Ergo AN OPEN ACCESS JOURNAL OF PHILOSOPHY

Constitution, Non-Causal Explanation, and Demarcation

MICHAEL BAUMGARTNER
University of Bergen

LORENZO CASINI
University of Bologna

In philosophy of science, constitutive explanations have attracted much attention since Craver's influential book *Explaining the Brain* (2007). His Mutual Manipulability (MM) theory of constitution aimed to explicate constitution as a non-causal explanatory relation and to demarcate between constituents and non-constituents. But MM received decisive criticism. In response, Craver et al. (2021) have recently proposed a new theory, called Matched Interlevel Experiments (MIE), which is currently gaining traction in various fields. The authors claim that MIE retains "the spirit of MM without conceptual confusion." Our paper argues that this claim is not borne out: neither does MIE meet MM's objectives nor is it free of conceptual confusion. At the same time, we show that it is possible to meet MM's objectives in a conceptually sound manner—by adopting the so-called No De-Coupling theory of constitution.

1. Introduction

The core question to be answered in Carl Craver's influential book *Explaining the Brain* (2007), according to its blurb, is: "What distinguishes good explanations in neuroscience from bad?" The answer, in short, is that neuroscientific explanations are *constitutive* and good constitutive explanations "describe all and only the component entities, activities, properties, and organizational features that are relevant to the multifaceted phenomenon to be explained" (Craver 2007: 111). The explanatory heavy-lifting in such explanations is done by the relation

Contact: Michael Baumgartner <michael.baumgartner@uib.no>
Lorenzo Casini <lorenzo.casini3@unibo.it>

https://doi.org/10.3998/ergo.6924

of constitution or constitutive relevance,[1] of which Craver's book offers a widely received account, *viz.* the Mutual Manipulability theory (MM). MM states, in essence, that a phenomenon's constituents are the activities of those of its spatio-temporal parts that can be changed by ideal interventions on the phenomenon as a whole and that can change the phenomenon if they are themselves changed by ideal interventions.

MM establishes constitution as a relevance relation between wholes and their parts, different from causation, and it is intended to demarcate between constituents and non-constituents. On the basis of these demarcations, the quality of a neuroscientific explanation can then be cashed out in terms of the degree to which it successfully identifies all and only the explanandum's constituents. But MM has been exposed to severe criticism (e.g., Baumgartner & Gebharter 2016; Baumgartner & Casini 2017; Harbecke 2010; Harinen 2018; Leuridan 2012; Romero 2015). Craver et al. (2021) acknowledge that MM has certain problems. They insist, however, that a lot of the criticism is unfounded because critics fail to properly distinguish between metaphysical and epistemological aspects of constitution. In response, they develop a new theory stating that, metaphysically, constituents are characterized by being causally between a phenomenon's input and output conditions and that, epistemologically, constituents can be discovered by "a novel 'matched interlevel experiments' (MIE) criterion that retains the spirit of MM without conceptual confusion" (8809).

This new theory is currently gaining considerable traction in various fields. In the philosophy of medicine, for example, Varga (2023: 13) advocates it as a means of clarifying how to generate biomedical understanding by keeping separate "a horizontal (causal) dimension and a vertical (part-whole) dimension" in difference-making relationships. In the philosophy of biology, it has been endorsed by Weber (2022) and, in the literature on mechanistic explanation, by Krickel et al. (2023). Moreover, the new theory has been widely adopted in the literature on extended cognition and mind. Parise et al. (2023: 5) maintain that MIE is helpful in planning experiments determining the boundaries of cognitive processes. Smart (2022) argues that MIE establishes that a wearable holographic device and the holograms it produces are constituents of the human cognitive process of protein structure prediction. Gillett et al. (2022: 7) contend that the "reformulated version of MM works in practise and can be used effectively to determine the boundaries of cognition" and they "urge proponents and opponents of [extended cognition] to use this modified version of MM in other putative cases to continue to push the debate forward."

1. We use the notions of constitution and constitutive relevance interchangeably in this paper.

The first part of our paper critically assesses the merits of the new theory relative to the core question Craver set out to answer in his (2007). On the one hand, we find that it does not retain the spirit of MM. Contrary to MM, the new theory does not establish constitution as a relevance relation different from causation; rather, it yields that constitutive explanation just is (a kind of) causal explanation. In addition, MIE does not demarcate between constituents and non-constituents, as it remains entirely silent about non-constituents. Unlike MM, therefore, it provides no basis for distinguishing good from bad constitutive explanations either. On the other hand, we argue that the new theory is not without conceptual confusion, for it is inherently unclear what Craver et al. (2021) mean by "being causally between" a phenomenon's input and output conditions. We distinguish three possible interpretations of that metaphysical thesis and find that none of them combines coherently with MIE. Overall, this critical part of the paper demonstrates that all attempts to distinguish between causal and non-causal difference-making relations or to demarcate mechanisms—cognitive or other—based on MIE are futile, and it rebuts the value of MIE for generating mechanistic understanding or explanations in the philosophy of medicine, biology, and elsewhere.

The paper's second part then shows that the upshot of these negative findings is not that any attempt to explicate constitution as a distinctly non-causal relation in a way that demarcates between constituents and non-constituents necessarily falls prey to conceptual problems like MM or MIE. We demonstrate that there does in fact exist a theory of constitution that successfully achieves these objectives, *viz.* the No De-Coupling theory (NDC; see Baumgartner & Casini 2017).

2. Mutual Manipulability

According to the new mechanist literature (Machamer et al. 2000; Glennan 2002; Bechtel & Abrahamsen 2005; Craver 2007), a mechanistic explanation describes how the joint operation of lower-level entities and activities is responsible for an upper-level phenomenon. This responsibility has several dimensions—spatio-temporal, causal, and hierarchical. In contrast to previous accounts of explanation, which gave a dominant role to causation (cf. Lewis 1986; Salmon 1984; Woodward 2003), causation is but one explanatory dimension among others for mechanists. A phenomenon is mechanistically explained by the hierarchical arrangement of its component entities and activities with respect to the phenomenon and its behaviour as well as by the spatio-temporal configuration of these components and their causal interactions. Hierarchy is the dimension of a mechanistic explanation that is theoretically least understood.

Craver (2007) famously spelled the hierarchical component of an explanation out in terms of the relation of *constitution* or *constitutive relevance*. He conceives of

constitution as a relevance relation that obtains between an upper-level activity Ψ of a system *S*, *S*'s Ψ-ing, and a lower-level activity Φ of a spatio-temporal part *X* of *S*, *X*'s Φ-ing. Craver analyzes that relation in his Mutual Manipulability theory (MM) of constitutive relevance. The following passage provides the most explicit formulation of MM:

> In sum, I conjecture that to establish that *X*'s Φ-ing is relevant to *S*'s Ψ-ing it is sufficient that one be able to manipulate *S*'s Ψ-ing by intervening to change *X*'s Φ-ing (by stimulating or inhibiting) and that one be able to manipulate *X*'s Φ-ing by manipulating *S*'s Ψ-ing. To establish that a component is irrelevant, it is sufficient to show that one cannot manipulate *S*'s Ψ-ing by intervening to change *X*'s Φ-ing and that one cannot manipulate *X*'s Φ-ing by manipulating *S*'s Ψ-ing. (Craver 2007: 159)

There are (at least) two reasons why MM is attractive to many. First, it establishes constitution as a relevance relation different from causation. Causation obtains between spatio-temporally non-overlapping entities and it is a one-directional dependence relation, meaning that effects depend on causes but not vice versa. By contrast, according to MM, constitution obtains between spatio-temporally overlapping entities and is a bi-directional dependence relation such that phenomena and constituents mutually depend on each other. Hence, MM grounds a distinctive non-causal type of explanation, which Craver claims to be important and widespread in many sciences, including neuroscience.[2] Second, MM aims to demarcate between constituents and non-constituents by echoing experimental practice in neuroscience, as reconstructed by Craver (2007: 60, 105, 131, 144). He argues that neuroscientists tend to combine "top-down" and "bottom-up" experiments, which, respectively, manipulate the upper-level phenomenon and measure its lower-level effects, and manipulate a phenomenon's parts and measure their upper-level effects.

> Neuroscientists use these experiments to establish which parts are components in a mechanism and which are not, that is, to distinguish relevant components from mere constitutive correlates, […] sterile effects, and background conditions. (144)

2. Craver's (2007) book provides one of the most prominent expositions of the view that there is more to mechanisms than causation. The chapter introducing his account of constitution starts with a distinction between two types of explanation: "Some mechanistic explanations are etiological; they explain an event by describing its antecedent causes. […] Other mechanistic explanations are constitutive or componential; they explain a phenomenon by describing its underlying mechanism" (107-108).

But Craver (2007) does not merely want MM to reflect actual experimental practice, which is error prone and not ideal, rather he wants MM's demarcations to be normatively adequate (106-107). Therefore, he requires that experimental manipulations meet the standards of *ideal interventions* as defined by Woodward (2003).[3] Ideal interventions on *S*'s Ψ-ing with respect to *X*'s Φ-ing must surgically cause *S*'s Ψ-ing, such that all causal influence on *X*'s Φ-ing, if any, goes via *S*'s Ψ-ing—and analogously for ideal interventions on *X*'s Φ-ing (see Figure 1a). However, recent literature has shown that such ideal interventions on *S*'s Ψ-ing are impossible. Since *S*'s Ψ-ing supervenes on its lower-level constituents, all changes in *S*'s Ψ-ing must be associated with changes in some constituent or other. Hence, manipulations causing changes in *S*'s Ψ-ing also cause changes on the lower level. If we assume—with Craver (2007) and many others—that constitution is a non-causal relation and that phenomena are not reducible to their constituents, it follows that all manipulations of *S*'s Ψ-ing are connected to the lower level on causal paths not going through *S*'s Ψ-ing, in violation of surgicality. Therefore, only non-ideal interventions on phenomena, as in Figure 1b, are possible (Baumgartner & Gebharter 2016). This, in turn, depending on the logical form MM is taken to have, either reduces MM to absurdity or renders it vacuous (Baumgartner & Casini 2017).

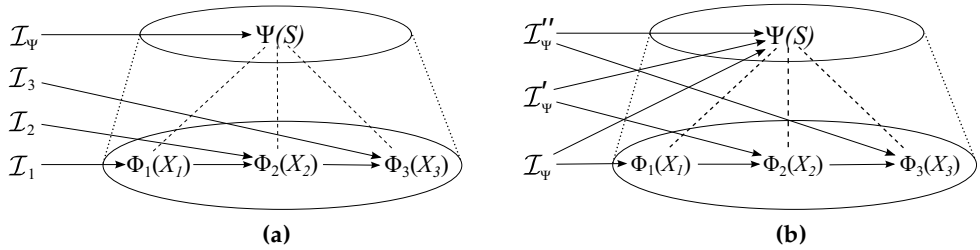


Figure 1: Mechanistic diagrams adapted from Craver (2007). Specific variables (e.g., Ψ(*S*)) represent entities' activities (e.g., *S*'s Ψ-ing). Dotted lines stand for spatio-temporal overlap and dashed lines for constitutive relevance. Surgical interventions as in (a) are required by MM, but only non-surgical manipulations as in (b) are possible.

3. The New Proposal

Craver et al. (2021: 8809) acknowledge that the critics of MM "identify some key obstacles in the way of a coherent and systematic understanding of both the metaphysics and the epistemology of constitutive relevance and inter-level experiments," and they set out to separately clarify epistemological and

3. There exist other definitions of the notion of an intervention (e.g. Pearl 2009), but the one relevant for Craver (and our argument) is Woodward's.

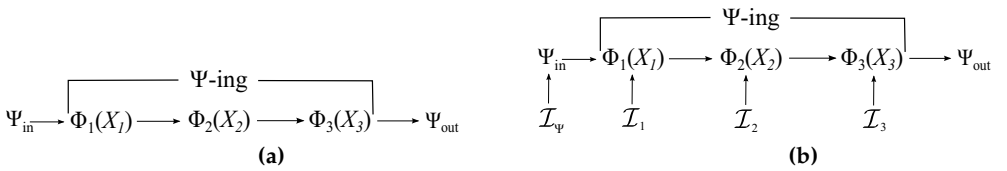


Figure 2: Revised mechanistic diagrams adapted from Craver et al. (2021).

metaphysical issues with constitutive relevance. They suggest that the problems with MM stem from misrepresentations in diagrams as the ones in Figure 1a. The crisp boundary of the upper-level ellipsis is misleading because phenomena often do not involve a well-defined localized entity *S*, rather, "there only is Ψ-ing" (8811). Moreover, these diagrams conflate "synchronous spatial inclusion relations between entities and their parts (i.e., between *S* and the *X*s), and temporal inclusion relations between mechanistic processes and their parts (i.e., between Ψ-ing and Φ-ing)" (8812).

To avoid these problems, Craver et al. replace diagrams as in Figure 1 by diagrams as in Figure 2.

> Here, Ψ-ing is represented as a process beginning with an input, Ψ_{in}, and terminating with an output, Ψ_{out}. Between these temporal endpoints, and a mechanistic level down, is a temporally sequenced causal chain of events, involving the X_i and their various Φ_i. [Figure 2a] makes clear that the problem of constitutive relevance is that of identifying the components of the process bridging Ψ_{in} and Ψ_{out} […] (Craver et al. 2021: 8812)

Contrary to constituents, which Craver et al. still conceive of as activities Φ_i of entities X_i, phenomena are newly conceptualized as mere input-output relations (without acting entities). The phenomenon Ψ-ing is defined by its input and output variables, Ψ_{in} and Ψ_{out}.[4] It is a dependence relation between values of Ψ_{in} and values of Ψ_{out}. In between, there is a causal process, or sequence of events instantiating X_1's Φ_1-ing, X_2's Φ_2-ing, …, X_n's Φ_n-ing, each of which may be a complex object, often involving not only the activity of single components but also interactions among multiple components (8812). For notational convenience, we will henceforth abstain from explicitly relativizing

4. Craver et al. (2021) sometimes say that Ψ_{in}, Ψ_{out} and *X*'s Φ-ing are "events." By that they cannot mean spatio-temporally located token events, which occur exactly once, because the interventionist tools they employ are inapplicable to token events. By "event" they must mean event *types* like "pedaling" or "moving," which are instantiated by many token events and which are modeled by variables. To avoid misunderstandings, we directly speak of "variables" throughout.

variables Φ_i to their acting entities X_i and simply write "Φ_i" instead of "$\Phi_i(X_i)$" and "X_i's Φ_i-ing."

Against that background, Craver et al. propose a new theory of constitutive relevance with an epistemological and a metaphysical component. The former provides a sufficient condition to identify constituents. Reconceptualizing phenomena as input-output relations renders ideal interventions on phenomena no more problematic than they are in ordinary causal modeling contexts, because the surgicality requirement only has to be imposed on the variables in the causal structure connecting Ψ_{in} and Ψ_{out} (cf. Figure 2b). An ideal intervention on the phenomenon Ψ-ing is simply a surgical cause of Ψ_{in}. Craver et al. contend that three types of ideal interventions and a 'matching' condition are jointly sufficient to establish that some Φ is part of the causal structure connecting Ψ_{in} to Ψ_{out} and, thereby, constitutively relevant to Ψ-ing. The result is the matched interlevel experiments (MIE) criterion (Craver et al. 2021: 8822, adjusted to our notational convention):

(MIE) To establish that Φ is constitutively relevant to a mechanism that Ψs, the following experimental results and matching condition are jointly sufficient:

> *(CR1i)* If an experiment initiates conditions Ψ_{in} while a bottom-up intervention, $\mathcal{I}$, prevents or inhibits Φ, alterations to or prevention of Ψ's terminal conditions, Ψ_{out}, are detected.
>
> *(CR1e)* If a bottom-up intervention, $\mathcal{I}$, stimulates Φ, Ψ's terminal conditions, Ψ_{out} are detected.
>
> *(CR2*)* If a top-down experiment initiates conditions Ψ_{in} and detects Ψ's terminal conditions Ψ_{out}, Φ is also detected.
>
> *(Matching)* The activities Φ_i activated or inhibited in bottom-up experiments (CR1i and CR1e) must be of the same kind as, and occur within quantitatively overlapping ranges with, the activities Φ_i detected in top-down experiments (CR2*).

The sufficiency of MIE for constitutive relevance is grounded in the metaphysical component of the new theory, which stipulates that *constitutive relevance is causal betweenness* (8807). Causal betweenness is "the truth-maker for claims about constitutive relevance" (8810). Craver et al. (2021: 8819-8820) take inspiration from Harinen (2018), who views a phenomenon as a causal path $\Psi_{\text{in}} \longrightarrow \Psi_{\text{out}}$ and the problem of constitutive discovery as that of experimentally finding intermediate variables on that path, using experiments as described by (CR1i), (CR1e), and (CR2*).

Before we turn to assessing the merits of Craver et al.'s new theory, one of its features deserves separate emphasis. MIE is sufficient but *not necessary* for the constitutive nature of Φ, because Craver et al. contend that there are constituents for which not both of (CR1i) and (CR1e) are satisfiable. They illustrate failures of (CR1i) with a redundant mechanism: "humans have two kidneys, each one of which can regulate plasma osmolality on its own" (8824). Even though both kidneys are constitutively relevant to the osmolality phenomenon, (CR1i)-experiments removing one kidney have no effect because the other kidney compensates. Failures of (CR1e) are illustrated with the bottom bracket of a bicycle, which holds the spindle connecting the two arms of the pedal crank system. Craver et al. allege (8824) that the bracket is constitutively relevant to the biking phenomenon even though it cannot be stimulated such that the bicycle moves, that is, even though (CR1e)-experiments are impossible.

4. MIE and the Spirit of MM

Craver et al. (2021: 8809) claim that MIE retains the spirit of MM. Whether that is true, of course, depends on what exactly MM's spirit is. We cannot provide an exhaustive analysis of that notion here, but it seems clear that the core objectives MM serves in Craver's (2007) book are core elements of its spirit: first, to establish constitution as a relevance relation that is different from causation; and second, to demarcate between constituents and non-constituents. In what follows, we argue that MIE meets neither of these objectives.

Contrary to MM, MIE does not construe a phenomenon Ψ-ing and its constituent Φ as mereologically related entities that bi-directionally depend on one another. Instead, Φ (one-directionally) depends on Ψ_{in}, and Ψ_{out} depends on Φ, but the phenomenon itself—the input-output relation—is no longer part of the dependence structure. Manipulating Ψ_{in} may induce changes in Φ and manipulating Φ may induce changes in Ψ_{out}. However, changes in Ψ_{in} and in Ψ_{out} are not changes in the input-output relation, that is, in the phenomenon, as this would require changing the mapping of values of Ψ_{in} onto values of Ψ_{out}, which in turn would amount to altering the causal structure itself. To illustrate with Craver et al.'s leading example, the input to the locomotion phenomenon in *C. elegans* can be manipulated by, say, tapping the worm on its head, whereas manipulating the locomotion phenomenon itself requires, for instance, changing *the wiring* of the worm's brain. Yet, manipulations of the phenomenon are irrelevant for MIE. Without mutual or bi-directional dependence, what we are left with is ordinary causal dependence among mereologically independent entities. Correspondingly, the problem of constitutive inference for MIE is

nothing other than the problem of finding the causal structure between Ψ_{in} and Ψ_{out}, for which the literature on causal discovery provides various ready-made solutions. This is incompatible with MM's objective of establishing a new type of explanation. For MIE, constitutive explanation just is (a kind of) causal explanation.[5]

The second core objective of MM is to demarcate between those spatio-temporal parts of a phenomenon that are constitutively relevant and those that are not. To this end, MM must ground the inference to both constitutive relevance and irrelevance; in other words, it must furnish both a sufficient and a necessary condition for constitution. Indeed, the passage most explicitly stating MM, which we quoted in §2, does exactly that: it introduces a sufficient condition for constitutive relevance and a sufficient condition for constitutive irrelevance, which, taken jointly, yields a sufficient as well as a necessary condition for constitution. Unfortunately, though, there are other passages where Craver says that MM offers only a sufficient condition for constitution and he formulates the account with a mere "if" (e.g., Craver 2007: 141, 154).

While the passages explicitly addressing the logical strength of MM are difficult to interpret consistently, the argumentative use Craver (2007) actually makes of MM leaves no room for interpretation. If MM were only sufficient, it could only be used to identify constituents but not non-constituents, because if a phenomenon and one of its parts fail to be mutually manipulable, MM, understood as a mere sufficient condition, entails nothing on the constitutive nature of that part. Yet, Craver unequivocally uses MM to identify non-constituents, too. Violation of MM is repeatedly taken to warrant an inference to a part of a phenomenon being a sterile effect, an isolated part, or a mere correlate (Craver 2007: 149, 156, 259). An exemplary passage is the following:

> One can exclude sterile effects and other mere correlates by requiring that the experiment satisfy CR1 [i.e., the bottom-up manipulability component of MM]. […] Performance of cognitive tasks, for example, is routinely correlated with hemodynamic changes, but this does not mean that the hemodynamic changes are part of the mechanism involved in task performance (as all MRI researchers know). Hemodynamic changes can be ruled out as components of the mechanism on the grounds that intervening to prevent the increase in blood flow during a task will not prevent one from performing the task. (156)

5. A similar point was made by Weinberger (2019) with respect to Harinen's (2018) inbetweenness theory. Craver et al. do not address Weinberger's criticism, even though it directly translates to MIE.

What is more, Craver is very clear that he wants his theory to demarcate constituents from non-constituents (Craver 2007: 60, 105, 112, 144) and to demarcate the boundaries of mechanisms (Craver 2007: 141, 259). These demarcations are then used for assessing to what degree neuroscientific explanations mention *all and only* constituents of the explanandum (Craver 2007: 111, 259), which, in turn, is Craver's quality criterion for these explanations. Of course, only an account that furnishes a sufficient as well as a necessary condition can deliver demarcations and identify all and only constituents.

That is, in order to make sense of a maximally large portion of Craver's discussion of constitution and neuroscientific explanation, MM must be interpreted to deliver a sufficient *and a necessary* condition for constitutive relevance, which, correspondingly, is the interpretation adopted by many commentators (e.g., Baumgartner & Gebharter 2016; Baumgartner & Casini 2017; Krickel, 2018a: 98; Krickel, 2018b; Harbecke 2010: 271). This interpretation is confirmed by Craver et al. (2021: fn 8) who say, in formal explicitness, that MM entails that the conjunction of non-manipulability from the top down and of non-manipulability from the bottom up is sufficient for constitutive irrelevance, which is equivalent to the disjunction of top-down and bottom-up manipulability being necessary for constitutive relevance.[6]

By contrast, it is clear that MIE only furnishes a sufficient condition for constitutive relevance and no sufficient condition for irrelevance, that is, no necessary condition for constitutive relevance. MIE allows to identify constituents but remains silent about non-constituents. If Φ violates one of (CR1i), (CR1e), (CR2*), and (Matching), nothing follows on its constitutive nature. Correspondingly, if a scientist explains a phenomenon Ψ-ing with recourse to some Φ_i that does not satisfy all conditions of MIE, it does not follow that the explanation is bad, because Φ_i might still be a constituent, yet one that cannot be recovered by the experiments described in MIE. Contrary to MM, MIE provides no inferential leverage to demarcate between constituents and sterile effects, isolated parts, fictional posits, etc.[7] Whether constitutive explanations contain all and only the

6. Unfortunately, a coherent picture still does not emerge from Craver et al. (2021) who, in another footnote (fn 7), also insist that MM only provides a sufficient condition for constitution. One rationale offered for this insistence is that the word "sufficient" appears twice in the passage stating MM in (Craver 2007: 159), which we quoted in §2. As the second claim in that passage, *viz.* "non-MM is sufficient for constitutive irrelevance," is equivalent to "MM is necessary for constitutive relevance," this rationale is such a non-starter that charity demands to ignore it, and with it the insistence on 'sufficiency only'.

7. In light of this, it is a mystery why so many advocates of extended cognition (see introduction; in particular Gillett et al. 2022) believe that MIE could neutrally arbitrate whether an extra-bodily entity is a constituent of a cognitive process or not. Being merely sufficient for constitution, MIE is unequivocally biased in favor of advocates of extended cognition, as—for mere logical reasons—it could never rule that extra-bodily entities fail to be constituents.

constituents of phenomena cannot be determined based on MIE. In replacing MM by MIE, the primary purpose Craver's theory of constitution was designed to serve is given up: to distinguish good constitutive explanations (in neuroscience) from bad ones.

In sum, whatever the full spirit of MM might be, a theory like Craver et al.'s new proposal, which treats constitution as a form of causation and provides no handle for demarcation, is undoubtedly *not* in the spirit of MM.

5. Causal Betweenness and Conceptual Confusion

Not following the spirit of a predecessor theory is not itself problematic for a new theory, if that theory turns out to be valuable in its own right. Unfortunately, Craver et al.'s new proposal raises more questions than it answers. In fact, despite avoiding MM's reliance on impossible interventions, it still fails to provide a coherent account of constitution. The source of the problem is that it is unclear how exactly to interpret the metaphysical thesis that constitution is causal betweenness. There is textual support for three different interpretations, each being problematic for different reasons.

5.1. The Path Interpretation

According to the first and most straightforward interpretation, a constituent Φ being causally between a phenomenon's input and output conditions, Ψ_{in} and Ψ_{out}, simply means that Φ is *on a directed causal path* from Ψ_{in} to Ψ_{out}. This interpretation is supported by all of Craver et al.'s (2021) diagrams and by the fact that they follow Harinen's (2018) understanding of the three experimental procedures in MIE (8819). Harinen endorses the "path" interpretation of causal betweenness (as already emphasized by Weinberger 2019) and views the goal of those procedures to be to experimentally find intermediate variables on the causal path $\Psi_{in} \longrightarrow \Psi_{out}$. It is thus plausible that Craver et al. interpret betweenness in the same way.

The path interpretation is in line with Craver et al.'s claim that MIE is sufficient for constitution, because (CR1i), (CR1e), and (CR2*), jointly, indeed suffice to infer a path $\Psi_{in} \longrightarrow \Phi \longrightarrow \Psi_{out}$. To see this, consider Figure 3, which collects the causal structures compatible with the evidence generated by the experiments in (CR1i), (CR1e), and (CR2*). The result of (CR1i) can be accounted for by one or more paths from Ψ_{in} to Ψ_{out}, at least one via Φ, or by a collider at Ψ_{out}. Experiment (CR1e) establishes a path from Φ to Ψ_{out}, while remaining silent about Ψ_{in}. The result of (CR2*) is compatible with there being at least one path from Ψ_{in} to

Ψ_{out} via Φ or a common cause structure. It follows that any model compatible with the evidence produced by all experiments must feature at least one path $\Psi_{in} \longrightarrow \Phi \longrightarrow \Psi_{out}$.

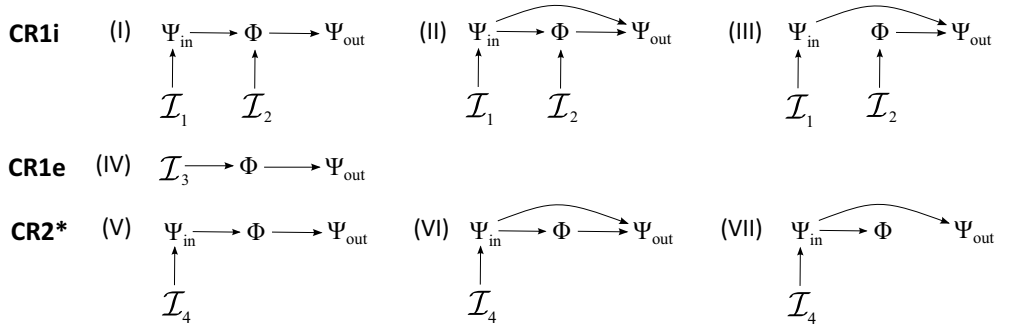


Figure 3: Causal structures compatible with the data generated by the experiments described in (CR1i), (CR1e), and (CR2*), respectively. $\mathcal{I}_1$ to $\mathcal{I}_4$ represent the interventions required by those experiments.

However, Figure 3 also shows that this conclusion does not require all of (CR1i), (CR1e), and (CR2*). Either one of (CR1i) and (CR1e) establishes a path $\Phi \longrightarrow \Psi_{out}$, while (CR2*) alone demonstrates that there is a path $\Psi_{in} \longrightarrow \Phi$, which, jointly, yields a path $\Psi_{in} \longrightarrow \Phi \longrightarrow \Psi_{out}$. Hence, if Craver et al. (2021) really have the path interpretation of causal betweenness in mind, their list of experiments contains a redundancy. Choosing either the combination of (CR1e) & (CR2*) or the combination of (CR1i) & (CR2*) would be sufficient to establish a path. And MIE contains yet another condition: (Matching). Relative to the path interpretation, it is mysterious why Craver et al. impose (Matching), as there is no work left for that condition. All that matters for the existence of a path is that some value of Ψ_{in} makes a difference to some value of Φ, and that some value of Φ, not necessarily the same one, makes a difference to some value of Ψ_{in}. That Craver et al. explicitly require (Matching) on top of three experiments, two of which would already suffice, suggests that they might not interpret causal betweenness in terms of being on a directed path from Ψ_{in} to Ψ_{out} after all.

5.2. The Process Interpretation

A second interpretation can be derived from what Craver et al. say about their preferred theory of causation. Although MIE exclusively draws on inference tools from a causal modeling tradition that understands causation in difference-making terms, Craver et al. (2021) do not subscribe to a difference-making the-

ory of causation. Instead, they express a preference for a production theory of causation (e.g. Glennan 2017). They contend that a "causal dependence requires a productive process between an effect and its cause" (8823). There is a large and diverse literature on causal processes (Russell 1948; Salmon 1984; Dowe 2000; Steel 2007). Processes are widely understood as relations between events or states (or chains thereof) rather than variables. While Craver et al. do not provide a full-blown analysis of productivity, they do say that they "understand productivity in a way that allows for productive continuity" (8823). They stipulate that for "B to lie causally between A and C, there must be a process by which A contributes to the production of B, and a process by which B contributes to the production of C" (8823).

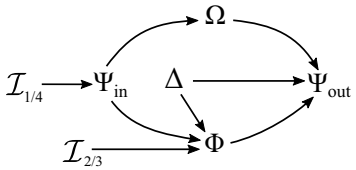
In this second interpretation, Φ being causally between Ψ_{in} and Ψ_{out} could mean that events represented by a value ψ^i_{in} of Ψ_{in}, a value ϕ^j of Φ and a value ψ^k_{out} of Ψ_{out}, respectively, can be *connected by a process with productive continuity*. Adopting the notational convention that "ϕ^j" is short for "$\Phi = \phi^j$," a process can be modeled as an ordered sequence of events, or "stages" (8819), $\langle \psi^i_{\text{in}}, \dots, \phi^j, \dots, \psi^k_{\text{out}} \rangle$, where a change at each stage—alone or jointly with other background conditions—makes a difference to its successor.

Variables may well be connected by a directed causal path via mediators without there being a process connecting concrete values of these variables. An example of such a scenario is due to Neapolitan (2004: 49, 111): finasteride (F) is a drug that lowers testosterone levels (T), and testosterone levels below some threshold θ cause hair growth (G); meaning there is a directed path $F \longrightarrow T \longrightarrow G$. Yet, if finasteride cannot lower testosterone below θ, then F, T, and G have no values f^i, t^j, and g^k that could be connected by a process $\langle f^i, t^j, g^k \rangle$. That is, even though there is a directed path from F to G, no value of F can make a difference to G because of a threshold effect at the mediating link.

It follows that, if causal betweenness is understood in terms of the "process" interpretation, establishing betweenness requires not only establishing the existence of a directed path $\Psi_{\text{in}} \longrightarrow \Phi \longrightarrow \Psi_{\text{out}}$, but also that values of these variables can be connected by a process $\langle \psi^i_{\text{in}}, \phi^j, \psi^k_{\text{out}} \rangle$. In other words, it must be shown that there is some context—meaning some configuration of values of background conditions—where a value ϕ^j of Φ is both the effect of a value ψ^i_{in} of Ψ_{in} and the cause of a value ψ^k_{out} of Ψ_{out}. If that holds, the path $\Psi_{\text{in}} \longrightarrow \Phi \longrightarrow \Psi_{\text{out}}$ is said to be instantiated by an *active causal route* (Hitchcock 2001) connecting the input to the output via the constituent. If we interpret the constraint, expressed in (Matching), that activities Φ_i activated in bottom-up and top-down experiments must match as demanding that the same values of Φ_i figure in both types of experiments, ensuring the existence of an active causal route can be seen to be exactly the purpose of (Matching).

Establishing the existence of an active causal route is a notoriously hard problem, since it presupposes much background knowledge about the structure in question (see below). Still, there exist well-known scientific approaches to address this problem, for instance, *back-door adjustment* in the graphical literature (Pearl 2009: 79), *mediation analysis* in the potential outcomes framework (e.g. Vanderweele et al. 2014), or *tests of contributing causation* in the experimental literature (e.g. Woodward 2003: 50). The latter approach—on the assumption that the existing paths between a cause and an effect are known, that interventions are surgical, and that the background conditions are stable throughout the experiment—is particularly well suited to test whether one such path can be instantiated by an active route. Craver et al. do not mention any of the existing approaches but instead attempt to develop their own. In the remainder of this subsection, however, we will argue that MIE is not sufficient to establish that there is an active causal route from Ψ_{in} to Ψ_{out} via Φ.

To demonstrate this, we introduce a concrete causal structure, in Figure 4, for which all conditions of MIE are satisfied and yet there exists no active route through Φ. The structure contains two paths from Ψ_{in} to Ψ_{out}, one via Φ and one via another intermediate link Ω.[8] Moreover, it features a context variable Δ that represents configurations of background conditions. Δ can take three values $\{\delta^1, \delta^2, \delta^3\}$; all other variables are binary. Ψ_{in} and Δ are exogenous and thus independent of one another; all other variables are endogenously determined as described in Equations (1), (2), and (3). In addition, all variables, except for the background condition Δ, can be altered through surgical interventions, which, for simplicity, are not explicitly modeled in the equations. Moreover, to avoid unnecessary complications, we assume that the structure is deterministic.



$$\omega^1 \leftrightarrow \psi^1_{in} \quad (1)$$

$$\phi^1 \leftrightarrow \delta^1 \vee \delta^3\psi^1_{in} \quad (2)$$

$$\psi^1_{out} \leftrightarrow \delta^1\omega^1\phi^1 \vee \delta^2\phi^1 \vee \delta^3\omega^1 \quad (3)$$

Figure 4: A context-sensitive, two-path structure and its governing equations. Concatenation means conjunction, "∨" disjunction, and "↔" equivalence.

The property of this structure relevant to our argument is that the paths $\Psi_{in} \longrightarrow \Phi$ and $\Phi \longrightarrow \Psi_{out}$ can only be activated in different background contexts. According to Equation (2), $\Psi_{in} \longrightarrow \Phi$ is only active when Δ happens to take value δ^3, whereas (3) stipulates that $\Phi \longrightarrow \Psi_{out}$ is only active when Δ takes values δ^1 or δ^2. That is, Φ is only sensitive to changes in Ψ_{in} in context δ^3, and

8. As the previous subsection has shown, two-path structures are not excluded by MIE.

Ψ_{out} is only sensitive to changes in Φ in contexts δ^1 and δ^2. It follows that there exists no context where causal influence is transmitted from Ψ_{in} to Ψ_{out} via Φ.

The experiments of MIE only succeed if they happen to be conducted in backgrounds where Δ takes values such that the experimental interventions on Ψ_{in} and Φ have the downstream effects expressed in Equations (1) to (3). If experiment (CR1i) is performed when Δ takes value δ^1, an intervention $\mathcal{I}_1$ first sets Ψ_{in} to ψ^1_{in}, which triggers ω^1 by virtue of Equation (1). According to (2), δ^1 causes ϕ^1 independently of ψ^1_{in}, which yields $\delta^1\omega^1\phi^1$ and, as per (3), causes Ψ_{out} to take value ψ^1_{out}. Next, an intervention $\mathcal{I}_2$ sets Φ to ϕ^0, such that $\delta^1\omega^1\phi^1$ is no longer given, meaning that Ψ_{out} turns to ψ^0_{out}. It follows that the conditions of (CR1i) are satisfied without an active route through Φ. Experiment (CR1e) is straightforward: if conducted in a context where δ^2 is given, an intervention $\mathcal{I}_3$ sets Φ to ϕ^1, and $\delta^2\phi^1$ causes ψ^1_{out} subject to (3). Finally, if experiment (CR2*) is run when the background is δ^3, an intervention $\mathcal{I}_4$ sets Ψ_{in} to ψ^1_{in}, which produces ω^1 and ϕ^1. However, when Δ takes value δ^3, Ψ_{out} is only sensitive to changes in Ω, not in Φ. That means that also the conditions in (CR2*) are satisfied without an active route through Φ. Moreover, the values to which Φ is set in the activation and the detection experiments match, in compliance with (Matching). In sum, MIE is satisfied without causal influence ever being transmitted on an active route through Φ. And if there cannot be such an active route, there cannot be events instantiating the path $\Psi_{in} \longrightarrow \Phi \longrightarrow \Psi_{out}$ such that they are connected by a process with productive continuity.

This example shows that, under the process interpretation, MIE is insufficient to establish constitution, contrary to what Craver et al. claim. Ensuring that variables on a causal path are set to matching values does not ensure that this path can be instantiated by an active route. But a test of contributing causation could help out. If the paths $\Psi_{in} \longrightarrow \Omega \longrightarrow \Psi_{out}$ and $\Psi_{in} \longrightarrow \Phi \longrightarrow \Psi_{out}$ have been uncovered via path tests as described in the previous subsection, the existence of an active route along one of these paths can then be investigated by blocking the other path (e.g. by suppressing the middle link) in a stable background context and checking if a surgical change of Ψ_{in} correlates with a change in Ψ_{out}. In the case of our example, this sort of test reveals that, if we surgically intervene on Ψ_{in} (by setting it to ψ^1_{in} or ψ^0_{in}) and block the path $\Psi_{in} \longrightarrow \Omega \longrightarrow \Psi_{out}$ by forcing Ω to take the value it would take if it was not for the intervention on Ψ_{in} (by fixing it to ω^0 or ω^1, respectively), then the intervention on Ψ_{in} cannot make a difference to Ψ_{out} in any of the background contexts δ_1, δ_2, and δ_3. This shows that there is no active route, and thus no process, from Ψ_{in} to Ψ_{out} via Φ. By contrast, if we intervene to block Φ by fixing it, say, to value ϕ^0, it is still possible for Ψ_{in} to make a difference to Ψ_{out} via Ω, namely in context δ^3, for when Δ happens to take value δ^3, interventions on Ψ_{in} are associated with changes in Ψ_{out} despite Φ being fixed to ϕ^0. It follows that the path $\Psi_{in} \longrightarrow \Omega \longrightarrow \Psi_{out}$ can be instantiated

by an active route. In sum, subject to the process interpretation, only Ω is a constituent, but MIE erroneously identifies Φ as a constituent as well.

5.3. The Modulation Interpretation

Not only does there exist a standard test for contributing causation that supports the process interpretation, but that test is—in ideal experimental contexts, at least—both sufficient *and necessary* to detect constituents under that interpretation. That Craver et al. do not use that test and emphatically deny the necessity of their proposal suggests that they might have yet a third interpretation of causal betweenness in mind. While their first counterexample to necessity—the kidney example—could easily be handled by the standard test, their second counterexample—the bicycle bracket—raises questions. Craver et al. define the phenomenon of biking as having an input of pedaling and an output of moving. The bottom bracket is a structural component of the bicycle's drive train, "a stable structure that makes Ψ-ing possible" (8824). They say it is a *standing condition*, which is constitutively relevant to the moving.[9] The bracket holds the spindle in place against the frame and minimizes the friction between the spindle and the frame by means of ball bearings. Importantly, the bracket being a standing condition entails that any variation in the friction or the bracket's capacity to hold the spindle is due to external factors—such as dust, moisture, rust, vibrations, etc.—and not to the operation of the bike's moving parts. In particular, contrary to the rotation of the spindle, the bracket's functionality is not sensitive to changes in pedaling, meaning that the bracket causally influences the moving but is not influenced by the pedaling. Accordingly, neither MIE nor a standard route (or path) test can possibly reveal that the bracket is a constituent of the biking phenomenon. The reason is not a shortcoming of these tests, but simply that *there is no causal path* from pedaling via the bracket to the bicycle's moving, and without a path, there cannot be an active route or a process either.

Craver et al. add that "[t]his toy example is replicated in every mechanism we know: they require (relatively) stable structures as standing conditions to work" (8824). In the causal graph literature exploited by MIE, relatively stable standing conditions are modeled as exogenous causes into a target effect that can modulate the influence of other causes while being independent of the

9. Notice the departure of the new theory from MM, for which background conditions (e.g., the beating heart in the mechanism of word-stem completion) are *not* constituents, even though they may be necessary to the phenomenon (Craver 2007: 152); cf. also the passage from page 144 we quoted in §2.

latter, such that the target effect is a so-called *collider*. For instance, the presence of oxygen *O* is a standing condition for the causal relation between a burning match *M* and a forest fire *F*, because *F*'s dependence on *M* is sensitive to the state of *O*: the match burning makes a difference to the fire only if oxygen is present. Thus, *M* interacts with *O* to jointly cause *F*. However, *M* does not cause *O*. While there can be small variations in atmospheric oxygen concentration, these fluctuations are due to independent and identically distributed latent causes, whose mixture results in a normally distributed error term. Since *M* makes no difference to *O*, *O* is not located on a path from *M* to *F*. Rather, *O* is on an alternative path to *F* forming a collider structure $M \longrightarrow F \longleftarrow O$. Craver et al.'s insistence that standing conditions can be constituents indicates that they might not interpret causal betweenness in terms of either of the two interpretations considered so far. Instead, what they could mean by Φ being causally between Ψ_{in} and Ψ_{out} is that there is path $\Psi_{in} \longrightarrow \Psi_{out}$ (realizable by a process) that either contains Φ or that can be *modulated* by Φ.[10]

Plainly, there also exist tests that are both sufficient and necessary for collider structures. In fact, colliders empirically manifest in such a unique way that they are among the easiest structures to empirically detect. Their characteristic feature is that conditioning on a variable, such as the bike's movement Ψ_{out}, introduces a probabilistic dependence between otherwise unconditionally independent variables, such as the pedaling Ψ_{in} and the bracket Φ (Spirtes et al. 2000: 85). Hence, combining a standard test for paths or active routes with a standard collider test would yield a straightforward epistemic criterion for constitution under this "modulation" interpretation of causal betweenness. If that is really Craver et al.'s intended interpretation, it is thus again mysterious why they propose a new criterion of their own.

What is worse, the modulation interpretation seems to be conceptually incoherent to begin with. In order for Φ to be meaningfully referred to as being "causally between Ψ_{in} and Ψ_{out}" it is minimally required that Φ bears some causal relation to both Ψ_{in} and Ψ_{out}. Yet, if Φ is a standing condition, and thus insensitive to changes in Ψ_{in}, that requirement is not satisfied. Being exogenous to a modeled causal system is incompatible with being causally between any two elements of that system.

In support of their analysis of the biking phenomenon, Craver et al. argue that the bracket "is causally between Ψ_{in} and Ψ_{out}, because it is part of an event between the pedaling and the moving" (8824), where the complete event

10. This interpretation finds support in other passages from Craver (2007), for instance: "[Neuroscientists] say that they discover *systems* and *pathways* in the flow of information, and molecular *cascades*, *mediators*, and *modulators*. The term mechanism could do the same work" (3; emphases in original).

arguably involves not only the spindle rotating but also the bracket interacting with the spindle (8812). However, while the bracket being part of an event occurring between pedaling and moving may be said to place the bracket (spatio-) temporally between pedaling and moving, it does not place it *causally* between pedaling and moving. In the same way, the friction of the bike's tires or the weight of its frame are involved in events that occur between pedaling and moving and that influence the moving by modulating the response of the spindle to the pedaling, but they are not caused by pedaling. The bracket, the friction, and the weight feature in standing conditions that are insensitive to changes in Ψ_{in}, they are exogenous, they are not on a causal process from Ψ_{in} to Ψ_{out}, meaning they are not causally between Ψ_{in} and Ψ_{out}.

Craver et al. will insist that standing conditions are involved in events that are between Ψ_{in} and Ψ_{out} in a sense that is more substantive than just spatio-temporal betweenness. These events can be said to belong to the substrate, or the realizer, or the organization that grounds the causal process from Ψ_{in} to Ψ_{out} —perhaps in the way the brain is the substrate for, or the realizer of, mental states, or a legal principle grounds the ruling on a crime.[11] However, reconstructing standing conditions along these lines amounts to removing them from the causal process connecting Ψ_{in} and Ψ_{out} entirely. Substrates, realizers, or grounds of a causal process do not themselves figure as causes or effects in that process. Hence, even though there may be many other substantive notions of betweenness, these are all *non-causal* notions, meaning that none of them can reconcile the modulation interpretation with Craver et al.'s metaphysical thesis that constitution is *causal* betweenness. What we are thus left with is an incoherent interpretation of that thesis allowing for constituents that are not causally between a phenomenon's inputs and outputs.

To maintain the modulation interpretation that incoherence must be resolved by weakening the metaphysical thesis: being a constituent is *either* being causally between the phenomenon's input and output conditions *or* being a standing condition of that causal process. That is, there are two kinds of truth-makers of claims on constitutive relevance, namely paths or processes, which are causally between, and modulators (i.e., standing conditions), which are not. Obviously though, such a weakening leads to an explosion in the number of constituents for any phenomenon. The output of any input-output relation on earth has countless standing conditions in its spatio-temporal reach—gravity, electromagnetic and nuclear forces, moderate temperatures, presence of oxygen, etc. All of these become constituents of any phenomenon if we adopt the weakening of the

11. This reading aligns with other statements of Craver (2007), such as: "[…] the brain is composed of mechanisms. […] Neuroscientists sometimes use other terms to describe their explanatory achievements. They say that they are searching for the neural *bases*, the *realizers*, and the *substrates* of a phenomenon." (2-3; emphases in original).

metaphysical thesis. That is highly counterintuitive, and such a massive overgeneration borders on trivializing the notion of constitution.

To avoid that consequence, the eligibility of standing conditions as constituents must be suitably constrained. What comes naturally to mind here is the approach taken by MM, which constrains constituents to spatio-temporal proper parts of the upper-level entity *S* performing the activity Ψ-ing. For the biking phenomenon, this approach would entail that only standing conditions that are proper parts of the bicycle are candidate constituents. That is, the bracket would qualify but gravity or temperature would not. But plainly, that approach is not available to Craver et al.'s new theory because it explicitly reconceptualizes a phenomenon as an input-output relation *without* an upper-level acting entity *S*. As we have seen in §3, Craver et al. (2021: 8812) contend that MM's reliance on upper-level entities *S* is the ultimate source of MM's problems, which their new theory is designed to avoid. And they consider the fact that spatio-temporal parthood is no longer necessary for MIE "a side-benefit of our revision" (8822, fn 15).

But alternative constraints are difficult to come by. Requiring standing conditions eligible for constitution to be involved in events occurring temporally between Ψ_{in} and Ψ_{out} is not restrictive enough. In the biking phenomenon, not only gravity and temperature are standing conditions modulating events occurring temporally between pedaling and moving, but so are rain, wind, or road friction. It seems that standing conditions eligible for constitution must also be properly related spatially to Ψ_{in} and Ψ_{out}. But it is utterly unclear how to render that precise. What makes the bike's bracket properly spatio-temporally related to pedaling and moving, whereas gravity or road friction are not? The only generalizable answer that comes to mind here is the one unavailable to MIE, *viz.* that the bracket is a proper part of the acting entity *S* of the biking phenomenon, the bike, while gravity and road friction are not.

Hence, the modulation interpretation faces a dilemma: either all standing conditions modulating events occurring temporally between Ψ_{in} and Ψ_{out} are declared eligible for constitution or phenomena are, again, conceptualized as featuring acting entities *S* and standing conditions eligible for constitution are constrained to proper parts of *S*. The first horn runs the danger of trivializing the notion of constitution, while the second reintroduces the very source of the problems into the new theory that this theory sought to avoid in the first place.

6. No De-Coupling

We end this paper with a positive outlook on the prospects of meeting MM's original objectives. The severe problems encountered by MM and MIE are not

to be taken as evidence that it would be impossible to render constitution as a distinctly non-causal relation in a way that demarcates between constituents and non-constituents without falling into conceptual traps. In fact, there exists a theory of constitution that accomplishes exactly that, *viz.* the *No De-Coupling theory* (NDC; Baumgartner & Casini 2017).[12] What is more, NDC has been argued to allow for more faithful reconstructions of actual scientific practice than MM (van Eck 2019; Serban & Holm 2020). This section thus reviews the basics of NDC and shows how it meets MM's objectives.

The main problem of MM arises from its attempt to define "Φ is constitutively relevant to Ψ" in terms of the possible existence of an ideal intervention on Ψ that changes Φ. As phenomena and constituents spatio-temporally overlap and, thus, are assumed not to be causally related, all causes (i.e. intervention candidates) Z_i of Ψ that change Φ cause Ψ and Φ on two different directed paths $\langle Z_i, \Psi\rangle$ and $\langle Z_i, \Phi\rangle$, where a directed path simply is an ordered n-tuple of variables (Spirtes et al. 2000: 7). That $\langle Z_i, \Psi\rangle$ and $\langle Z_i, \Phi\rangle$ are different paths follows from the fact that the n-tuples are different because their second elements, Ψ and Φ, are non-identical. A cause Z_i with two effects Ψ and Φ on two different paths is a *common cause* of Ψ and Φ (Spirtes et al. 2000: 22).[13] Since Z_i is just a placeholder for any cause of Ψ that changes Φ, it follows that all such causes are common causes of Ψ and Φ. But common causes are not ideal interventions. Therefore, every cause of Ψ that changes Φ fails to be an ideal intervention, meaning there cannot be any ideal interventions as required by MM (Romero 2015; Baumgartner & Gebharter 2016).

This predicament of MM is the starting point of NDC. According to NDC, the characteristic feature of constitution is not the possibility to manipulate phenomena and constituents by ideal interventions but the *impossibility* to do so. An upper-level phenomenon Ψ is realized on a lower level by a set of constituents $\mathbf{\Phi} = \{\Phi_1, \ldots, \Phi_n\}$. If, as is usual, Ψ is assumed to non-reductively supervene on $\mathbf{\Phi}$ (e.g. Glennan 1996; Eronen 2012), it follows that every change induced on

12. There might exist other theories that meet MM's objectives without conceptual problems. For instance, Krickel (2018b) contends that her *causation-based CR* theory accomplishes the same, or Harbecke's (2010) regularity theory of constitution could be another contender. But as both of these proposals analyze constitution in close analogy to causation, we suspect—without having the space to discuss the issue in detail here—that they do not yield distinctly non-causal notions of constitution. In any case, Craver et al. (2021) do not engage with NDC or any of these other theories that have been claimed to meet MM's objectives while avoiding its problems. It is thus not clear why they believe that their new theory is needed in the first place.

13. Woodward (2022) rejects this terminology. He contends that both effects of a common cause must be independently fixable, which is violated by $\langle Z_i, \Psi\rangle$ and $\langle Z_i, \Phi\rangle$ because Ψ and Φ are related by supervenience. However, nothing in the argument presented here depends on the label "common cause"; all that matters is that the paths $\langle Z_i, \Psi\rangle$ and $\langle Z_i, \Phi\rangle$ are different (because Ψ and Φ are non-identical). Hence, a reader who shares Woodward's terminological intuition can just replace the term "common cause" by "two-path cause" (or whatever other term appears suitable) in our discussion.

Ψ is necessarily associated with a change in some $\Phi_i \in \mathbf{\Phi}$. As phenomena and their constituents are not causally related, every cause inducing a change in Ψ is necessarily causing a change in some $\Phi_i \in \mathbf{\Phi}$ on a different path. That is, every cause of Ψ necessarily is a common cause of Ψ and at least one $\Phi_i \in \mathbf{\Phi}$, meaning a phenomenon and the set of its constituents are *unbreakably coupled via common causes*. According to NDC, this is the defining feature of constitution.

To flesh this basic idea out into a full-blown theory of constitution, two further constraints and an operationalization of the notion of unbreakability are needed. First, a parthood constraint must be imposed: not all sets of variables that are coupled via common causes with the phenomenon Ψ are constituents but only those that contain spatio-temporal parts of Ψ. Second, a minimality constraint is required ensuring that $\mathbf{\Phi}$ does not contain redundant variables, that is, variables that can be removed from $\mathbf{\Phi}$ such that the remaining set is still coupled via common causes with the phenomenon. Third, unbreakability can be operationalized in terms of the phenomenon Ψ and the elements of $\mathbf{\Phi}$ remaining coupled via common causes across all expansions of the set of analyzed variables $\mathbf{V}$. In other words, no matter what additional causes of Ψ are integrated into the analysis, all of these added causes of Ψ are common causes of Ψ and at least one $\Phi_i \in \mathbf{\Phi}$—no surgical cause of Ψ can be found by expanding $\mathbf{V}$.

If we let a *complex instance* of a variable set $\mathbf{\Phi}$ designate the occurrence or process (in a particular spatio-temporal region) represented by a complex value assignment $\Phi_1 = \phi_1, \ldots, \Phi_n = \phi_n$ to all elements of $\mathbf{\Phi}$, the No De-Coupling theory of constitution can be stated as follows (Baumgartner & Casini 2017):

(NDC) Φ_1 is constitutively relevant to Ψ if, and only if, there exists a variable set $\mathbf{V}$ containing Ψ and a proper subset $\mathbf{\Phi} = \{\Phi_1, \ldots, \Phi_n\}$, such that the following conditions hold:

> *(Parthood)* For every complex instance of $\mathbf{\Phi}$, there is an instance of Ψ such that the former is a spatio-temporal part of the latter.
>
> *(Coupling)* Every cause of Ψ in $\mathbf{V}$ is a common cause of Ψ and at least one Φ_i in $\mathbf{\Phi}$.
>
> *(Minimality)* There does not exist a $\Phi_k \in \mathbf{\Phi}$ such that $\mathbf{\Phi} \setminus \{\Phi_k\}$ satisfies (Coupling).
>
> *(No De-Coupling)* The Coupling of $\mathbf{\Phi}$ and Ψ subsists in all expansions of $\mathbf{V}$.

Less formally, NDC defines a constituent to be a member of a minimal set of spatio-temporal parts of the phenomenon that is unbreakably common-cause coupled with the phenomenon.

NDC has various features setting it apart from other theories of constitution, but most of them are not relevant for our current purposes (for details see Baumgartner & Casini 2017; van Eck 2019; Serban & Holm 2020). What matters here are only those of its characteristics that need to be understood in order to see (I) that NDC does in fact establish constitution as distinctly non-causal dependence relation, (II) that it demarcates between constituents and non-constituents, and (III) that it is conceptually sound. Let us hence take these three points in turn.

(I) We begin with the non-causal nature of constitution. NDC defines constitution based on (Parthood), which requires phenomena and constituents to be mereologically related. As causation is commonly assumed to obtain among mereologically unrelated entities only, (Parthood) entails that no Φ that NDC identifies as constituent of a phenomenon Ψ can be an ordinary cause or effect of Ψ. However, despite being widely assumed, only few theories of causation contain explicit clauses stipulating that causes and effects must not spatio-temporally overlap. Some authors even allow for abnormal forms of causal dependence that obtain among overlapping entities (e.g. Leuridan 2012). Hence, (Parthood) does provide some, albeit not conclusive, support for the non-causal nature of NDC-defined constitution.

Further support is offered by the fact that NDC defines constitution in terms of *all* causes of the phenomenon, across *all* variable set expansions, not being interventions but common causes. That is, constitution is a universally defined relation, or alternatively, it is defined based on the non-existence of possible interventions. By contrast, theories of causation analyze causation in terms of the *existence* of $\mathcal{P}$, where $\mathcal{P}$, depending the theoretical framework, stands for possible interventions, for probabilistic or counterfactual difference-making scenarios, or for processes or energy transfer, etc. That is, causation is routinely rendered as existentially defined relation. This, in turn, means that the negation of causation—causal irrelevance—becomes universally defined. More specifically, if "X is causally relevant to Y" is defined via $\exists x\mathcal{P}x$, then, by contraposition, "X is causally irrelevant to Y" is rendered in terms of $\neg\exists x\mathcal{P}x$, *viz.* in terms of $\forall x\neg\mathcal{P}x$. For example, Woodward's (2003) interventionist theory of causation cashes out "X is causally irrelevant to Y" based on the non-existence of ideal interventions on X with respect to Y. This is exactly how NDC defines "Φ is constitutively relevant to Ψ." Of course, NDC does not yield a notion of constitution that is co-extensional with causal irrelevance. After all, constitution is a robust dependence relation, whereas the extension of causal irrelevance encompasses many independent entities. Still, the extension of NDC-constitution is a proper subset of the extension of the notion of causal irrelevance (as commonly defined).

In addition, Baumgartner and Casini (2017) characterize NDC as an *abductive* theory because it reconstructs an inference to constitution as an inference to the best explanation. A phenomenon and its constituents have highly correlated behavior patterns. That strong correlation can be accounted for by the mere fact that phenomena and constituents are coupled via common causes. That is, pure causal models, which do not contain any constitutive dependencies, can faithfully reproduce the empirical data generated by mechanistic systems. But pure causal models have an important explanatory gap: they cannot explain why the coupling of phenomena and constituents cannot be broken. NDC supplies a relevance relation, different from causation, that fills precisely this explanatory gap. Models introducing NDC-constitution not only account for the empirical data relative to some given set of analyzed variables but they also explain why upper- and lower-level variables remain coupled across all expansions of analyzed variable sets.

(II) The next point is demarcation. That NDC furnishes a criterion demarcating constituents and non-constituents follows from the fact that its definiens is both sufficient and necessary for constitutive relevance. Φ is constitutively relevant to Ψ if, and only if, there exists a variable set $\mathbf{V}$ containing Ψ and a subset $\mathbf{\Phi}$, such that $\Phi \in \mathbf{\Phi}$ and NDC's four conditions are satisfied. In other words, containment in a set complying with (Parthood), (Coupling), (Minimality), and (No De-Coupling) demarcates between constituents and non-constituents. According to this criterion, for example, both kidneys are constituents of the osmolality phenomenon because both of them are contained in some minimal set $\mathbf{\Phi}$—possibly not both in the same set—of unbreakably common-cause coupled parts of the phenomenon. By contrast, if we assume with Craver et al. (2021: 8824) that a bike's bracket is a standing condition, NDC rules that it is not a constituent of the bike's movement. As the bracket and the movement are unconditionally independent, they do not share common causes.

Of course, whether NDC's demarcation criterion is satisfied by some $\Phi \in \mathbf{\Phi}$ and a phenomenon Ψ can only be decided inductively. Establishing (No De-Coupling) calls for progressive expansions of analyzed variable sets by additional causes of Ψ and continuous testing whether these added causes break the common-cause coupling of $\mathbf{\Phi}$ and Ψ. After a finite number of expansions and rigorous but unsuccessful attempts at breaking this coupling, the satisfaction of (No De-Coupling) can be inductively inferred. This, in turn, licenses an abductive inference to every element of $\mathbf{\Phi}$ being a constituent of Ψ. By contrast, if variable set expansions reveal a surgical cause of Ψ that does not target any element of $\mathbf{\Phi}$, the common-cause coupling of $\mathbf{\Phi}$ and Ψ is falsified. That, however, does not entail that a particular $\Phi \in \mathbf{\Phi}$ might not be contained in *another* set $\mathbf{\Phi'}$ for which common-cause coupling cannot be broken. Hence, that a particular Φ is a

non-constituent can likewise only be inductively corroborated, namely by means of an extended unsuccessful search for a constituting set comprising Φ. But its inductive nature notwithstanding, NDC furnishes a clear criterion demarcating between constituents and non-constituents.

(III) Finally, that NDC faces none of the conceptual problems of MM and MIE is apparent from the mere fact that neither the problematic notion of an ideal intervention nor that of causal betweenness play any role in NDC. Instead, the core notion based on which NDC defines constitution is that of a common cause, which is standard, unambiguous, and readily applicable to mechanistic systems. The only notion appearing in NDC that is not straightforward is that of spatio-temporal parthood; it raises certain logical and metaphysical questions commonly addressed in the field of mereology. These issues, however, are sidestepped in the literature on mechanistic explanation. Mechanists simply assume a metaphysically thin notion of parthood according to which x is a part of y if, and only if, x occupies a spacetime region that is contained in the region occupied by y, and they take clarity on such parthood relations among analyzed behaviors for granted. Craver (2007) avails himself to that background assumption for MM and so do Baumgartner and Casini (2017) for NDC.

Overall, barring mereological questions concerning parthood, this section has shown that NDC indeed establishes constitution as a distinctly non-causal explanatory relation, that it provides a criterion demarcating between constituents and non-constituents, and that it accomplishes this without falling into conceptual traps.

7. Conclusion

To solve the problems of Craver's (2007) Mutual Manipulability theory (MM), Craver et al. (2021) presented a new theory of constitution consisting of the metaphysical thesis that constitution is causal betweenness and the epistemic condition MIE, which the authors claim to be sufficient (but not necessary) to identify constituents. The main part of our paper argued that this new theory, contrary to the assertions of Craver et al., neither retains the spirit of MM nor is free of conceptual confusion.

On the one hand, contra MM, the new theory renders constitutive explanation as a form of causal explanation and it reduces the problem of finding constituents to the problem of finding the causes or the processes between a phenomenon's input and output conditions. There exist many well-tested methodological approaches to solve this problem, but MIE attempts to develop its

own solution from scratch. Moreover, whereas MM is meant to rule out, for example, that hemodynamic changes are constituents of cognitive task performance, MIE remains entirely silent about non-constituents. In consequence, MIE does not provide a basis for distinguishing between good and bad constitutive explanations.

On the other hand, the new theory's metaphysical component is inherently ambiguous. We discussed three viable interpretations of what Craver et al. could mean by constituents being causally between input and outcome conditions of phenomena: the path, the process, and the modulation interpretation. We found that all of them are defective and none of them combines coherently with MIE. Relative to the path interpretation, MIE is indeed sufficient to identify constituents but contains two redundant conditions that serve no purpose whatsoever. By contrast, relative to the process interpretation, MIE is not sufficient for identifying constituents, contrary to what Craver et al. claim. Finally, the modulation interpretation either trivializes the notion of constitution, by allowing too many standing conditions to count as constituents, or it is forced to reintroduce defining criteria for acting upper-level entities, which—according to Craver et al.—are the very source of the problem that the new theory set out to avoid. We conclude that the new theory is not free of conceptual confusion either.

Since the flaws of MM have become a discussion point in the literature, different theories of constitution have been proposed, some of which explicitly tailored to meet the same objectives as MM. Unfortunately, Craver et al. (2021) do not engage with these alternative theories. It hence remains unclear which aspects of the available alternatives Craver et al. deem unsuitable and why they believe that yet another theory of constitution would be needed in the first place. To substantiate that there is no actual need for a new theory if the goal is to meet MM's objectives, we reviewed the basics of the No De-Coupling theory (Baumgartner & Casini 2017) and demonstrated its ability to deliver on its promises. By defining constitution in terms of unbreakable common-cause coupling, NDC provides a distinctly non-causal relevance relation grounding a hierarchical form of explanation, which fills an important explanatory gap of pure causal models. Contrary to the latter, constitutive models are capable of explaining why behavior patterns of phenomena and their constituents cannot be de-coupled. This validates a core tenet, not only of Craver's account of neuroscientific explanation, but of the new mechanist program more generally: constitution is a standalone explanatory dimension on a par with causation. A mechanistic explanation of a phenomenon is good if, and only if, it describes the hierarchical arrangement and causal interactions of all and only its constitutively relevant component entities and activities.

Acknowledgements

We thank two anonymous reviewers for helpful comments on an earlier draft. Moreover, we are grateful to the Trond Mohn Foundation (grant 811886 for M.B.) and the Italian Ministry of University and Research (grants 201743F9YE, 20177FX2A7, and the PRO3 grant for the project "Understanding public data: experts, decisions, epistemic values" for L.C.) for generous support of this research.

References

Baumgartner, Michael and Lorenzo Casini (2017). An Abductive Theory of Constitution. *Philosophy of Science*, *84*(2), 214–233.

Baumgartner, Michael and Alexander Gebharter (2016). Constitutive Relevance, Mutual Manipulability, and Fat-Handedness. *The British Journal for the Philosophy of Science*, *67*(3), 731–756.

Bechtel, William and Adele Abrahamsen (2005). Explanation: A Mechanist Alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*(2), 421–441. https://doi.org/10.1016/j.shpsc.2005.03.010

Craver, Carl (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press.

Craver, Carl, Stuart Glennan and Mark Povich (2021). Constitutive Relevance & Mutual Manipulability Revisited. *Synthese*, *199*(3), 8807–8828.

Dowe, Phil (2000). *Physical Causation*. Cambridge University Press.

Eronen, Markus (2012). Pluralistic Physicalism and the Causal Exclusion Argument. *European Journal for the Philosophy of Science*, *2*, 219–232.

Gillett, Alexander J., Christopher J. Whyte, Christopher L. Hewitson, and David M. Kaplan (2022). Defending the Use of the Mutual Manipulability Criterion in the Extended Cognition Debate. *Frontiers in Psychology*, *13*, 1043747. https://doi.org/10.3389/fpsyg.2022.1043747

Glennan, Stuart (1996). Mechanisms and the Nature of Causation. *Erkenntnis*, *44*(1), 49–71.

Glennan, Stuart (2002). Rethinking Mechanistic Explanation. *Proceedings of the Philosophy of Science Association*, *69*(3), 342–353.

Glennan, Stuart (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), 267–285.

Harinen, Totte (2018). Mutual Manipulability and Causal Inbetweenness. *Synthese*, *195*, 35–54. https://doi.org/10.1007/s11229-014-0564-5

Hitchcock, Christopher (2001). The Intransitivity of Causation Revealed in Equations and Graphs. *Journal of Philosophy*, *98*, 273–299.

Krickel, Beate (2018a). *The Mechanical World: The Metaphysical Commitments of the New Mechanistic Approach*. Springer International Publishing.

Krickel, Beate (2018b). Saving the Mutual Manipulability Account of Constitutive Relevance. *Studies in History and Philosophy of Science Part A*, *68*, 58–67. https://doi.org/10.1016/j.shpsa.2018.01.003

Krickel, Beate, Leon de Bruin, and Linda Douw (2023). How and When are Topological Explanations Complete Mechanistic Explanations? The Case of Multilayer Network Models. Synthese, 202, 14. https://link.springer.com/article/10.1007/s11229-023-04241-z

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.

Lewis, David (1986). Causal Explanation. In David Lewis (Ed.), *Philosophical Papers* (Vol. 2, 214–240). Oxford University Press.

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

Neapolitan, Richard E. (2004). *Learning Bayesian Networks*. Pearson Prentice Hall.

Parise, André G., Gabriela F. Gubert, Steve Whalan, and Monica Gagliano (2023). Ariadne's Thread and the Extension of Cognition: A Common but Overlooked Phenomenon in Nature? *Frontiers in Ecology and Evolution*, *10*, 1069349. https://doi.org/10.3389/fevo.2022.1069349

Pearl, Judea (2009). *Causality. Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press.

Romero, Felipe (2015). Why There Isn't Inter-Level Causation in Mechanisms. *Synthese*, *192*(11), 3731–3755.

Russell, Bertrand (1948). *Human Knowledge*. Simon and Schuster.

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

Serban, Maria and Sune Holm (2020). Constitutive Relevance in Interlevel Experiments. *The British Journal for the Philosophy of Science*, *71*(2), 697–725.

Smart, Paul (2022). Minds in the Metaverse: Extended Cognition Meets Mixed Reality. *Philosophy & Technology*, *35*(4), 87. https://link.springer.com/article/10.1007/s13347-022-00580-w

Spirtes, Peter, Clark Glymour, and Richard Scheines (2000). *Causation, Prediction, and Search* (2nd ed.). MIT Press.

Steel, Daniel (2007). *Across the Boundaries: Extrapolation in Biology and Social Science*. Oxford University Press.

van Eck, Dingmar (2019). Constitutive Relevance in Cognitive Science: The Case of Eye Movements and Cognitive Mechanisms. *Studies in History and Philosophy of Science Part A*, *73*, 44–53.

Vanderweele, Tyler J., Stijn Vansteelandt, and James M. Robins (2014). Effect Decomposition in the Presence of an Exposure-Induced Mediator-Outcome Confounder. *Epidemiology*, *25*(2), 300–306.

Varga, Somogy (2023). Understanding in Medicine. *Erkenntnis*, *89*, 3025–3049. https://doi.org/10.1007/s10670-023-00665-8

Weber, Marcel (2022). *Philosophy of Developmental Biology*. Cambridge University Press.

Weinberger, Naftali (2019). Mechanisms without Mechanistic Explanation. *Synthese*, *196*(6), 2323–2340. https://link.springer.com/article/10.1007/s11229-017-1538-1

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

Woodward, James (2022). Modeling Interventions in Multi-Level Causal Systems: Supervenience, Exclusion and Underdetermination. *European Journal for Philosophy of Science*, *12*(4), 1–34.