the carbon atoms with which it overlaps: this doesn't work. More generally, it is not implausible to require that changing a realized property $\psi$ in an individual always changes some properties $\phi_i$ of the $\psi$-realizing constituents of that individual. This is, of course, very similar to the basic idea driving formulations of supervenience.[6] It also results in exactly the same kind of trouble for the possibility of ideal top-down interventions.

## 5 Interlevel interventionism

The ideality of top-down interventions seems to be compromised under two familiar accounts of the metaphysics of mechanistic models. Of course there may be many more concepts in addition to supervenience and realization that we could use to make sense of the relation between a mechanism and its components-taken-together, and it is not my aim here to provide an exhaustive analysis of such possibilities. Nevertheless, I will make the following conjecture: given that we would want the mechanistic interlevel relation to be one in which the overall behaviour of a mechanism and the organized behaviours of its components are metaphysically close-knit in the sense of ruling out spooky emergence and the like, the prospects are dim for ideal top-down interventions. The reason is that the relevant higher-level changes are associated with lower-level

Footnote 5 continued
see also footnotes 9 and 10). If needed, my conclusion about the lack of ideal interventions on $\psi$ w.r.t SB($\psi$) could be reached trough that route too. In what follows, I will continue to assume that an intervention on $\psi$ is *also* an intervention on SB($\psi$). I thank anonymous reviewer for drawing my attention to this issue.

[6] Indeed, as David Papineau has pointed out (personal communication), it is even possible to understand realization as the *converse* of supervenience: $\phi$ realizes $\psi$ if and only if $\psi$ supervenes on $\phi$. Under this analysis, it is hardly surprising if both relations turn out to be equally problematic for ideal interlevel interventions. But even under the more complex definitions of realization put forward in recent debates, which often make no overt reference to supervenience, it is still plausibly the case that changes in $\psi$ require changes in $\phi$. Note that I am here intentionally working with a very abstract characterization of the realization relation. The reason is that there is currently considerable dispute as to the 'appropriate' type of realization relation in the mechanistic context (Polger 2010). In this dispute, Craver himself seems to side with those advocating the 'dimensioned' concept against the 'flat' one (Craver 2007, p. 212). But it is also questionable whether there is a genuine disagreement between these two notions to begin with (Endicott 2011).

changes in a way that conflicts with the requirements of ideality. This association is due to the metaphysical close-knittedness of the mechanism and its components.[7]

What should we then think about mutual manipulability? For as we recall, establishing the mutual manipulability relation between S's $\psi$-ing and some particular X's $\phi$-ing very much requires top-down interventions. One possibility would be to bite the bullet and say that there is just less mutual manipulability, and consequently constitutive relevance, than what was initially expected. But this is very implausible indeed, given that almost nothing would then count as constitutively relevant. We could also give up on Craver's criterion, in which case it would remain puzzling as to why something like the mutual manipulability criterion seems to be implicit in the practice of science. Moreover, simple example cases such as the braking mechanism in a bicycle suggest that the appropriate combination of top-down and bottom-up interventions delivers the *prima facie* correct results. Finally, a third option is to leave the mutual manipulability criterion as it is and instead develop the notion of *ideal interventions*. As it happens, interventionists working on the causal exclusion problem have recently done just that.[8]

To remind ourselves of the causal exclusion problem, let us look at the classic diagram in Fig. 2, which represents a system in which some relations are causal and some relations are non-causal. The usual interpretation is that $P_1$ and $P_2$ are some physical properties and that the arrow between them represents a causal relation. $M_1$ and $M_2$ are assumed to be mental properties and the dotted lines between the Ps and the Ms are thought to represent supervenience or realization relations, or perhaps some other non-causal relations. The standard problem is, of course, whether it is legitimate to draw an arrow from the mental $M_1$ to the physical $P_2$. Since the relationship between the mental and the physical is not of special concern here, we need not assume anything substantive about the nature of the Ps and the Ms. Thus, I am again going to talk as if all the relations in models like the one in Fig. 2, including the non-causal ones,

[7] The argument in the previous section also shows why the worries raised by Leuridan (2012, p. 407–409) about interlevel interventions satisfying conditions (I1)–(I4) are premature. There are no such interlevel interventions because conditions (I1) or (I2) will be violated as the result of the close-knitted metaphysical relationship between the mechanism as a whole and its components-taken-together, such as supervenience or realization. Leuridan's argument is that if we assume that Craver's interlevel interventions satisfy conditions (I1)–(I4), then it very much looks as if those interventions pick out causal relevance relations. My answer here is that we should not make such an unrealistic assumption in the first place. In any case, Leuridan then goes on to ask whether Craver could use a parthood criterion to argue that mechanistic interlevel interventions are not causal. In order to do that, he thinks, Craver would need to be able to assume that (i) 'if X is part of S, then an intervention I on X directly changes S' whilst denying that (ii) 'if X is part of S, then an intervention I on X's $\phi$-ing directly changes S's $\psi$-ing'. The problem is, according to Leuridan, that it is hard to make the former assumption without making the latter. As it happens, in Sect. 6.1 I will demonstrate why what Leuridan claims is difficult for Craver is not difficult at all: the key move is to understand mutual manipulability as involving *three* variables. For further discussion of Leuridan's argument, see Footnotes 16 and 18.

[8] Note that the problem does not depend on whether the mutual manipulability criterion is interpreted 'metaphysically' or 'epistemologically' (cf. Schindler 2013). Either way, the interlevel interventions turn out non-ideal. Moreover, the following question would remain even under the weaker epistemological reading: what *is* the metaphysical structure that grounds top-down and bottom-up experiments?

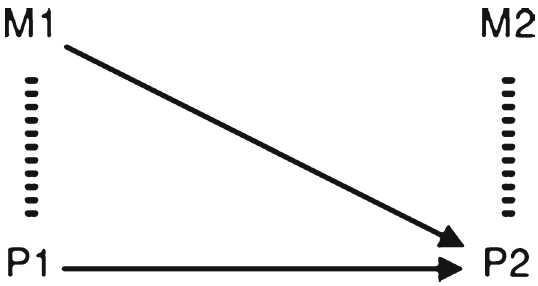


**Fig. 2** Classic diagram representing the problem of high-level causation

hold between variables. Additionally, to keep things simple, I will assume that the non-causal relation in question is supervenience.[9]

Now, the problem here, too, is that a change in $M_1$ will be accompanied by a change in $P_1$, given the nature of supervenience. By now we know very well why this suggests that there are no ideal interventions on $M_1$ with respect to $P_2$: the intervention on $M_1$ automatically changes variable $P_1$, which again is causally related with variable $P_2$. This appears to conflict with condition (I2) on ideal interventions. One conclusion would be that we have just formulated an interventionist version of Kim's (cf. 2000) 'causal exclusion argument'. Baumgartner, for example, has promoted this view (cf. 2010). The reasoning is that, since there are no ideal interventions on $M_1$ with respect to $P_2$, there cannot possibly be a causal relation between $M_1$ and $P_2$, given that the presence of causality in the interventionist framework requires the possibility of such an intervention. This follows from the way in which interventionists typically define their causal notions. Woodward, for example, gives the following necessary and sufficient condition for variable X's being a type-level direct cause of variable Y (relativizing causal claims to some set of variables V):

> There [must] 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 (Woodward 2005, p. 59).

His necessary and sufficient condition for X's being a type-level contributory cause of Y with respect to V is that:

> There be a directed path from X to Y such that each link in this path is a direct causal relationship [...and] 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 2005, p. 59).

The notion of intervention in these definitions is assumed to be precisely the kind of ideal intervention satisfying conditions (I1)–(I4). The non-existence of such an

[9] One further assumption here and elsewhere in this paper is that a supervenient variable cannot be causally related with its supervenience base. This is trivially the case if the relationship between the two violates condition (I1) for ideal interventions, as I think it does. But again the same conclusion can be achieved by observing that the two variables violate the independent fixability condition (see footnotes 5 and 10).

intervention on $M_1$ with respect to $P_2$ is then taken, by Baumgartner, to suggest that $M_1$ cannot be causally related with $P_2$ under interventionism.[10]

Many philosophers (Sober and Shapiro 2007; Woodward 2014; Shapiro 2012) disagree with the above argument. An obvious objection is that the notion of an ideal intervention of the type satisfying conditions (I1)–(I4) was developed in the context of causal models. The conditions seek to capture a view of interventions that can be reasonably thought to underlie causal inference in such models, and this is one of the reasons why we might want to understand causal notions in terms of ideal interventions in the first place. Note that the motivation for criteria (I1)–(I4) did *not* arise from how causal inference works in cases such as the one in Fig. 2, in which we are dealing with causal as well as non-causal relations. So there is no immediate reason to assume that criteria (I1)–(I4) must apply, unchanged, in the context of models that contain non-causal relations. Another problem with Baumgartner's argument is that the counterfactual that we are supposed to evaluate when assessing the causal status of $M_1$ with respect to $P_2$ has a metaphysically impossible antecedent: 'If an ideal intervention (on $M_1$ w.r.t. $P_2$) were to change the value of $M_1$, the value of $P_2$ would change.' The antecedent is metaphysically impossible because the supervenience relations in the model ensure that there are no interventions on $M_1$ with respect to $P_2$ that would satisfy conditions (I1)–(I4). This is true, *mutatis mutandis*, of all of the relations between variables to which the interventionist variant of the exclusion argument is supposed to apply. But under standard semantics (cf. Lewis 1973), this shows *at best* that the counterfactual in question is vacuous (or vacuously *true*)—not that $M_1$ cannot possibly cause $P_2$.[11]

The alternative view that interventionists of the *non-exclusionist* persuasion have put forward (cf. Sober and Shapiro 2007; Woodward 2014) is that if our models contain non-causal relations in addition to causal ones, we can employ the interventionist techniques of assessing causality by letting variables related via logical or 'metaphysical' relations vary in a way that 'respects' those relations. What I mean by a metaphysical relation in this section includes supervenience, realization and other relevantly similar relations. So, for example, when we intervene on $M_1$, we let $P_1$ change at the same time, and exclude $P_1$ from those variables that our intervention on $M_1$ with respect to $P_2$ must leave intact. Under this extension of the interventionist approach, we do have an ideal intervention on $M_1$ with respect to $P_2$. Whether $M_1$ is causally related with $P_2$ then depends on what would happen to the value $P_2$ if the value $M_1$ would

[10] The debate gets more complex than this. For instance, it is usually framed in terms of the independent fixability condition that many (cf. Baumgartner 2010; Halpern and Hitchcock 2011; Woodward 2014) take to be an important interventionist assumption and that is also discussed in footnotes 5 and 9. A roughly similar assumption can be found in the work of Craver (2007, p. 156–157), although it is not clear whether Craver's version of the condition is equivalent to those put forward by others (cf. Woodward 2014, p. 14). (I thank anonymous referee for pointing out the apparent lack of equivalence.) For the purposes of this paper, however, we can happily continue to frame the debate in terms of the conditions for ideal interventions.

[11] This is just how Lewis treats counterfactuals with impossible antecedents. The motivation that Lewis gives for this involve the intuition, which some might have, that if something impossible were to be true, then anything could be true (Lewis 1973, p. 24). Another thing worth noting is that Woodward himself requires that, in order for X to cause Y, there must *be* an ideal intervention on X with respect to Y (Woodward and Hitchcock 2003; Woodward 2005). But it is not clear whether this commitment is essential for the interventionist programme.

be changed by means of that ideal (under the extended approach) intervention. Suppose that the relationship between M1 and P2 does indeed satisfy, say, Woodward's definition of M1's being causally related with P2. Then that relation is what some people will call 'downward causation'. Note the compatibility of this type of downward causation with there being supervenience relations in the model (represented by the dotted arrows in Fig. 2). Even though M1 supervenes on P1, that does not in any way prevent M1's being causally related with P2 according to extended interventionism. The above approach can be stated in the form of the following two principles:

> (EI1) If variable X and variable SB(X) are related via supervenience or some other metaphysical or logical relation, an intervention on X may at the same time be an intervention on SB(X); and
> (EI2) when assessing whether X is a direct or contributory cause of some variable Y, SB(X) should *not* be regarded as belonging to the set of variables that must be left intact or controlled for in the intervention on X.

(EI1) and (EI2) are meant to describe how we can go on conducting ideal interventions in models that contain non-causal relations in a way that 'respects' those relations. Note that when I say, in condition (EI1), that the intervention is an intervention on X and SB(X) at the same time, this should be read literally: it is the same one intervention that changes both variables simultaneously, as requested by the supervenience relation and other relevantly similar relations (see Woodward 2014, p. 25-30 for further rationale and discussion). It is easy to see how simple the approach is: we ignore the variables that are related via metaphysical or logical relations with the variable on which we intervene, and allow them to vary in whichever way is necessary as a result of the nature of the relevant relation.[12]

What is the motivation for this extension of interventionism? Here both Woodward (2014) and Sober and Shapiro (2007) appeal to scientific practice. According to them, when researchers evaluate the causal contribution of some variable X, they do not regard the supervenience base SB(X) of that variable as something that needs to be left intact or 'controlled for' in the intervention on X. Controlling for, in this context, can be understood as holding the values of some subset of variables in one's model 'fixed' while the value of X is being varied (see Woodward's conditions for direct and contributory causes above). In assessing whether X causes Y, scientists are clearly worried about some third variable Z genuinely independent of X that is also causally related with Y. These types of variables are immediately regarded as potential confounders. But scientists do *not* seem similarly worried about the supervenience base SB(X) of X, even if SB(X) is causally related with Y, just as $P_1$ is causally related with $P_2$ in Fig. 2.

[12] One might worry that the proposed extension is *ad hoc*, and Woodward himself (2014, p. 30) concedes that his conditions for extended interventions are 'cumbersome'. But I believe the only reason (EI1) and (EI2) sound *ad hoc* is that they weren't included in the original set of complex conditions used to formulate interventionism in the first place. The crucial test here is whether they capture the scientific practice, which I've argued they do. I thank an anonymous reviewer for drawing my attention to this worry.

To drive the point home, Woodward (2014) devises a thought experiment about a medical researcher who argues that ingesting a certain drug D cannot cause recovery R because it is not possible to intervene on D without at the same time changing the physical state of affairs SB(D) on which D supervenes. All interventions on whether a subject ingests D are bound to change whether the subject has a substance with the microproperties of D present in her system. And that state of affairs SB(D) is for the researcher a confounding variable with respect to R. The punchline of this story is that the scientific community would regard these claims as silly and the researcher a crank. The reason apparently is that scientists do not regard subvenient variables as causal competitors of those variables that supervene on them in anything like the manner that they regard genuinely distinct variables as potential causal competitors.[13] This may be because variables bearing logical or metaphysical relations with each other are tied together more closely than variables bearing causal relations only. A typical requirement imposed on causal relata is that they should be wholly distinct, and it is clear that this is not the case with variables related via the kinds of non-causal relation that I have discussed in this section.

In the remaining part of this paper, I want to see how we can go about understanding mechanistic models *if* we adopt the 'extended interventionism' of Woodward (2014) and Sober and Shapiro (2007). More specifically, I want to see what happens if we accept principles (EI1) and (EI2). In the next section, my argument will be that doing so leads us to a radical reinterpretation of the metaphysics of mutual manipulability.

## 6 Towards causal inbetweenness

So far in this paper I have sought to establish three points. First, I showed that the notion of mutual manipulability is based on the assumption that we are able to perform ideal interventions in models that contain causal as well as non-causal relations. Second, I pointed out that a typical way of understanding the non-causal dimension in mechanistic models is to treat it as a species of supervenience or realization. I argued that the presence of these types of relation foils the ideality of interventions on the behaviour of a mechanism as a whole with respect to the behaviours of its components under standard conditions. But then, third, I observed that the standard conditions for ideal interventions should in fact be extended if our models contain non-causal relations. In this last section of the paper my aim will be very simple: I will adopt the extended notion of ideal interventions and apply it in the context of mutual manipulability. The resulting picture I will call the idea of 'causal inbetweenness'.[14]

[13] Baumgartner (2013) argues that scientists might want to regard D and SB(D) as competitors if they suspect that D does not *reduce* to SB(D). If there would be some sort of primitive and inexplicable supervenience relation between the two, then we might actually see scientists requiring interventions on D with respect to R that would at the same time leave SB(D) unchanged. Perhaps this is so. But the reason is the suspect nature of the *supervenience* relation between D and SB(D) and not because this is the correct way to do causal modeling.

[14] I thank Carl Craver for suggesting this name for my approach, as well as for the many valuable comments he provided regarding the ideas put forward in this section.

### 6.1 Mutual manipulability as a three variable affair

My argument begins from a further explication of S's $\psi$-ing. Thus far, I haven't discussed in any detail what it means specifically to intervene on $\psi$. As I hinted earlier, the phenomenon exhibited by a mechanism as whole is usually thought to be some regularity, identified in terms of the conditions under which it occurs, doesn't occur, or is altered. According to Craver, that regularity can be thought of as holding between some complex sets of inputs and outputs (Craver 2007, p. 145), and it is commonplace to think of it as causal (cf. Menzies 2012; Soom 2012; Glennan 1996). Adopting the interventionist convention, such causal regularities are to be understood as holding between the values of variables. Ignoring the complexity of real-life cases, it is possible to represent the phenomenon exhibited by a mechanism as a causal relation holding between two variables, $\psi_{in}$ and $\psi_{out}$, corresponding to the input and output conditions characteristic of the relevant regularity. These variables could have as many values as we like, but for my purposes it suffices to imagine them as having just two values corresponding to the presence or absence of the appropriate conditions; let these values be '1' for the presence and '0' for the absence.

Adopting the above simplified interpretation of the behaviour exhibited by a mechanism as a whole, how should we understand interventions on such behaviour? Here is the key: 'One intervenes on S's $\psi$-ing by intervening to provide the conditions under which S regularly $\psi$s' (Craver 2007, p. 146). According to the the interpretation sketched here, this amounts to changing the value of $\psi_{in}$ to 1, as a result of which $\psi_{out}$ should take the value 1. In other words, we intervene in order to make the input conditions occur, as a result of which the output conditions typically occur. Of course, we might also intervene to make the input conditions disappear, by setting $\psi_{in} = 0$, in which case we would expect $\psi_{out}$ to take value 0, too. And so on. If our variables would have multiple values—as they would in real-life applications—we could establish all sorts of pattern between $\psi_{in}$ and $\psi_{out}$. We could investigate the behaviour of the mechanism under 'modulating' conditions, by making minute changes in the values of the input variables and observing the resulting changes (if any) in the output variables.

But note the implications of this very simple move. If S's $\psi$-ing is understood as a causal regularity holding between the values on $\psi_{in}$ and $\psi_{out}$, then it appears that assessments of mutual manipulability are best understood as involving *three* variables. We have the higher-level input and output variables $\psi_{in}$ and $\psi_{out}$, which are together individuative of S's $\psi$-ing, and then we have some lower-level variable $\phi_i$ which engages in a manipulability relation with both of the higher-level variables. Since interventions on the behaviour of a mechanism as a whole are here understood as targeting the value of the *input* variable $\psi_{in}$, there now surfaces a natural way to interpret the top-down experiments relevant for mutual manipulability: one wiggles the value of the *input* variable $\psi_{in}$ and observes whether there occur any changes in the value of the lower-level variable $\phi_i$. What about the bottom-up experiments also required for mutual manipulability? I contend that in such experiments one typically wiggles the value of the lower-level variable $\phi_i$ and observes whether there occur any changes in

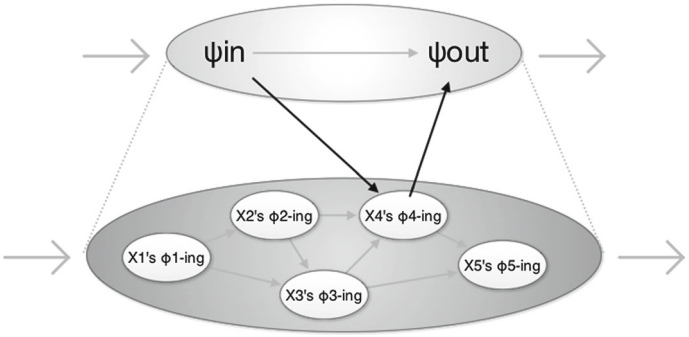


**Fig. 3** Top-down and bottom-up experiments further unpacked. Here, a hypothetical investigator wants to establish whether $X_4$'s $\phi_4$-ing (or '$\phi_4$' for short) stands in the mutual manipulability relation with S's $\psi$-ing. The top-down intervention required for mutual manipulability changes the value of $\psi_{in}$ and detects changes in the value of $\phi_4$; the bottom-up intervention similarly required changes the value of $\phi_4$ and detects changes in the value of $\psi_{out}$. The intralevel arrows, which are greyed out for expository purposes, represent causal relations. Assuming that causes temporally precede their effects, it follows that the arrow of time in the model points from left to right

the *output* variable $\psi_{out}$.[15] Note, finally, that the top-down and bottom-up experiments as understood here needn't be two-variable affairs. For example, one can top-down intervene on $\psi_{in}$ with respect to some $\phi_i$ while also detecting changes in $\psi_{out}$. Or, alternatively, one can bottom-up intervene on some $\phi_i$ with respect to $\psi_{out}$ while at the same time manipulating the value of $\psi_{in}$: for example, sometimes a change in $\phi_i$ will result in a change $\psi_{out}$ only if the input condition is also present. Craver (2007, personal communication) suggests that this is what often happens in neuroscientific experiments. Figure 3 captures this interpretation of mutual manipulability in terms of the hypothetical mechanistic model discussed in Sect. 4 (and depicted in Fig. 1).[16]

Recall the bicycle braking mechanism discussed earlier. Let the input variable $\psi_{in}$ correspond to whether force is applied on the lever and the output variable $\psi_{out}$ to whether the pads in the brake caliper press against the rim surface. Suppose the lower level variable $\phi_i$ of interest corresponds to whether the cable connecting the lever to the caliper moves relative to its housing. According to the interpretation proposed here, the mechanistic investigation works as follows. In a top-down intervention, one

[15] I am not excluding the possibility that a researcher might sometimes wiggle the value of some $\phi_i$ and detect changes in the value of the *input* variable $\psi_{in}$. This type of case might occur, for example, if the $\phi_i$ in question is among the variables on which $\psi_{in}$ supervenes. However, even in this case it is plausible that the researcher would *also* require changes in $\psi_{out}$ under the intervention on $\phi_i$.

[16] This three-variable nature of mutual manipulability is missed by Leuridan (2012), whose argument was discussed earlier in Footnote 7. As we recall, his claim was that, in order to argue that mechanistic interlevel relations are not causal, Craver would have to maintain that (i) an intervention on part X of S directly changes S while simultaneously denying that (ii) an intervention on X's $\phi$-ing directly changes S's $\psi$-ing—something Leuridan thought Craver would have a hard time doing. However, the unpacking of S's $\psi$-ing as involving *two* variables and the consequent interpretation of mutual manipulability as a three-variable affair show why one could easily hold (i) while denying (ii). The issue is a red herring. Why mechanistic interlevel relations come out as causal isn't to do with the difficulty of holding the above combination of beliefs; it's to do with what is a plausible account of interlevel interventions. My argument is that it is the extended one. See also Footnote 18.

applies force on the lever ($\psi_{in}$) and observes changes in the position of the cable ($\phi_i$). In a bottom-up intervention, one changes the position of the cable ($\phi_i$) and observes changes in the brake caliper ($\psi_{out}$). If necessary, the investigator can switch to three-variable analysis by varying the position of the cable and observing changes in the caliper while at the same time applying force on the lever.

With this interpretation of top-down and bottom-up interventions at hand, we can now see how the extended notion of ideal interventions works with mechanisms. In order to do so, let us go back to the model depicted in Figs. 1 and 3. Let us suppose that S's $\psi$-ing supervenes on $X_{1,\ldots,n}$'s $\phi_{1,\ldots n}$-ing and we want to investigate whether $X_4$'s $\phi_4$-ing (variable '$\phi_4$' for short) is constitutively relevant for S's $\psi$-ing. This requires us to perform an ideal intervention on $\psi_{in}$ (w.r.t. $\phi_4$) and see whether there occurs a change in $\phi_4$. If $\psi_{in}$ and $\phi_4$ covary, it would under the 'narrow' interpretation of ideal interventions suggest that the intervention on $\psi_{in}$ is not ideal with respect to $\phi_4$. The reason is that the likely explanation for the covariance would be: either $\phi_4$ is among the variables in the supervenience base of $\psi_{in}$ and change directly in the intervention; or then the variables in the supervenience base of $\psi_{in}$ that change directly in the intervention include some variable(s) causally related with $\phi_4$.

With extended interventionism, the situation is completely different. True, if the intervention on $\psi_{in}$ directly changes $\phi_4$ because $\phi_4$ is in the supervenience base of $\psi_{in}$, then it isn't ideal with respect to $\phi_4$. This violates condition (I1) just as before, and the extended approach doesn't make any amendments with regard to this type of case. But if the intervention on $\psi_{in}$ directly changes some variables in the supervenience base of $\psi_{in}$ that are causally related with $\phi_4$, then that does *not* rule out the ideality of that intervention under the extended approach. For according to conditions (EI1) and (EI2), we let such variables vary in a way that respects the supervenience relation in question and do *not* demand that our intervention on $\psi_{in}$ must leave them intact. So suppose that, say, $X_1$'s $\phi_1$-ing (variable '$\phi_1$' for short) in our model is among the variables in the supervenience base of $\psi_{in}$ and as a result changes directly in an intervention on $\psi_{in}$. Then, even though $\phi_1$ is causally related with $\phi_4$ in the model, the intervention on $\psi_{in}$ *can* be ideal with respect to $\phi_4$. Whether it *is* ideal depends on whether the other conditions of ideality are satisfied.

But notice this: *if* the intervention on $\psi_{in}$ satisfies all of the other conditions of ideality with respect to $\phi_4$, and *if* the change introduced in the value of $\psi_{in}$ via such an intervention results in a change in the value of $\phi_4$, then the relationship between $\psi_{in}$ and $\phi_4$ counts as *causal*. This is an instance of what some call downward causality, and it is a straightforward consequence of the fact that the relevant intervention now satisfies the conditions of ideality. Similarly, although less controversially, if an intervention on the value of $\phi_4$ changes the value of $\psi_{out}$ and satisfies the conditions of ideality, then that relation too counts as causal. This, then, is an instance of upward causality. Thus, adopting extended interventionism renders ideal interventions in mechanistic models viable, but in doing so it treats an important subset of the interlevel manipulability relations as causal.

Craver holds that the manipulability relation between the behaviour of a mechanism as a whole and the behaviours of its components cannot count as causal because it is synchronic, symmetric and involves variables that are not wholly distinct (Craver and Bechtel 2007; Craver 2007). But if this is the case, then we are going to end up with

more trouble for ideal interventions. The problem is that we still want the behaviour of the mechanism to supervene on the behaviours of its components. But if one treats the behaviour of a mechanism as a *single* variable supervening on everything that goes on at the level of its components, then it appears as if *all* interventions on the mechanism as a whole change the behaviours of its components directly (if at all) and *vice versa*. This, of course, is against the conditions of ideality.[17] The picture that I have argued for doesn't have this problem because the overall behaviour is unpacked as involving *two* variables. This is what makes it possible for me to have asynchronic, asymmetric manipulability relations between distinct variables. In so far as ideal interventions are to play a role in assessing constitutive relevance, I believe it is essential to see mutual manipulability as a three variable affair involving interventionist interlevel causation. In the following final section of this paper, I will briefly sketch how my account works with a neuroscientific example.[18]

### 6.2 Causal inbetweenness at work: explaining neuronal communication

To illustrate the account sketched above, I want to focus on the phenomenon of neuronal communication, which provides a paradigmatic example of mechanistic explanation. Suppose we want to explain a neuron's ability to transmit the signals that it receives from other neurons. The input here is the reception of a signal at the dendrites and the output is the release of a signal at the axon terminals. A much simplified lower-level story could go something like this. The signals come in the form of neurotransmitters such as glutamate that bind in receptors typically in the dendrites of the neuron. This binding causes the opening of various voltage-dependent gates that are embedded in the neuron's membrane. When the neuron is in its resting state it functions as a battery: there is a higher potassium concentration inside the cell and a higher sodium concentration outside. Specific transmembrane sodium-potassium pumps work for this purpose. The differing sodium and potassium concentrations inside and outside of the cell keep the neuron slightly negatively charged. When the neurotransmitters bind in the receptors in the dendrites, the voltage-dependent gates open in a way that enables sodium influx and potassium efflux. This raises the neuron's potential.

If the neuron receives enough signals so that their combined effect raises its potential above a certain threshold level, this generates a positive feedback loop where more and more voltage-dependent gates open at an increasing rate, typically in the neuron's axon hillock area. This results in a rapid increase fol-

[17] Or violates other important interventionist assumptions. See footnotes 5, 9 and 10.

[18] An anonymous reviewer asks what the relationship between constitution and causality is in the account that I'm giving. Elsewhere (unpublished) I have developed an approach to constitution under which parts must be causally in between the inputs and outputs defining the phenomenon for which the whole of which they are parts is responsible. My view is that this is a natural step to take if one is already willing to decide issues of constitutive relevance in terms of mutual manipulability. An obvious implication of this is that I do not accept the view that constitutive relations cannot be causal. On the contrary, I have outlined a perfectly good way in which there is a causal relationship between a mechanism's behaviour as a whole and the individual behaviours of its components, even though the latter are constituents in the former. As the consequence of this, putative counterexamples to views that Craver may hold, such as the case of endosymbiosis discussed by Leuridan (2012, p. 412), are not counterexamples at all from my point of view.

lowed by a swift decrease in the neuron's potential, i.e. generates an action potential. The shape of the potential is the result of the differing opening and closing times of potassium and sodium channels. The action potential then has the ability to propagate through the neuron's axon owing to the axon's physical structure, which may include a surrounding wrap of myelin that has tiny gaps where the action potential reoccurs. When the action potential finally arrives at the axon terminals, it causes calcium channels in the membrane to open, which triggers a chain of chemical events in which vesicles containing neurotransmitters fuse in the membrane, releasing the neurotransmitters in the process. This completes the transmission of the signal in this caricaturized example of a mechanistic explanation.

Suppose we are interested in whether, say, the opening and closing of a voltage-dependent gate at some region in the neuron's axon is a part of the mechanism for signal transmission. That is, we want to find out whether the behaviour of a lower-level component is constitutively relevant with respect to a phenomenon exhibited by the mechanism at a higher level. In order to conduct the required top-down intervention, we need to vary between the condition in which the input for the neuron (the reception of a signal) is present and the one in which it isn't. The presence or absence of this input corresponds to the value of $\psi_{in}$. We must then see whether the status of the voltage-dependent gate covaries with those conditions. The status of the gate corresponds to the value of $\phi_i$. As the above story suggests, the presence or absence of the input condition supervenes on the presence or absence of neurotransmitters in a thousand receptors in the dendrites of the neuron. That is why, when we wiggle the value of the variable corresponding to the presence or absence of the input condition ($\psi_{in}$), we at the same time wiggle the values of the numerous variables corresponding to the statuses of these receptors. From these lower-level variables starts an *ordinary intralevel causal chain* that may or may not lead to the voltage-sensitive gate under investigation. If the value of the gate variable ($\phi_i$) covaries with the presence or absence of the signal, it is indeed likely that there is such a causal chain from the receptors to the voltage-dependent gate. This in itself doesn't preclude the ideality of the intervention on the neuron as a whole with respect to that voltage-dependent gate. If the intervention otherwise satisfies the conditions of ideality (as it should), the interlevel relation between the input variable ($\psi_{in}$) and the gate variable ($\phi_i$) counts as causal under extended interventionism.

The relevant bottom-up intervention proceeds from the voltage-sensitive gate to the output condition, which is the release of a signal by the neuron. The presence or absence of this condition corresponds to the value of $\psi_{out}$. Again we know from the above story that the value of this higher-level output variable supervenes on the values of lower-level variables that correspond, among other things, to the statuses of the vesicles containing the neurotransmitters in the axon terminal membranes. If there is a covariance between whether the voltage-dependent gate is open (the value of $\phi_i$) and whether the output condition is present (the value of $\psi_{out}$), then it is likely that there is an *ordinary intralevel causal chain* from the voltage-sensitive gate to the variables corresponding to the statuses of the vesicles containing neurotransmitters. Note that the covariance between the gate variable and the output variable need only be 'visible'

when the values of a number of other variables, possibly including the input variable, are held fixed.[19] Again, if the intervention on the value of the gate variable changes the value of the output variable and otherwise satisfies the extended conditions of ideality (as it should), then this interlevel relation too counts as causal. Imagine Fig. 3 depicts the mechanistic model for neural communication. Then variable $\psi_{in}$ in the model corresponds to the arrival of a signal, $\psi_{out}$ to the release of a signal and $\phi_4$ to the status of the voltage-sensitive gate. The gate variable is *causally in between* the input and output variables.

## 7 Conclusion

The problem of constitutive relevance is that of picking out exactly those components that should be included in a mechanistic explanation for a phenomenon and of specifying the nature of the relation that holds between those components and that *explanandum*. Craver's mutual manipulability criterion is a major step forward in this regard because it addresses these issues by building upon the scientific practice and the well worked out interventionist account of causation. In this paper I have pointed out the tensions that remain in the mutual manipulability criterion as the result of the fact that mechanistic models contain causal as well as non-causal relations. Further, I have demonstrated how those tensions can be resolved with the help of extended interventionism when mutual manipulability is understood as a three variable affair. The resulting picture differs from Craver's in that many of the relevant manipulability relations in top-down and bottom-up interventions come out as causal. I regard this as welcome because it suggests that the question concerning the nature of these relations reduces to the question concerning the nature of causation. If this is right, there is no *special* problem about the metaphysics of mutual manipulability.

**Acknowledgements** I thank Carl Craver, Eleanor Knox, David Papineau and two anonymous reviewers for valuable comments on previous versions of the manuscript. I'm also grateful to all participants and organizers of the 'Mind and Life: Mechanistic and Topological Perspectives' conference in Belgrade. The research that went into this paper was supported by Kone Foundation.

## References

Baumgartner, M. (2010). Interventionism and epiphenomenalism. *Canadian Journal of Philosophy*, *40*(3), 359–383.

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

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Science Part C*, *36*(2), 421–441.

[19] In practical terms, investigating the causal contribution of some *single* voltage-dependent gate is clearly unrealistic. A real-life researcher would be likely to intervene on a number of voltage-dependent gates in a given region and see whether *that* results in changes in the output (real-life interventions *are* most of the time ham-fisted). But recall the counterfactual element in interventionism. The appropriate question here is: would the neuron release a signal if an ideal intervention were to change the status of the voltage-dependent gate under investigation, whilst the values of various off-path variables, including those corresponding to the statuses of the other voltage-dependent gates in the relevant region, would be held fixed.

Bechtel, W., & Richardson, R. C. (1993). *Discovering complexity: Decomposition and localization as scientific research strategies*. Princeton: Princeton University Press.
Craver, C. F. (2007). *Explaining the brain*. Oxford: Oxford University Press.
Craver, C. F., & Bechtel, W. (2007). Top-down causation without top-down causes. *Biology and Philosophy*, *22*(4), 547–563.
Cummins, R. C. (2000). "How does it work" versus "what are the laws?": Two conceptions of psychological explanation. In F. Keil & R. A. Wilson (Eds.), *Explanation and cognition* (pp. 117–145). Cambridge: MIT Press.
Endicott, R. P. (2011). Flat versus dimensioned. *Journal of Philosophical Research*, *36*, 191–208.
Gillett, C. (2013). Constitution, and multiple constitution, in the sciences: Using the neuron to construct a starting framework. *Minds and Machines*, *23*(3), 309–337.
Glennan, S. S. (1996). Mechanisms and the nature of causation. *Erkenntnis*, *44*(1), 49–71.
Halpern, J.Y., Hitchcock, C. (2011). Actual causation and the art of modeling. arXiv:1106.2652.
Harbecke, J. (2013). The role of supervenience and constitution in neuroscientific research. *Synthese*, *191*, 1–19.
Harinen, T. (unpublished). Towards an account of scientific constitution. Unpublished manuscript.
Hausman, D. M., & Woodward, J. (1999). Independence, invariance and the causal markov condition. *The British Journal for the Philosophy of Science*, *50*(4), 521–583.
Kim, J. (1984). Concepts of supervenience. *Philosophy and Phenomenological Research*, *45*(2), 153–176.
Kim, J. (1988). Supervenience for multiple domains. *Philosophical Topics*, *16*(1), 129–150.
Kim, J. (2000). *Mind in a physical world: An essay on the mind-body problem and mental causation*. Cambridge: MIT press.
Leuridan, B. (2012). Three problems for the mutual manipulability account of constitutive relevance in mechanisms. *The British Journal for the Philosophy of Science*, *63*(2), 399–427.
Lewis, D. K. (1973). *Counterfactuals*. Oxford: Blackwell Publishers.
Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of science*, *67*, 1–25.
McKitrick, J. (2003). A case for extrinsic dispositions. *Australasian Journal of Philosophy*, *81*(2), 155–174.
McLaughlin, B., & Bennett, K. (2014). Supervenience. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/archives/spr2014/entries/supervenience/.
Menzies, P. (2012). The causal structure of mechanisms. *Studies in History and Philosophy of Science Part C*, *43*(4), 796–805.
Mumford, S. (1994). Dispositions, supervenience and reduction. *The Philosophical Quarterly*, *44*(177), 419–438.
Polger, T. W. (2010). Mechanisms and explanatory realization relations. *Synthese*, *177*(2), 193–212.
Schindler, S. (2013). Mechanistic explanation: asymmetry lost. In V. Karakostas & D. Dieks (Eds.), *EPSA11 perspectives and foundational problems in philosophy of science. The European Philosophy of Science Association Proceedings* (pp. 81–91). Springer.
Shapiro, L. A. (2012). Mental manipulations and the problem of causal exclusion. *Australasian Journal of Philosophy*, *90*(3), 507–524.
Shapiro, L., & Sober, E. (2007). Epiphenomenalism - The do's and the don'ts. In P. Machamer & G. Wolters (Eds.), *Thinking about Causes* (pp. 235–264). Pittsburgh: University of Pittsburgh Press.
Soom, P. (2012). Mechanisms, determination and the metaphysics of neuroscience. *Studies in History and Philosophy of Science Part C*, *43*(3), 655–664.
Woodward, J. (2005). *Making things happen: A theory of causal explanation*. New York: Oxford University Press.
Woodward, J. (2014). Interventionism and causal exclusion. *Philosophy and Phenomenological Research*. doi:10.1111/phpr.12095.
Woodward, J., & Hitchcock, C. (2003). Explanatory generalizations, part i: A counterfactual account. *Nous*, *37*(1), 1–24.