Saving the Mutual Manipulability Account of Constitutive Relevance

Beate Krickel

ABSTRACT

Constitutive mechanistic explanations are said to refer to mechanisms that *constitute* the phenomenon-to-be-explained. The most prominent approach of how to understand this constitution relation is Carl Craver's mutual manipulability approach to constitutive relevance. Recently, the mutual manipulability approach has come under attack (Leuridan 2012; Baumgartner and Gebharter 2015; Romero 2015; Harinen 2014; Casini and Baumgartner 2016). Roughly, it is argued that this approach is inconsistent because it is spelled out in terms of interventionism (which is an approach to causation), whereas constitutive relevance is said to be a non-causal relation. In this paper, I will discuss a strategy of how to resolve this inconsistency, so-called *fat-handedness approaches* (Baumgartner and Gebharter 2015; Casini and Baumgartner 2016; Romero 2015). I will argue that these approaches are problematic. I will present a novel suggestion of how to consistently define constitutive relevance in terms of interventionism. My approach is based on a causal interpretation of mutual manipulability, where manipulability is interpreted as a causal relation between the mechanism's components and *temporal parts* of the phenomenon.

1. Introduction

Defenders of the new mechanistic approach highlight the importance of so-called *constitutive* mechanistic explanations for the life sciences and other special sciences. Constitutive mechanistic explanations are taken to explain a phenomenon by a mechanism that *constitutes* or *underlies* the phenomenon (Bechtel 2008; Bechtel and Abrahamsen 2005; Craver 2007b; Craver and Darden 2013; P. Machamer, Darden, and Craver 2000; Illari and Williamson 2012). A standard view is that this constitution-relation is a *non-causal dependency relation* that involves a *part-whole relation* between the mechanism's components and the phenomenon. Furthermore, according to a popular approach, constitutive relevance[1] is analyzed in terms of *mutual manipulability* between the mechanism's components and the phenomenon (Craver 2007b, 153). Thereby, it is held that mutual manipulability can be spelled out in terms of Woodwardian *ideal interventions* (Woodward 2003).

[1] Mechanistic constitution and constitutive relevance are different but related notions. Mechanistic constitution is a relation between a mechanism and a phenomenon; constitutive relevance is a relation between a *component* of a mechanism and a phenomenon.

At a first glance, this combination of interventionism and the mechanistic approach adequately captures many aspects of the explanatory and experimental practice of the life sciences, especially the practice of interlevel experiments (Craver 2002; Kaplan 2012; Romero 2015).[2] Still, recently different authors have argued that the combination is problematic (Leuridan 2012; Baumgartner and Gebharter 2015; Romero 2015; Harinen 2014; Casini and Baumgartner 2016). Roughly, it is argued that applying interventionism to mechanisms leads into a dilemma: Either constitutive relevance turns out to be a causal relation, since constitutive relationships satisfy the interventionist criteria for causation. Or one upholds the assumption that constitutive relevance is a non-causal dependency relation, then ideal interventions into mechanisms turn out to be impossible.

There are different strategies for how to react to this dilemma.[3] First, one might opt for the first horn. Bert Leuridan (2012) suggests to bite the bullet and accept that constitutive relevance just is a special kind of causal relevance. Totte Harinen (2014) highlights the idea that phenomena are often characterized as input-output relations realized by mechanisms. He, then, argues that mutual manipulability can be interpreted in terms of causal relations between an input and the mechanism, and the mechanism and an output. Harinen, thus opts for the first horn by modeling constitutive mechanistic explanations as a special kind of etiological mechanistic explanations, where a phenomenon is explained by the sequence of its preceding causes.

Other authors opt for the second horn. For example, Felipe Romero argues that ideal interventions into mechanisms are simply impossible if one wants to uphold the claim that constitutive relevance is a non-causal relation (Romero 2015). Instead, he defines constitutive relevance in terms of fat-handed interventions[4]. According to his *fat-handedness criterion*, mechanistic components of phenomena are such that, first, every intervention into the phenomenon is a fat-handed intervention on the phenomenon and at least one component. Second, every intervention into a component is a fat-handed intervention into the component and the phenomenon (2015, 3746).

[2] Although, in the life sciences, non-interventionist studies are surely crucial as well (Kästner 2015).

[3] A strategy I will not discuss in this paper is to simply reject the idea that constitutive relevance can be spelled out in terms of interventionism in general. For example, Couch (2011) and Harbecke (2010) defend a regularity account of mechanistic constitution. Gillett (2013) argues that his dimensioned realization approach best accounts for mechanistic constitution. Gebharter (2016) investigates to which extend the PC algorithm, that was originally developed for detecting causal relations, can be used to discover constitutive relevance relations.

[4] Woodward characterizes interventions as fat-handed if they affect "not just *X* and other variables lying on the route from *I* to *X* to *Y*, but also other variables that are not on this route and that affect *Y*" (Woodward 2008, 209).

A third strategy is to resolve the dilemma. One way to do so that can be found in the literature is by introducing a modified notion of an ideal intervention (based on Woodward (2015); I will call the resulting notion *ideal* intervention*) that renders ideal* interventions into mechanisms possible while accepting that constitutive relevance is a non-causal relation. Michael Baumgartner and Alexander Gebharter (2015) (similar to Baumgartner and Casini (2016)) argue that in order to avoid the consequence that constitution turns out to be causal, one has to introduce time into the definition of interventionist causation. The existence of an ideal* intervention allows for an inference to causation only if the putative cause-variable changes temporally before the putative effect-variable. In order to get a definition of constitutive relevance in terms of this modified version of interventionism they introduce a *fat-handedness criterion* similar to Romero's. A crucial difference is that, according to Baumgartner and Gebharter, this criterion only requires that the relation between a mechanistic component and the phenomenon is such that every intervention into the phenomenon-variable is a fat-handed intervention into the phenomenon and one of the component-variables (Baumgartner and Gebharter 2015, 752).

The distinction between causation and constitution is generally accepted among the new mechanists. Hence, they do not follow Leuridan's suggestion (I will present some arguments in favor of the distinction in the next section). Furthermore, the distinction between constitutive and etiological mechanistic explanation has been accepted since Salmon (1984) and it is useful for epistemic and metaphysical purposes. Hence, opting for the first horn of the dilemma does not seem to be a good option. The second and third strategy, which are in fact rather similar, seem to provide better options. Thus, in this paper, I will focus on solutions based on *fat-handedness*. I will argue that both versions of the fat-handedness approach presented above fail as adequate accounts of constitutive relevance.

I will provide a novel account of constitutive relevance that also rests on a *resolution* of the dilemma based on the modified notion of an ideal* intervention. My account avoids the problems of the fat-handedness approach by interpreting mutual manipulability in terms of causal relations between the constituents and *temporal parts* of the constituted phenomenon. Still, my solution avoids the first horn of the dilemma, and thus, differs from Leuridan's and Harinen's, since I maintain that constitutive relevance is not a causal relation (but manipulability is).

The paper proceeds as follows: in Section 2, I will introduce the general idea of constitutive mechanistic explanation and Carl Craver's mutual manipulability account. In Section 3, I will present the problems that are usually taken to arise when combining

constitutive relevance and interventionism in more detail. I will argue that in order to solve these problems two challenges have to be met. First, one has to provide a notion of an intervention whose application accounts for the difference between causation and constitution. Second, one has to show how constitutive relevance can be spelled out in terms of interventionism without rendering it a causal relation. In Section 4, I will discuss different strategies for meeting the first challenge. In Section 5, I will present answers to the second challenge. I will first (5.1) present and criticize the fat-handedness approaches. Then (5.2), I will present my own solution. Section 6 concludes.

2. Constitutive Explanations, Mechanisms, and Mutual Manipulability

There are various different characterizations of mechanisms on the market. Despite their differences, one common assumption is that mechanisms consist of entities/parts/objects and their activities/interactions/operations in a certain organization (P. Machamer, Darden, and Craver 2000; Craver 2007b; Illari and Williamson 2012; Craver and Darden 2013; Glennan 2017). Based on this characterization, usually two kinds of mechanistic explanations are distinguished: first, in *etiological* mechanistic explanations, a phenomenon is explained by the mechanism that *causes* it. Second, in *constitutive* mechanistic explanations, a phenomenon is explained by the *underlying* mechanism. Here, mechanism and phenomenon are not related by causation. Rather, the mechanism is taken to *constitute* the phenomenon.

One prominent example of a constitutive mechanistic explanation is the explanation of spatial memory. Spatial memory is often investigated by observing mice navigating the Morris water maze (a pool filled with an opaque liquid; the mouse is supposed to find a platform that is hidden under the surface of the liquid). Spatial memory is usually described as being instantiated in the mouse's navigating the Morris water maze (the phenomenon), and the mouse's hippocampus generating spatial maps is supposed to be a component of the mechanism responsible for the navigation behavior (Craver 2007b, 165–70; Bechtel 2008, 49–88; Bechtel and Richardson 2010, 134–44). Other examples of phenomena that are constitutively explained are the action potential (Craver 2007b, 114–22), the human heart pumping blood (Bechtel and Abrahamsen 2005, 425; Bechtel 2006, 29–30; Glennan 2010, 257; Craver and Darden 2013, 98–117), a cell synthesizing proteins (P. Machamer, Darden, and Craver 2000, 18–21; Darden 2002, 357; Craver and Darden 2013, 31–34, 164–67), and long-term potentiation at synapses of neurons (P. Machamer, Darden, and Craver 2000, 8–11; Craver and Darden 2001, 115–17; Craver and Darden 2013, 167–72; Craver 2007b, 65–72).

What exactly does it mean to *constitutively* explain a phenomenon? When does a mechanism *constitute* a phenomenon? Craver illustrates the notion of this constitution-relation with the help of the following figure:

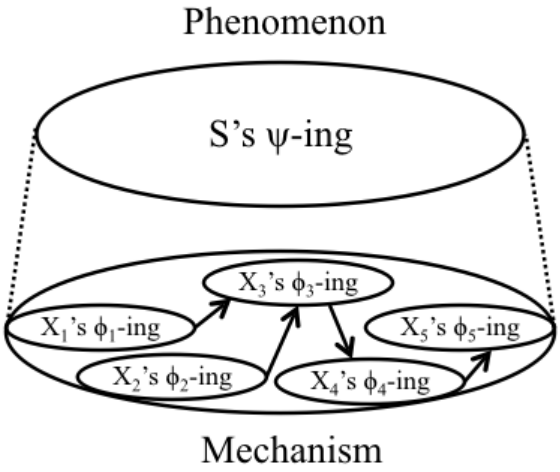

Figure 1: Mechanisms that constitute phenomena according to Craver (adapted from Craver 2007a, 7).

According to Craver, "S's Ψ-ing is explained by the organization of entities $\{X_1, X_2, \ldots, X_i\}$ and activities $\{\Phi_1, \Phi_2, \ldots, \Phi_n\}$" (Craver 2007b, 7). "S" refers to the mechanism as a whole; Ψ is the behavior of S that is to be explained. The X_is represent the entities that are components of the mechanism, and the Φs are the activities performed by the entities. The arrows stand for the different interactions between the X_is' Φ_i-ing. Although the picture alone clearly does not provide a full understanding of what constitutive explanations are, it provides us with some information: first, the phenomenon and the mechanism occur at the same time (indicated by the fact that the phenomenon is located above the mechanism rather than on the right hand side). Second, the relation between the mechanism and the phenomenon is not causal (indicated by the dotted lines; causal relations are represented by arrows). Both assumptions are commonly accepted among the new mechanists (Baumgartner and Gebharter 2015; Casini and Baumgartner 2016; Leuridan 2012; Romero 2015; Craver 2007b).

What exactly is the relation between the mechanism/the mechanism's components and the phenomenon? According to Craver (2007b; 2007a), in constitutive mechanistic explanations one refers to components of mechanisms that are *constitutively relevant* for the phenomenon. He specifies this notion in his mutual manipulability account:

(*Constitutive Relevance*) X's Φ-ing is constitutively relevant for S's Ψ-ing iff:

(i) X's Φ-ing is a part of S's Ψ-ing,[5] and

(ii) X's Φ-ing and S's Ψ-ing are mutually manipulable. (Craver 2007b, 153)

This definition of constitutive relevance provides us with some information about what it means for a mechanism to constitute a phenomenon: first, the components of the mechanism are *spatiotemporal parts* of the phenomenon. For example, the hippocampus is a spatial part of the mouse, and the hippocampus's activity occurs during the mouse's navigation behavior. Second, constitutive relations involve mutual manipulability. Mutual manipulability is taken to consist in two manipulative steps:

(*Mutual Manipulability*) X's Φ-ing and S's Ψ-ing are mutually manipulable iff:

(i) there is an ideal intervention on X's Φ-ing with respect to S's Ψ-ing that changes S's Ψ-ing;

(ii) there is an ideal intervention on S's Ψ-ing with respect to X's Φ-ing that changes X's Φ-ing. (Craver 2007b, 153)

Ideal interventions, thereby, are spelled out in terms of Woodwardian interventionism (Woodward 2003). Roughly, an intervention *I* on X's Φ-ing with respect to S's Ψ-ing is ideal if and only if X's Φ-ing changes only due to the influence of I and S's Ψ-ing is not changed by *I* directly (I will say more on that issue in the next section).

Besides these two features that are explicitly stated, constitutive relevance, according to Craver, has the following further features: first, constitutive relevance is a non-causal relation. If an X's Φ-ing is constitutively relevant for S's Ψ-ing, then it is impossible that X's Φ-ing causes S's Ψ-ing (Craver 2007b; Craver and Bechtel 2007). Most arguments in favor of this claim are based on the fact that X's Φ-ing and S's Ψ-ing are not wholly distinct events because X's Φ-ing occupies a sub-region of the spatiotemporal region of S's Ψ-ing (condition i.) and both are mutually dependent (condition ii.) (Craver and Bechtel 2007; Romero 2015).[6]

[5] Craver's original definition of constitutive relevance requires only X to be a part of S (condition (i)). But it is plausible to assume that constitutive relevance does not only require the component-entities to be spatial parts of the phenomenon-entity. Additionally, the behavior that is to be explained and the behavior of the components that explain the phenomenon have to occur *at the same time* (Leuridan 2012; Baumgartner and Gebharter 2015; Krickel under review).

[6] A further argument in favor of the claim that constitutive relevance is a non-causal relation is that otherwise we are confronted with causal loops (Craver and Bechtel 2007; Baumgartner and Gebharter 2015; Romero 2015). Causal loops are held to be problematic because, roughly, their existence would imply that something could be the cause of its own cause (Kim 1999; Romero 2015). Causal loops are unproblematic if they are taken to be feedback loops where an effect is a cause of an effect that is of the same type as the cause of the first effect. The problematic element comes in when assuming that the putative cause and effect occur at the same time. A further argument Romero provides is based on exclusion worries arguing that if constitutive relevance

One platitude about causation is that its relata have to be wholly distinct (Craver and Bechtel 2007; Lewis 1973; 1986). Second, constitutive relevance (and mechanistic constitution) are not relations of strong, or "more robust kinds of" emergence (Craver 2015, 2). The properties of the phenomenon are determined by the properties of the parts and it is impossible to change the properties of the phenomenon without changing the properties of at least some of the parts. Third, constitutive relevance should allow for multiple realizability (or multiple constitutability) in the sense that there might be two different mechanism types that constitute the same phenomenon type, or two different components that might be interchangeably involved in the same mechanism. As a consequence, it might be the case that some changes in a mechanistic component have no effect on the phenomenon.

3. The Incompatibility of Constitutive Relevance and Interventionism

As Leuridan (2012), Baumgartner and Gebharter (2015), and Romero (2015) argue, there is a problem for Craver's mutual manipulability account: if one want to uphold the intuition that constitutive relevance is not a causal relation, constitutive relevance relations are incompatible with interventionism. To see this, let me introduce the most central notions of the interventionist framework.

First, note that interventionists use variables to model causal structures. Interventionism states that a variable X is a (direct) cause of a variable Y if and only if there is an *ideal intervention* on X with respect to Y that changes Y given that all variables that are not on the causal path between I, X and Y are held fixed (Woodward 2003). An intervention I on X with respect to Y is ideal if and only if X changes only due to changes in I, I does not change Y directly, there is no further variable Z that is not on the causal path between I and X that is changed by I and influences Y, and I is not probabilistically correlated with a further variable Z' that causes Y over a path not going through X (see Figure 2 for an illustration of an ideal intervention).

would be a causal notion that would imply that there is downward causation, which is not compatible with the causal closure of the physical. This argument does not depend on the assumption that the relata of constitutive relevance are parts and wholes. I will not discuss this objection in the present paper. Surely, given that I will argue that constitutive relevance can be spelled out in causal terms and given that this implies that there is downward-causation, exclusion worries have to be addressed. For a discussion of exclusion worries in the context of interventionism see Baumgartner (2009), Woodward (2015), and Gebharter (2015). A discussion of exclusion worries concerning causation between mechanistic levels see Craver (2007b, chap. 6), and between parts and wholes see Kim (1998, 84).

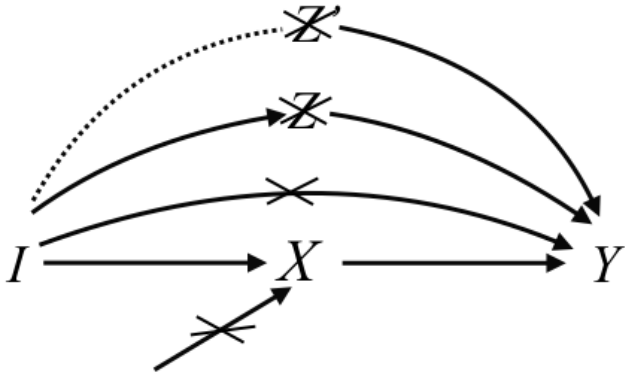

Figure 2 Ideal intervention *I* on *X* with respect to *Y*.

Now consider mutual manipulability, which requires there to be ideal interventions on the phenomenon-variable with respect to each component-variable, and ideal interventions on each component-variable with respect to the phenomenon-variable (see Figure 3). I use variables Φ_1-Φ_3 to represent the mechanism's components and variable Ψ to represent the phenomenon. The interventions $I_{\Phi 1}$, $I_{\Phi 2}$ and $I_{\Phi 3}$ are interventions on the respective component-variable with respect to the phenomenon; I_{Ψ} indicates that there must be interventions on the phenomenon-variable with respect to each component-variable.

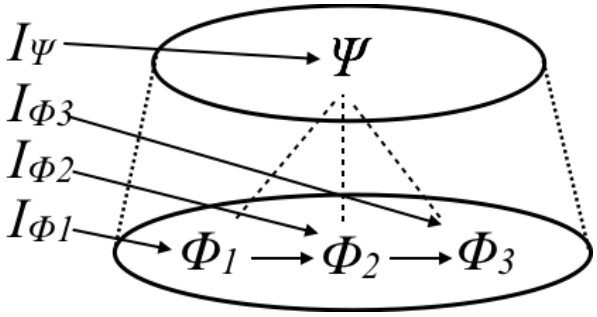

Figure 3 Mutual manipulability requires there to be ideal interventions on the phenomenon-variable Ψ with respect to every component-variable and ideal interventions on every component-variable with respect to the phenomenon-variable (adopted from Baumgartner and Gebharter (2015, 741).

The first problem is that in constitutive relationships (as depicted in Figure 3) it is impossible to ideally intervene on the phenomenon. The reason is that, since the phenomenon-variable constitutively depends on the components, any intervention on Ψ is also an intervention into one of the components (i.e. it is *fat-handed*; Baumgartner and Gebharter 2015; Romero 2015). To see that, assume that Φ_1 is constitutively relevant for Ψ and that there is a correlation between changes in Φ_1 and Ψ when an intervention I_{Ψ} on Ψ is performed. Now, according to interventionism (and Reichenbach's common cause assumption; see Romero 2015), this correlation is either due to the fact that Ψ causes Φ_1, or that Φ_1 causes Ψ, or that *I* is a common cause of Φ_1 and Ψ. Since constitutive relevance is supposed to be a non-causal relation, the only possible explanation for the correlation between Φ_1 and Ψ is that *I* is a

common cause. But ideal interventions cannot be common causes (see definition and Figure 2). Hence, there are no ideal interventions into phenomena that are constituted by mechanisms.

Baumgartner and Gebharter (2015) suggest that this problem can be solved by introducing Woodward's modified definition of an ideal intervention (Woodward 2014; I will talk about *ideal* interventions*). Roughly, Woodward argues that an intervention *I* on *X* with respect to *Y* is ideal* if *I* does not influence any variable that is not on the causal path between *I* and *X* except *for variables that* X *non-causally depends on* (by e.g. definitional dependence, supervenience, realization, constitution). Figure 4 illustrates an ideal* intervention (the dotted line between *X* and *X** indicates that there is a non-causal dependency relation between them).

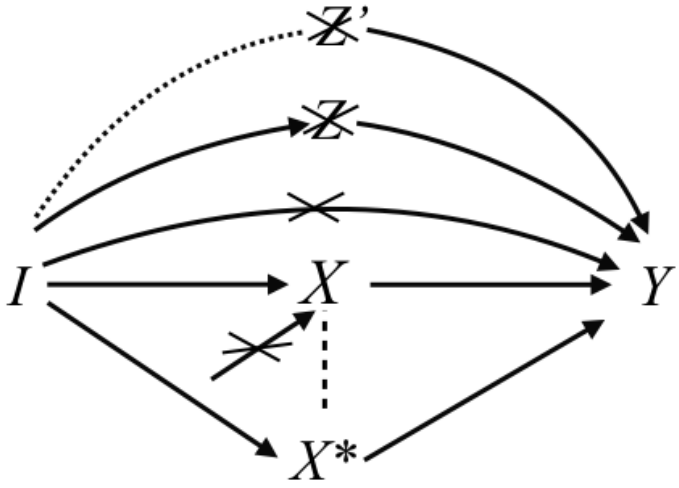

Figure 4 Ideal* interventions according to Woodward (2014); ideal* interventions can be common causes of two variables that are related by a non-causal dependency relation.

This strategy renders ideal* interventions into mechanisms possible because ideal* interventions can be common causes of variables that are non-causally related.

Still, this strategy gives rise to a second problem (Baumgartner and Gebharter 2015): based on this modification, constitutive relevance turns out to be causal. To see this, note that Woodward does not only modify the notion of an ideal intervention but also the definition of causation. The modified interventionist account of causation states that *X* is a (direct) cause of *Y*, iff there is an ideal or ideal* intervention *I* on *X* with respect to *Y* that changes *Y* and all other variables that are not on the causal path between *I*, *X* and *Y* are kept fixed *except for variables that* I, X *and* Y *non-causally depend on*. Now, it turns out that according to this modified account, the existence of an ideal* intervention on Ψ with respect to a component-variable Φ_i establishes that Ψ is *a cause* of the respective component-variable. This is due to

the fact that the modified account of interventionist causation is compatible with the fact that the other component-variables change if we ideally* intervene on Ψ with respect to Φ_i. Since Ψ non-causally depends on all component-variables, the modified account of interventionist causation states that we need not keep them fixed. Hence, the two necessary and together sufficient conditions for causation are satisfied for interventions on the phenomenon with respect to its components. This contradicts the assumption that constitutive relevance is a non-causal relation (see Section 2).

In order to avoid these problems we have to meet two challenges: first, we have to find a way to formulate interventionism such that it does not render constitutive relevance causal. Second, we have to provide a formulation of interventionism such that it also allows for interventions into constitutive relationships and that can be used to spell out constitutive relevance in interventionist terms. I will address both challenges in the next two sections.

4. First Challenge

In order to solve the problems presented in Section 3, the first challenge is to find a definition of interventionism that does not render constitutive relevance causal. Baumgartner and Gebharter (2015) address this challenge by introducing time into the interventionist definition of a cause: if we are dealing with causation in mechanisms, changes in the cause variable occur earlier than changes in the effect variable. In contrast to that, they argue, constitutive relevance requires simultaneous changes in the variables.

Romero addresses the challenge by simply rejecting Woodward's modification of the notion of an ideal intervention. For Romero, interventions into constitutive relevance relations are not ideal and, thus, do not indicate causal relations. Of course, then Romero is confronted with the first problem that there are no ideal interventions into mechanisms. I will discuss his solution to this problem in the next section.

Another possible way to address the first challenge is to make sure that the variables in our causal model only represent events that are wholly distinct (this is usually an implicit requirement for causal models). "Wholly distinct" means that the relata must not logically, definitionally, or metaphysically depend on each other (Lewis 1973; 1986). Woodward and Hausman (1999) argue that

> [w]hen variables bear conceptual or logical connections to one another, or when their located values have parts in common, then they may bear probabilistic relations to one another that have no causal explanation. (523)

They specify in a footnote that

> [t]oken causal relations obtain among distinct events; that is, among distinct instantiations of properties at particular spatio-temporal locations or among spatio-temporally distinct located values of variables. (523, fn. 4)

For events to be wholly distinct they have to occupy distinct space-time regions or the existence of one of the events does not depend on the existence of the other. The relata of constitutive relevance are not wholly distinct because they occupy the same space-time region (or rather, one occupies a sub-region of the space-time region of the other) and the existence of the constituted event depends on the existence of the constituting event. Hence, they are not wholly distinct. Now, one could argue that an ideal intervention I on X with respect to Y that changes Y indicates causation only if all values of I, X and Y represent events that are wholly distinct. Since the values of variables that are constitutively related are not wholly distinct, interventions on these variables that are correlated with changes in the other do not indicate that the variable is a cause of the other.

I will accept this last solution for the purposes of the present paper. Furthermore, I will accept Woodward's modified definition of an ideal* intervention and interventionist causation. Hence, I reject Romero's suggestion for how to address the first challenge. Baumgartner's and Gebharter's solution is similar to the one I accept. The requirement that cause and effect have to be wholly distinct is satisfied if causation takes time. But my requirement is less specific because it is open to whether causes precede their effects, occur afterwards, or occur simultaneously.

In the following section, I will discuss two approaches to meeting the second challenge: how to define constitutive relevance in terms of interventionism.

5. Second Challenge: Constitutive Relevance in Terms of Interventionism

5.1. Approaches Based on Fat-handedness

We still need a positive account of constitutive relevance in terms of interventionism. Baumgartner and Gebharter (2015) and Romero (2015) develop approaches that make use of the fact that interventions on mechanistic phenomena are always common causes. Interventions that are common causes of two (or more) variables are called "fat-handed." What has to be shown is how common causes that are due to constitutive relevance between Φ and Ψ differ from accidental common cause structures and how this difference can be expressed in terms of interventionism. For example, imagine you intervene on a driving car by smashing it with a huge hammer, and thereby you stop the car and you destroy the windshield. Your intervention is a common cause of the car's stopping and the destruction of

the windshield. But this does not show that the latter is constitutively relevant for the former. Rather, it is what I call an "accidental" common cause. Fat-handedness approaches rely on the idea that common cause structures that are due to a constitutive relevance relation differ from accidental common cause structures in that, in the former case, *every* intervention is fat-handed (call this the "fat-handedness criterion"), while in the latter case only some interventions are fat-handed. As I will show, it is unclear how to spell out this idea in detail in order to make sense of the difference between the hammer smashing example mentioned above, and cases that involve constitutive relevance relations.

As already indicated in the introduction, the approaches developed by Romero (2015), and Baumgartner and Gebharter (2015) differ. Roughly, according to Baumgartner and Gebharter's approach, the fat-handedness criterion applies to interventions on the phenomenon with respect to any of its components. In other words, they state that in order to establish constitutive relevance every intervention on the phenomenon-variable has to be a fat-handed intervention into the phenomenon-variable and *at least one* of its components (Baumgartner and Gebharter 2015, 152). This way of spelling out the fat-handedness criterion does not provide us with any resources to exclude the windshield from being constitutively relevant to the car's driving (see example above). The reason is that the account is compatible with some components to be manipulable by means of *only one* fat-handed intervention on the phenomenon. But what distinguishes this component from the windshield in the example mentioned above?

In contrast to Baumgartner and Gebharter, Romero's approach applies to individual components. Romero suggests the following working definition of constitutive relevance in terms of fat-handed interventions (Romero 2015, 3746; my reconstruction)[7]:

(*Fat-handed CR*) A variable Φ_i of a variable set $V = \{\Phi_1, \ldots, \Phi_n\}$ is constitutively relevant for a variable Ψ iff:

(i) Φ_i represents a spatiotemporal part of what is represented by Ψ,

[7] Romero argues that *Fat-handed CR* provides only a necessary condition for constitutive relevance because "[a]ny time we have a fat-handed intervention between two variables we cannot infer that we are talking about a constitutive relevance relation in a mechanism (…) some fat-handed interventions are so because they aren't refined enough to have a localized effect in one variable only" (Romero 2015, 3748). This comment is a bit confusing because his definition states requirements concerning *every* intervention. Surely, it is an open question how we can empirically test for constitutive relevance based on Romero's approach. But the approach seems to exclude the consequence that accidentally fat-handed interventions turn out to establish constitutive relevance because in these cases we can expect to find at least one intervention that is not fat-handed. Hence, it is not clear why the account should provide only necessary conditions.

(ii) every intervention on Ψ is necessarily a fat-handed intervention on Ψ and one of the variables of set V,[8]

(iii) every intervention on Φ_i is necessarily a fat-handed intervention on Φ_i and Ψ.

The first condition corresponds to the first condition of Craver's mutual manipulability approach. The second and third conditions constitute a reinterpretation of Craver's mutual manipulability condition.

Fat-handed CR is problematic. The first reason is that it is not clear how the "necessarily" should be interpreted given the interventionist framework. It is not clear what the "necessarily" adds beyond the assumption that *every* intervention is fat-handed. According to the very idea of fat-handedness approaches to constitutive relevance, invoking a fat-handedness requirement that applies to *every* intervention of a certain kind (as formulated in (ii) and (iii)) just means to spell out what it means for a relation to hold necessarily in interventionist terms. Hence, adding "necessarily" seems to be redundant. Of course, this may not be a big problem for Romero's approach. But it is misleading at least.

The second problem is that, from an empirical perspective, the question of whether constitutive relevance holds in a given case turns out to be a non-definite issue. In order to find constitutive relevance, one has to establish that *every* intervention of a certain kind is fat-handed. From an empirical perspective it is impossible to give a definite answer. The only thing one can do is to test as many interventions as possible and if they turn out to be fat-handed one can abductively infer that this is due to the fact that the variables are constitutively related (Casini and Baumgartner 2016; Baumgartner and Gebharter 2015). It is questionable whether scientists really make these abductive inferences. Even worse, the fact that single fat-handed interventions are always underdetermined with regard to inferences to the underlying structure plus the fact that there is always a simple, purely causal interpretation (in terms of common causes) might raise doubts with regard to whether there is sufficient justification for the whole enterprise of looking for an extra dependency relation, i.e., constitutive relevance.[9]

The third reason why *Fat-handed CR* is problematic is that it does not allow for multiple realization. Multiple realization requires there to be interventions on components that do not

[8] In his definition, Romero does not explicitly mention that the intervention influences „one of the variables of set V." But since he mentions it later in his text and since it seems to be more plausible than requiring that each intervention into the phenomenon has to be an intervention into *one particular* component, I reconstruct his approach in this way.

[9] I want to thank an anonymous reviewer for highlighting the severe consequences of this feature of fat-handedness approaches.

change the phenomenon. This requires there to be interventions onto components that are not fat-handed interventions into a component and the phenomenon. Thus, condition (iii) is too strong. But if there can be constitutively relevant parts that can be surgically intervened on (i.e., not by a fat-handed intervention), and we, thus, have to reject condition (iii), *Fat-handed CR* fails as an account of constitutive relevance (indeed, Baumgartner and Gebharter's approach can account for multiple realization—this is why they do to invoke a condition analogous to (iii)). The problem cannot be solved by simply dropping condition (iii). Condition (ii) is silent with regard to particular single components but allows for inferences only with regard to sets of variables. Hence, Romero's approach would collapse into Baumgartner's and Gebharter's account—which, as mentioned above, is useless when it comes to specifying how to identify individual constituents independently of a complete set of constitutively relevant variables.

A fourth problem of the fat-handedness approach is that it cannot distinguish between top-down and bottom-up interventions since all interventions are always on both levels at the same time. This is problematic because it defeats central motivations of the original mutual manipulability account, and of the whole endeavor of combining constitutive relevance with interventionism. First, Craver introduced the notions of top-down and bottom-up interventions to make sense of important research strategies in the life sciences. According to Craver, interference experiments, stimulation experiments, and activation experiments differ in whether they are so-called *bottom-up experiments* (performing bottom-up interventions) or top-down experiments (performing top-down interventions) (Craver 2007b, 146–47). Second, spelling out constitutive relevance in terms of *mutual* manipulability was supposed to exclude mere *background conditions* from being mechanistic components. According to Craver, background conditions are not top-down manipulable. For example, the heart's beating is a background condition of the word-stem completion mechanism—it is necessary for the mechanism to run that the heart is beating. Still, it is not a component in the mechanism for word-stem completion because one cannot change the phenomenon (word-stem completion behavior) such that the heart's beating changes (Craver 2007b, 157). Still, background conditions seem to satisfy condition (ii): since they are necessary causes of mechanistic components (assuming that the phenomenon indeed occurs), every intervention into a background condition is a common cause of the background condition and the phenomenon (since the latter is constituted by an effect of the background condition). Therefore, based on fat-handedness approaches, it seems to be impossible to distinguish between background conditions and components.

A fifth problem arises from the choice of variables. Romero (and Baumgartner and Gebharter as well) represents the phenomenon S's Ψ-ing by just one variable. This seems to imply that either phenomena are non-complex things that can be individuated by one property only, or that a very complex variable is used that represents various different properties. None of these assumptions seems to be adequate. First, phenomena are taken to be *multifaceted* (Craver 2007b, 125). Craver illustrates this idea with help of the action potential:

> "[a]ction potentials are complex phenomena, when compared to shattering and dissolving. Part of characterizing the action potential phenomenon involves noting that action potentials are produced under a given range of *precipitating conditions* (for example, a range of depolarizations in the cell body or axon hillock). (…) there is more to be said about the *manifestations* of an action potential. It is necessary to describe its rate of rise, its peak magnitude, its rate of decline, its refractory period, and so on." (Craver 2007b, 125)

Explaining the action potential consists in, among other things, explaining its rate of rise, its peak magnitude, its rate of decline and its refractory period. Each component of the action potential mechanism is crucial for a different aspect; not every component is crucial for every aspect. In order to account for this complexity in an interventionist framework, one has to introduce multiple variables.

Second, using just one complex variable is not an option for the following reason: assume we introduce a variable $\Psi_{complex}$ that represents the rising rate, peak magnitude, rate of decline and the refractory period. Now, assume that, necessarily, there are fat-handed interventions on that variable and a particular component-variable – the criteria for constitutive relevance according to the fat-handedness approach are satisfied. We can infer that the component variable is constitutively relevant for a variable representing rising rate, peak magnitude, rate of decline, and refractory period. But this does not tell us anything about what the component is relevant for exactly. But when developing mechanistic explanations we want to know which component of the mechanism is responsible for which aspect of the phenomenon.

Sixth, a further consequence of the fact that the phenomenon is represented by one variable only is that the fat-handedness approach cannot account for the fact that phenomena often consist in changes over time (for a modeling account that integrates the temporal dimension see Gebharter and Schurz (2016)). I will call this the *temporal heterogeneity* of mechanistic phenomena. Different components of the mechanism are crucial for different features of the phenomenon at different times. Take again the action potential. The action potential is (at least partly; see quote above) characterized by its rising rate, peak magnitude, declining rate and its refractory period. These features are temporally ordered: first, the

potential rises, then it peaks, then it declines, then the refractory period is reached. Explaining the action potential means to account for this temporal order. Only some components are crucial for the rising phase, or the peak magnitude. No component is crucial for the action potential during the whole time of its occurrence but only for certain aspects of the action potential at specific time points. If the phenomenon is represented by just one variable, again, one has to choose between two options: either one reduces the complexity of the phenomenon and treats it as homogeneous through time; or one uses a complex variable that represents different properties and, therefore, different times of the phenomenon's occurrence. Again, we run into the problems already discussed above.

In the next section, I will present my own suggestion that does not rely on fat-handedness and that solves these problems.

5.2. Taking Mutual Manipulability Seriously

In the remainder of this paper, I will to develop a novel solution to the problem of how to define constitutive relevance in terms of interventionism. This approach starts with Craver's original definition of constitutive relevance. The inconsistency problem is solved by giving mutual manipulability a causal interpretation while maintaining the idea that constitutive relevance and causal relevance are mutually exclusive relations. The core idea of my approach is that manipulability consists in a causal relation between what I call *temporal acting-entity-parts* of the phenomenon and the mechanism's components. To get there, we first have to take a closer look at constitutive relevance from a metaphysical perspective.

Craver (2007b) assumes that the relata of constitutive relevance and mechanistic constitution are *acting entities*. What are acting entities? First, acting entities are concrete individuals (not universals, or the like). In some sense, acting entities are events (Kaiser and Krickel 2016): they involve an entity that is doing something, where this doing takes time. Acting entities are things that are spatially and temporally extended in such a way that they can be divided into different parts. For present purposes, the most relevant parts are what I will call *spatial acting-entity-parts* (Figure 5 a)) and *temporal acting-entity-parts* (Figure 5 b)). I call them "acting-entity-parts" to highlight the fact that both kinds of parts are themselves acting entities. Before I define these concepts, I will illustrate the idea with help of an example: the mouse navigating the Morris water maze is an acting entity (the mouse is the entity; the navigation behavior is the activity); its spatial acting-entity-parts are, for example, the mouse's muscles moving, or the hippocampus generating spatial maps while the mouse is navigating. Its temporal acting-entity-parts are, for example, the mouse being put

into the pool, the mouse swimming for a while in one direction, swimming into another direction, and finally finding the platform, where all these activities happen during the navigation behavior.

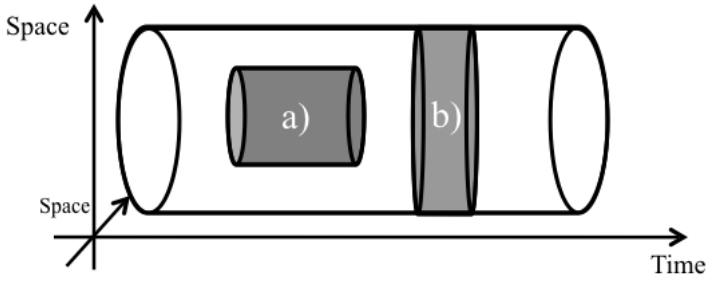

Figure 5 An acting entity and what I call a) a *spatial acting-entity-part*, and b) a *temporal acting-entity-part* of it.

Temporal acting-entity-parts are defined as follows:

> (*temporal acting-entity-part*) An acting entity E_1 is a temporal acting-entity-part of another acting entity E_2 iff:
>
> (i) the entity involved in E_1 is identical with the entity involved in E_2;
>
> (ii) the activity involved in E_1 occurs at the same time as the activity involved in E_2 and starts later or ends earlier than the latter.

Spatial acting-entity-parts of acting entities are defined as follows:

> (*spatial acting-entity-part*) An acting entity E_1 is a spatial acting-entity-part of another acting entity E_2 iff:
>
> (i) the entity involved in E_1 occupies a proper sub-region of the spatiotemporal region occupied by the entity involved in E_2;
>
> (ii) the activity involved in E_1 occurs at the same time as the activity involved in E_2.

Now, the crucial idea is that acting entities that are spatial acting-entity-parts and acting entities that are temporal acting-entity-parts *of one and the same* acting entity can, in principle, causally interact (if they occur in different spatio-temporal regions). This idea is illustrated in Figure 6.

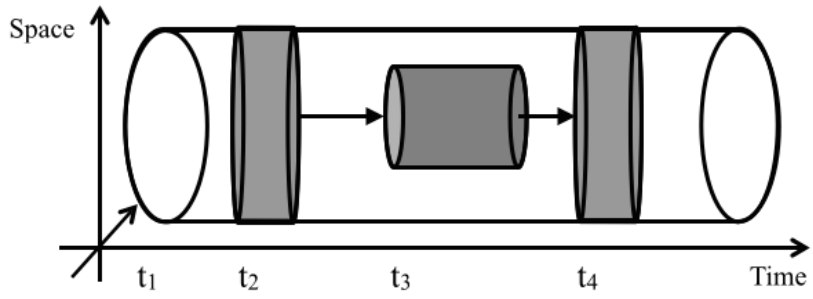

Figure 6 Causal interactions between spatial and temporal acting-entity-parts of one and the same acting entity. The x-axis shows time; the y- and z-axis represent spatial dimensions.

Consider the mouse navigating the Morris water maze. Assume this navigation behavior lasts from t_1 to t_4. Now, imagine that during this process the mouse turns left at t_2 (which is a temporal acting-entity-part of the mouse's navigation behavior). This event might be a cause of the behavior of the hippocampus starting at t_3 (which is a spatial acting-entity-part of the mouse's navigation behavior). This behavior, again, might be a cause of the mouse's finding the platform at t_4 (which is another temporal acting-entity-part of the mouse's navigation behavior).

Spatial and temporal parts of one and the same acting entity can in principle causally interact because they do not necessarily occupy the same space-time regions. In other words, they can be wholly distinct events. Remember that the general argument for why events related by constitutive relevance cannot be causally related was that they are not wholly distinct. This objection does not apply here.

Based on these metaphysical considerations, we can solve the problem of how to spell out constitutive relevance in terms of interventions. First, note that since the relata are taken to be concrete individuals, we have to model constitutive relevance in the interventionist framework Woodward developed for *actual causation* that holds between events that actually occurred. According to Woodward's account of actual causation (Woodward 2003, 77), a token event is represented by a variable taking a specific value (written X=x). A value of a variable X=x is a (direct) cause of a value of a variable Y=y iff there is an ideal or (ideal*) intervention on X with respect to Y that changes the value of Y given that all other causes of Y that are not on the route between I, X, and Y are kept fixed at their actual values (except for variables that I, X, or Y non-causally depend on). Second, since we are talking about temporal acting-entity-parts of the phenomenon, we have to represent the phenomenon by more than one variable, where each phenomenon-variable represents a temporal acting-entity-part of the phenomenon. Third, according to what I have argued in Section 4, the values of the variables

have to represent wholly distinct events in order for the variables to be able to stand in causal relations.

On the basis of these considerations, what I will call *Causation-based CR* can be defined as follows:

(*Causation-based CR*) X's Φ-ing is constitutively relevant for S's Ψ-ing iff

(i) X's Φ-ing is a spatial acting-entity-part of S's Ψ-ing,

(ii) there is a temporal acting-entity-part of S's Ψ-ing that is a cause of X's Φ-ing, and

(iii) there is a temporal acting-entity-part of S's Ψ-ing that it is an effect of X's Φ-ing.

In order to formulate causation-based CR in terms of interventionism, we have to represent the different acting entities by variables. I will represent X's Φ-ing by the variable Φ_i, and the temporal acting-entity-parts of S's Ψ-ing by Ψ_1 and Ψ_2. Since, as stated above, we are dealing with event/acting entity tokens, we have to apply Woodward's interventionist account of actual causation. In this terminology, condition (ii) is satisfied if and only if there is a variable $\Psi_1=\psi_1$ for which it is true that there is an ideal or ideal* intervention on Ψ_1 with respect to $\Phi_i=\varphi_i$ that changes Φ_i while all other variables not on the causal path between Ψ_1 and Φ_i are kept fixed except for variables that Ψ_1 and Φ_i non-causally depend on (Figure 7, (b) and (d)). Condition (iii) is satisfied if and only if there is a variable $\Psi_2=\psi_2$ for which it is true that there is an ideal or ideal* intervention on $\Phi_i=\varphi_i$ that changes Ψ_2 while all other variables not on the causal path between Φ_i and Ψ_2 are kept fixed except for variables that Ψ_2 and Φ_i non-causally depend on (Figure 7, (a) and (c)). Plausibly, the temporal acting-entity-parts of a phenomenon constitutively depend on different components of the underlying mechanism. Therefore, interventions in the present context are plausibly ideal* (Figure 7, (c) and (d)).

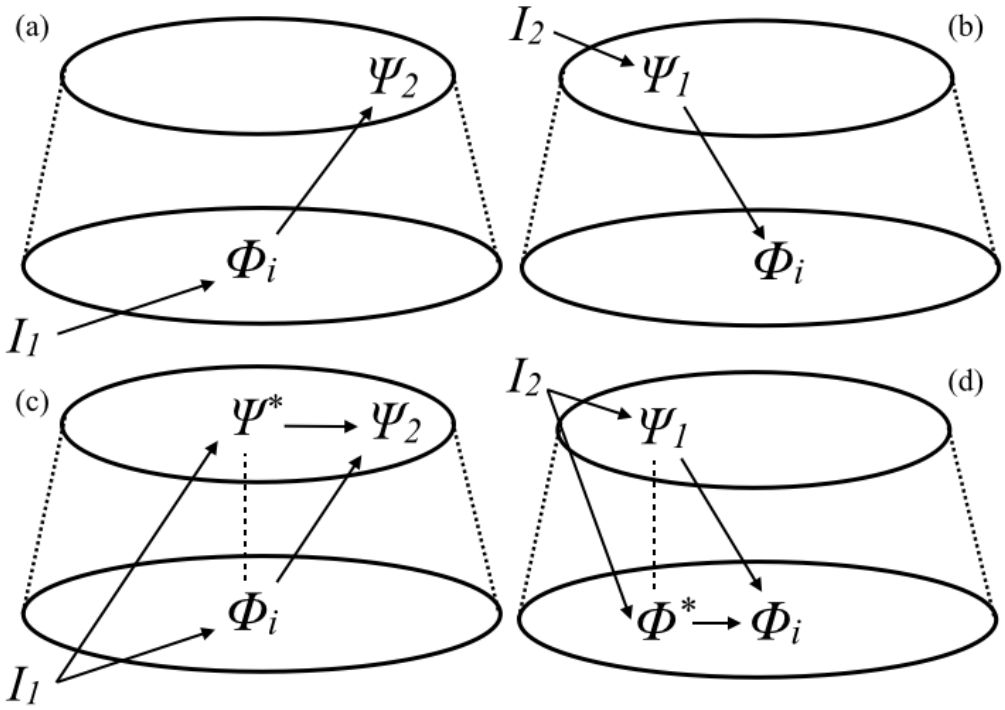

Figure 7 *Causation-based CR* in terms of interventionism requires either two ideal interventions (a) (bottom-up intervention) and (b) (top-down intervention), or two ideal* interventions (c) (bottom-up intervention) and (d) (top-down intervention)

The models depicted in Figure 7 allow for a straightforward interpretation of top-down and bottom-up interventions (in interventionist terms and in metaphysical terms) which was a crucial motivation for Craver's mutual manipulability approach. In bottom-up interventions, the variable that is intervened on represents events at the lower level, while the variable *with respect to which* one intervenes represents events at the higher level. In the case of top-down interventions it is the other way around.

How does my approach handle the example of the mouse navigating the Morris water maze? Imagine one particular mouse navigating through the maze from t_1 to t_4. Imagine we want to know whether the hippocampus's activity at t_3 was constitutively relevant to that phenomenon. In order to answer this question we have to verify, first, whether there was a temporal acting-entity-part of the mouse's navigation behavior for which it is true that had there been an ideal* intervention on that temporal acting-entity-part with respect to the hippocampus's activity at t_3, then the hippocampus's activity at t_3 would have been different. For example, one could imagine that had there been an ideal* intervention on the entering of the mouse into the maze at t_1 (e.g. a change in the location of where the mouse was put into the maze), then the hippocampus's activity would have been different at t_3 (e.g. different neural representations in the hippocampus would have been active). Second, we have to

verify whether there was a temporal acting-entity-part of the mouse's navigation behavior for which it is true that had there been an ideal* intervention on the hippocampus's activity with respect to that temporal part, the temporal acting-entity-part would have been different. Again, it seems plausible that had there been an ideal* intervention on the hippocampus's activity at t_3 the mouse's finding the platform at t_4 would have been different (e.g. it would have found the platform at a later time). Hence, my account renders the hippocampus constitutively relevant for the mouse's navigation behavior.

Now assume that the mouse's stomach was active at t_3. The stomach's activity does not seem to be constitutively relevant (assuming a normal mouse and a normal stomach). Plausibly, it is not the case that had there been an ideal* intervention into any temporal acting-entity-part before t_3 with respect to the stomach's activity at t_3, then the stomach's activity would have been different. Neither it seems to be plausible that had there been an ideal* intervention into the stomach's activity at t_3 with respect to, for example, the finding of the platform at t_4, then the finding of the platform at t_4 had been different.

There are two crucial challenges to my account: first, one has to tell a story about how scientists can actually test for causation-based constitutive relevance. The problem seems to be that there cannot be interventions into causal processes that already occurred. Second, one has to show how to get to general claims about constitutive relationships from here. Surely, scientists are not so much interested in whether the hippocampus's activity is relevant for the behavior of *that particular* mouse. Rather, they are interested in what constitutes this kind of behavior in general.

Both challenges might be met by adequately comparing sufficiently similar concrete particulars. For example, one compares two instances of a mouse navigating the Morris Water Maze that differ only with respect to the hippocampus's activity at a particular time. If these instances differ also with respect to the mouse's behavior afterwards, this indicates a causal relation between the hippocampus's activity at that time and the mouse's behavior at that later time. This would correspond to an ideal* bottom-up intervention. Then, one compares two instances of the navigation behavior that differ only with regard to, for example, where the mouse is put into the maze. If this is correlated with differences in the hippocampus's activity during the navigation behavior compared to the other instance, this indicates a causal relation between the first temporal acting-entity-part of the phenomenon (the mouse being put into the maze at a specific location) and the hippocampus's activity at a later time. In order to reach generalization, scientists have to perform these experiments with a number of mice (or other animals) in order to exclude the possibility that the effects they

observed where due to individual features of the mice that have been investigated. For example, there might be a mouse that has a rather weak stomach such that it starts growling every time the mouse moves quickly and that affects the movements depending on how loudly it growls. In this case, the stomach's growling would be constitutively relevant to the moving of that particular mouse. Since not all mice have weak stomachs in this way, the stomach's growling will not come out as constitutively relevant to the movements of mice in general.

This reinterpretation of Craver's mutual manipulability approach has several advantages over the fat-handedness approach presented in the previous section. First, my approach provides a clear criterion of how to distinguish between interventions that are accidental common causes and interventions that are common causes due to the fact that the variables are non-causally related. If the variables that the intervention is a common cause of are mutually manipulable, then these variables are constitutively related. If they are not mutually manipulable, the intervention is an accidental common cause.

Second, according to my approach, the question whether there is a constitutive relevance relation in a given case is in principle a definite enterprise. In contrast to fat-handedness approaches, my approach requires only that there are *two* ideal or ideal* interventions. Since, according to my approach, mutual manipulability is a causal notion, the empirical detectability of constitutive relevance is only as problematic as the empirical detection of causal relevance.

Third, my approach makes sense of Craver's notions of top-down and bottom-up manipulability. Since fat-handedness approaches assume that mutual manipulability consists in fat-handed interventions, these approaches are unable to make sense of this distinction (see Figure 7). Craver's notions of top-down and bottom-up interventions are taken to provide a useful and adequate account of the empirical practice of interlevel experiments. It is an advantage of my account that it maintains the descriptive adequacy of Craver's original approach.

Fourth, as argued in the previous section, fat-handedness approaches represent the phenomenon by only one variable. Thereby they cannot account for the complexity of mechanistic phenomena (they are multifaceted and temporally heterogeneous; see Section 5.1). I can avoid this problem by rejecting the idea that the phenomenon has to be represented by one variable only. Representing the phenomenon by introducing multiple variables representing different temporal parts of the phenomenon allows for a representation of

different properties of the phenomenon that are instantiated at different times during the occurrence of the phenomenon.

There might be one final objection: what about components that occur at the same time as the very first temporal acting-entity-part of the phenomenon, and what about those that occur synchronously with the very last temporal acting-entity-part of the phenomenon? Condition (ii) cannot be satisfied for the former, since there are no temporal parts of the phenomenon that occur earlier than the component. Condition (iii) cannot be satisfied for the latter, since there are no temporal parts of the phenomenon that occur later than the component. What does it mean for the first/last component of a mechanism to be constitutively relevant to the phenomenon? My answer is straightforward: We have to divide the temporal acting-entity-part of the phenomenon into temporal parts such that one turns out to be a cause/an effect of the component. In other words: we simply have to apply *Causation-based CR*.

6. Conclusion

It has been argued that Craver's mutual manipulability account of constitutive relevance is inconsistent because it is spelled out in terms of interventionism. Constitutive relevance relations do not seem to allow for ideal interventions if one wants to uphold the claim that constitutive relevance is a non-causal relation. I have discussed different strategies for how to solve this problem. Two challenges arose: first, one has to provide a definition of interventionism that provides an adequate account of causation that does not render constitutive relationships causal. Second, one has to provide a definition of constitutive relevance in terms of interventionism. Different strategies for meeting the first challenge were discussed. The strategy I followed in this paper is to require the variables in a causal model to represent wholly distinct events. Concerning the second challenge, different authors have introduced fat-handedness approaches. I argued that they are problematic. I developed a new suggestion of how to spell out constitutive relevance in terms of interventionism that accounts for Craver's original idea of mutual manipulability. Roughly, mutual manipulability, according to my account, consists in the causal manipulability of the mechanism's components by ideally or ideally* intervening into a temporal acting-entity-part of the phenomenon, and in the causal manipulability of a temporal acting-entity-part by ideally or ideally* intervening into the component. As I argued, this approach has several advantages over fat-handedness approaches.

Acknowledgements

I want to thank Alexander Gebharter, Felipe Romero, Jens Harbecke, Lena Kästner, Lorenzo Casini, and Marie Kaiser for comments on earlier versions of this paper. Furthermore, I want to thank Michael Baumgartner and my colleagues at RUB for helpful discussions regarding the topic of my paper. Furthermore, this paper profited a lot from the feedback I got at various workshops and conference such as the workshop *Ground in Philosophy of Science* (Geneva 2016), the workshop *Mechanistic Integration and Unification in Cognitive Science* (Warsaw 2016), the *GWP*-conference (Düsseldorf 2016), *EPSA*-conference (Düsseldorf 2015), and the workshop *Hempel and Beyond* (Cologne 2015).

References

Baumgartner, Michael. 2009. "Interventionist Causal Exclusion and Non-Reductive Physicalism." *International Studies in the Philosophy of Science* 23 (2): 161–78.

Baumgartner, Michael, and Alexander Gebharter. 2015. "Constitutive Relevance, Mutual Manipulability, and Fat-Handedness." *British Journal for the Philosophy of Science* 67 (3): 731–56. doi:10.1093/bjps/axv003.

Bechtel, William. 2006. *Discovering Cell Mechanisms*. Cambridge: Cambrdige University Press.

———. 2008. *Mental Mechanisms*. *Philosophical Perspectives on Cognitive Neuroscience*. New York/London: Routledge.

Bechtel, William, and Adele Abrahamsen. 2005. "Explanation: A Mechanist Alternative." *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences* 36 (2 SPEC. ISS.): 421–41. doi:10.1016/j.shpsc.2005.03.010.

Bechtel, William, and Robert C. Richardson. 2010. *Discovering Complexity. Decomposition and Localization as Strategies in Scientific Research*. Cambridge: MIT Press.

Casini, Lorenzo, and Michael Baumgartner. 2016. "An Abductive Theory of Constitution." *Philosophy of Science*. doi:10.1086/690716.

Couch, Mark B. 2011. "Mechanisms and Constitutive Relevance." *Synthese* 183 (3): 375–88. doi:10.1007/s11229-011-9882-z.

Craver, Carl F. 2015. "Levels." In *Open MIND*, edited by Thomas K Metzinger and Jennifer M Windt. Frankfurt am Main: MIND Group. doi:10.15502/9783958570498.

Craver, Carl F. 2002. "Interlevel Experiments and Multilevel Mechanisms in the Neuroscience of Memory." *Philosophy of Science* 69 (S3): S83–97. doi:10.1086/341836.

———. 2007a. “Constitutive Explanatory Relevance.” *Journal of Philosophical Research* 32 (Section II): 1–20. doi:10.5840/jpr_2007_4.

———. 2007b. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. New York: Oxford University Press.

Craver, Carl F., and William Bechtel. 2007. “Top-down Causation Without Top-down Causes.” *Biology & Philosophy* 22 (4): 547–63. doi:10.1007/s10539-006-9028-8.

Craver, Carl F., and Lindley Darden. 2013. *In Search of Mechanisms. Discoveries across the Life Sciences*. Chicago/London: University of Chicago Press.

Craver, Carl F, and Lindley Darden. 2001. “Discovering Mechanisms in Neurobiology: The Case of Spatial Memory.” In *Theory and Method in Neuroscience*, edited by P K Machamer, Rick Grush, and Peter McLaughlin, 112–37. Pittsburgh: University of Pitt Press.

Darden, Lindley. 2002. “Strategies for Discovering Mechanisms: Schema Instantiation, Modular Subassembly, Forward/Backward Chaining.” *Philosophy of Science* 59 (S3): 54-S365. doi:10.1086/341858.

Gebharter, Alexander. 2015. “Causal Exclusion and Causal Bayes Nets.” *Philosophy and Phenomenological Research*, 1–23. doi:10.1111/phpr.12247.

———. 2016. “Uncovering Constitutive Relevance Relations in Mechanisms.” *Philosophical Studies*. doi:10.1007/s11098-016-0803-3.

Gebharter, Alexander, and Gerhard Schurz. 2016. “A Modeling Approach for Mechanisms Featuring Causal Cycles.” *Philosophy of Science* 83 (5): 934–45.

Gillett, Carl. 2013. “Constitution, and Multiple Constitution, in the Sciences: Using the Neuron to Construct a Starting Framework.” *Minds and Machines* 23 (3): 309–37. doi:10.1007/s11023-013-9311-9.

Glennan, Stuart. 2010. “Mechanisms, Causes, and the Layered Model of the World.” *Philosophy and Phenomenological Research* 81 (2): 362–81. doi:10.1111/j.1933-1592.2010.00375.x.

———. 2017. *The New Mechanical Philosophy*. Oxford University Press.

Harbecke, Jens. 2010. “Mechanistic Constitution in Neurobiological Explanations.” *International Studies in the Philosophy of Science* 24 (3): 267–85. doi:10.1080/02698595.2010.522409.

Harinen, Totte. 2014. “Mutual Manipulability and Causal Inbetweenness.” *Synthese*, no. February (October). doi:10.1007/s11229-014-0564-5.

Hausman, Daniel M, and James Woodward. 1999. “Independence, Invariance and the Causal

Markov Condition." *The British Journal for the Philosophy of Science* 50 (4): 521. doi:10.1093/bjps/50.4.521.

Illari, Phyllis McKay, and Jon Williamson. 2012. "What Is a Mechanism? Thinking about Mechanisms across the Sciences." *European Journal for Philosophy of Science* 2 (1): 119–35. doi:10.1007/s13194-011-0038-2.

Kaiser, Marie I, and Beate Krickel. 2016. "The Metaphysics of Constitutive Mechanistic Phenomena." *The British Journal for the Philosophy of Science* 0: 1–35. doi:10.1093/bjps/axv058.

Kaplan, David M. 2012. "How to Demarcate the Boundaries of Cognition." *Biology and Philosophy* 27 (4): 545–70. doi:10.1007/s10539-012-9308-4.

Kästner, Lena. 2015. "Learning About Constitutive Relations." In *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, edited by Uskali Mäki, Ioannis Votsis, Stéphanie Ruphy, and Gerhard Schurz, 155–391. doi:10.1007/978-3-319-23015-3.

Kim, Jaegwon. 1998. *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. MIT Press.

———. 1999. "Making Sense of Emergence." *Philosophical Studies* 95 (1). Springer: 3–36.

Krickel, Beate. n.d. "Making Sense of Interlevel Causation in Mechanisms from a Metaphysical Perspective."

Leuridan, Bert. 2012. "Three Problems for the Mutual Manipulability Account of Constitutive Relevance in Mechanisms." *British Journal for the Philosophy of Science* 63 (2): 399–427. doi:10.1093/bjps/axr036.

Lewis, David K. 1973. "Causation." *Journal of Philosophy* 70 (17): 556–67. doi:10.2307/2025310.

———. 1986. "Events." In *Philosophical Papers Vol. II*, edited by David Lewis, 241–69. Oxford University Press.

Machamer, Peter, Lindely Darden, and Carl F. Craver. 2000. "Thinking About Mechanisms." *Philosophy of Science* 67 (1): 1–25.

Romero, Felipe. 2015. "Why There Isn't Inter-Level Causation in Mechanisms." *Synthese* 192 (11): 3731–55. doi:10.1007/s11229-015-0718-0.

Salmon, Wesley C. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

———. 2008. "Invariance, Modularity, and All That: Cartwright on Causation." In *Nancy*

Cartwright's Philosophy of Science, edited by Stephan Hartman, C. Hoefer, and L. Bovens, 198–237). New York: Taylor & Francis.

———. 2015. "Interventionism and Causal Exclusion." *Philosophy and Phenomenological Research* 91 (2): 303–47. doi:10.1111/phpr.12095.