Synthese
DOI 10.1007/s11229-017-1538-1

Mechanisms without mechanistic explanation

Naftali Weinberger[1]

Received: 9 February 2017 / Accepted: 16 August 2017

Abstract Some recent accounts of constitutive relevance have identified mechanism components with entities that are causal intermediaries between the input and output of a mechanism. I argue that on such accounts there is no distinctive inter-level form of mechanistic explanation and that this highlights an absence in the literature of a compelling argument that there are such explanations. Nevertheless, the entities that these accounts call 'components' do play an explanatory role. Studying causal intermediaries linking variables X and Y provides knowledge of the counterfactual conditions under which X will continue to bring about Y. This explanatory role does not depend on whether intermediate variables count as components. The question of whether there are distinctively mechanistic explanations remains open.

Keywords Mechanisms · Causation · Explanation · Causal mediation · Extrapolation · Constitution

1 Introduction

In discovering the double-helical structure of DNA, Watson and Crick advanced our understanding of how traits are inherited across generations. Yet one hesitates to claim that DNA molecules *cause* heredity. Rather, these molecules are part of the process by which traits are passed on. More formally, heredity is *constituted* by the replication and transmission of DNA across generations. The concept of constitution plays a key role in recent accounts of mechanistic explanation. Many success stories in the life

✉ Naftali Weinberger
Naftali.weinberger@gmail.com

[1] Tilburg Center for Logic, Ethics and Philosophy of Science, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands

Published online: 04 September 2017

sciences involve showing how the components of a mechanism interact to produce a phenomenon, and the relationship between the mechanism's activities and those of its components is a constitutive one. This fact alone may seem to justify the view that mechanistic phenomena are *explained* by the organized activity of a mechanism's components and that the relevant form of explanation is different from traditional causal explanations. In what follows, I argue that participants in debates over mechanistic explanation have been too quick to grant that there are such explanations. Analysis of existing accounts reveals either that mechanisms are not distinctively explained by the organized activity of their components, or that we still lack an account of this distinctive form of explanation.

The locus of my discussion is Craver's account of the constitutive relevance relation, as well as recent interpretations of his account on which this relation is a causal one [I focus on Harinen (2014)]. I argue that such accounts unintentionally render constitutive relevance to be identical to ordinary (within-level) causal relevance, and that this result is incompatible with treating constitutive relevance as a distinct explanatory relation. Nevertheless, we can make sense of the explanatory contribution of the entities that these accounts call 'components'. On these accounts, components are variables that are causally between the input and output of a mechanism. I show how, in general, studying a variable that is causally between two other variables X and Y helps one predict whether X would still cause Y under counterfactual circumstances. Accordingly, the 'components' in these accounts do play an explanatory role as intermediaries, but not because they are 'components'. If there is a way that the organized activity of components distinctively contributes to explaining a phenomenon, one is hard-pressed to find it in existing accounts.

This paper is organized as follows. Section 2 provides background regarding mechanistic explanation and argues that on Harinen's interpretation of Craver, constitutive relevance just is causal relevance. Section 3 presents a more general argument for the conclusion that there are no distinctively mechanistic explanations. Section 4 presents a non-mechanistic account of the explanatory role of intermediate variables. Section 5 concludes.

2 Mechanisms and constitutive relevance

In "Thinking about Mechanisms," Machamer, Darden and Craver define mechanisms as:

> [E]ntities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions. (2000, p. 3)

An example of a mechanism is a neuron's firing. When a neuron fires, it increases and then decreases in voltage. These voltage changes result from sodium and potassium ions moving across the cell membrane, thereby changing the proportions of sodium and potassium inside and outside the cell. To explain this process, one must identify the properties of the ion-channels–the entities that regulate the movement of ions across the membrane–and determine how they perform their functions (their "activities").

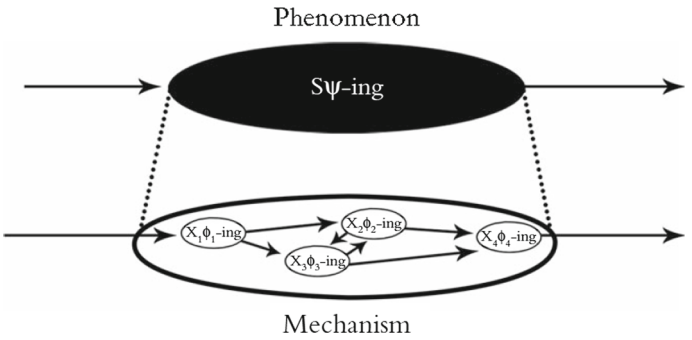

Fig. 1 (From Craver and Bechtel 2007, p. 7)

The entities that are organized to bring about the activity of a mechanism are its components, and one mechanism can be a component in a larger one. This suggests a hierarchical ordering of the world in which a whole mechanism counts as one level and its components are at a lower level. The concept of a mechanism level is distinct from other level-concepts in the literature such as levels of size (macro/micro), levels of abstraction and levels of properties (first-order/second-order etc.).

Craver and Bechtel (2007) argue that the relationship between mechanism levels is non-causal on the grounds that causes and effects must be spatially and temporally distinct. Since components and mechanisms stand in a part/whole relation, they cannot be causally related. The relationship between a mechanism and its components is *constitutive* rather than causal. An account of constitutive relevance specifies how an entity must contribute to the activity of a mechanism in order to count as a component.

Craver (2007) provides an account of mechanistic explanations in neuroscience. He refers to the mechanism as S and its components as X_1, $X_2 \ldots X_n$. S's activity is denoted by Ψ ("psi") and the activities of X's are denoted by φ_1, $\varphi_2 \ldots$ φ_n("phi-1" etc). A neuron firing is an S that Ψs. A sodium-ion gate opening is an X that φ. s. In Fig. 1, the within-level relationship among the φ-ing X's are causal and the inter-level relationship between a φ. ing X's and the Ψ-ing S is constitutive.

On Craver's account, an entity is a component in a mechanism if it is both a part of the mechanism and it is possible to change the behavior of the mechanism by changing that of the component, and vice versa. More precisely, X is constitutively relevant to S if (1) X is a part of S and (2) X's φ-ing and S's Ψ-ing meet the following conditions:

> (CR1) When φ is set to the value φ_1 in an ideal intervention, then ψ takes on the value f(φ_1). (155)
> (CR2): if Ψ is set to the value Ψ_1 in an ideal intervention, then φ takes on the value f(Ψ_1). (159)

f(φ_1) and f(Ψ_1) refer, of course, to different functions. CR1 and CR2 rely on Woodward's (2003) notion of an ideal intervention. In evaluating the effect of X on Y, an ideal intervention determines the value of X in such a way that it no longer depends on its direct causes (other than the intervention). Additionally, such interventions do not influence Y via variables on causal paths not going through X. On Woodward's account Xcauses Y if and only if one can change the value of Yvia an ideal intervention

on X.While the concept of an ideal intervention was developed for explicating causal relevance, Craver uses this notion to explicate constitutive relevance as well.

To illustrate Craver's account, consider his discussion (2002) of how scientists discovered the role of the hippocampus in spatial memory. When rats run a maze, their hippocampi are activated. Here Ψ is the process of running the maze. Intervening to make the rat run the maze causes a change in the activity of the hippocampus (a φ-ing X), fulfilling CR2. Yet learning that the hippocampus activates when the rat navigates the maze is insufficient for establishing that the hippocampus is a component in spatial memory–it is possible that its activity is a side effect of the running that plays no role in navigation. To rule out this possibility, one must intervene to either stimulate or disable the hippocampus and see if doing so changes the rat's maze-running ability. If so, then CR1 is fulfilled as well. It is through such experiments that scientists discovered the role of the hippocampus in spatial memory.

The subsequent discussion will be centered on Craver's account and the responses it has generated. I will focus on questions related to mechanistic explanation–in particular on whether there is an inter-level form of explanation–and Craver's book contains the best-developed inter-level account. Yet much of the mechanistic literature does not emphasize explanation. Levy (2013) helpfully distinguishes explanatory theses about mechanisms from those related to the metaphysics of causation and those related to discovery. As an example of the former, Glennan (1996) argues that mechanisms are ontologically more basic than causes (though he has subsequently revised his view). The literature on mechanistic discovery was spearheaded by Bechtel and Richardson (1993/2010) and is alive and well in works such as Craver and Darden (2013). This literature focuses on the role of mechanisms in strategies for discovery. Since I am concerned here with explanation, I focus on Craver's account and reference others as needed.

In recent years there have been many criticisms of Craver's account. These have become increasingly sophisticated, though the source of the problems they raise is simple: talk of interventions on the Ψ variable is ambiguous (Menzies 2012; Franklin-Hall 2016). Is an intervention on Ψ an intervention that triggers the mechanism into action, or an intervention on something that happens at a later stage in the activity of the mechanism? These are distinct interventions. Recall that an intervention *determines* the value of the variable that one intervenes upon. But ensuring that a rat begins a maze (and observing which components are subsequently activated) does not ensure that the rat will complete the maze. Moreover, there is the further complication that since the X that φs is supposed to be a part of the S that Ψs, these two variables are not distinct. But the conditions for an ideal intervention require that the variable upon wch one intervenes be distinct from other variables in the model (Harinen 2014; Baumgartner and Gebharter 2015; Romero 2015).

Harinen (2014) proposes a way to resolve both problems simultaneously. He disambiguates Ψ into two variables, Ψ_{in} and Ψ_{out}, corresponding to the input and output of the mechanism[1] (cf. Menzies 2012). Since these two variables are distinct from φ–the

[1] Harinen might not characterize himself as disambiguating Ψ into Ψ_{in} and Ψ_{out}, as he considers the phenomenon to be the causal relationship between Ψ_{in} and Ψ_{out} and thus to be distinct from either. Nevertheless, I argue below that the variable Ψ plays no role in his account.

activity of a component that occurs between them–there may be ideal interventions on Ψ_{in} that change φ and ideal interventions on φ that change Ψ_{out}. These, in fact, are the very interventions corresponding to CR2 and CR1. He concludes that–contrary to appearances–constitutive relevance is causal.[2]

Harinen takes himself to be saving Craver's account, but he in fact trivializes it. Although he follows Craver in referring to φ-ing X's as "lower-level" and Ψ-ing S's as "higher-level", and draws metaphysical conclusions about the possibility of inter-level causation, there is nothing in Harinen's account that elucidates what it means to say that these variables are at different levels. It is true that the account still requires that X's φ-ing be a part of S's Ψ-ing, but φ is *not* a part of Ψ_{in} or Ψ_{out}. If there is no basis for saying that Ψ_{in} and Ψ_{out} are at different levels from φ, then CR2 and CR1 identify two within-level causal relationships, one between Ψ_{in} and φ. and another between φ and Ψ_{out}. To be clear, the problem is not so much that his account yields the counterintuitive result that constitutive relevance is causal. Rather, it is that the causal relationships that make up the constitutive relevance relationship are *exactly the same ones* as those that est between a mechanism's components.

One might object that Harinen does explain how to distinguish between levels, since he says that Ψ_{in} or Ψ_{out} *supervene* on the activities of particular components. Yet the claim that Ψ_{in} supervenes on some φ is not by itself very informative. It entails that it is not possible for there to be differences in Ψ_{in} without there being differences in φ . Although supervenience is often invoked as a feature of asymmetric dependence relationships, the relationship itself is not asymmetric. If Ψ_{in} were type-identical to some φ such that any change in the value of one were a change in the value of the one were a change in the value of the other, then, trivially, Ψ_{in} would supervene on φ. So supervenience by itself does not help us determine when one entity should be characterized as being at one level rather than another. Moreover, any attempt to save Craver's account by explaining how particular components can be described as being at either one level or another should itself strike us as strange. While there are plausible reasons for distinguishing between the behavior of a component and that of the whole mechanism, it is unclear why we would appeal to mechanism-levels to distinguish between two ways of describing the same localized components.

The absence of a basis for treating Ψ_{in} and Ψ_{out} as being at a different level from φ, why should we should treat his account as giving an explication of constitutive relevance, rather than showing us that talk of constitutive relevance may *replaced* by talk of causal relevance? I anticipate the response that although Harinen requires us refer to Ψ_{in}, and Ψ_{out} in describing the interventions that are needed to *discover* constitutive relevance relations, the constitutive relationship itself is between φ. and Ψ. But what is 'Ψ', then? If it is just the proposition that Ψ_{in} causes φ, which in turn causes Ψ_{out}, what do we gain by treating Ψ as a distinct variable? If not, then what further necessary conditions are there for X's φ-ing being a component of S's Ψ-ing?[3]

[2] See Leuridan (2011) for a distinct argument for this conclusion.

[3] It may seem strange to require Harinen (or Craver) an answer to the question of what 'Ψ' is. It is up to scientists to pick out phenomena of interest in a domain. Moreover, some mechanists hold a 'perspectival' view on which the decision to give a higher-level characterization of a phenomenon depends on scientists' interest in rendering the world intelligible (Craver 2013). Here I am asking specifically about the variable

My reason for emphasizing Harinen's interpretation of Craver is not that it is especially problematic, but rather that it is especially clear. Any explication of Craver needs to deal with the ambiguity in Ψ and the proposed solution provides the most straightforward fix. Yet once one does so, it is no longer clear that Craver provides an account of constitutive relevance that distinguishes it from within-level causal relevance. In the next section, I argue that this reveals a more general problem with mechanistic explanation.

3 Constitutive relevance without constitutive explanation

It is uncontroversial that there are entities scientists model as components in a mechanism, and that these components play various explanatory roles. Minimally, components stand in causal relations to other components. But I take it that mechanistic accounts of exanation are not merely committed to saying that components in mechanisms play various explanatory roles. Presumably, the motivation for providing mechanistic explanations is that mechanistic concepts matter for explanation, and enable one to present an account that is different from accounts not relying on mechanistic concepts. As shorthand for this idea, we can say that on such accounts mechanisms "distinctively explain". In this section I argue that mechanists have not shown that mechanisms distinctively explain. In fact, attention to existing accounts suggests that they do not.

Mechanists tend not to be explicit about whether they are committed to the claim that mechanisms distinctively explain. One widespread belief is that mechanistic phenomena cannot be explained using laws, but rather require a 'causal-mechanical' explanation (Bechtel and Abrahamsen 2005, Craver 2007). It is not generally clear if and how 'causal-mechanical' explanations differ from causal explanations more generally. The idea that law-based accounts should be rejected in favor of causal ones is not at all new. If all the mechanists were doing was emphasizing the pervasiveness of causal explanations across sciences that describe mechanisms, it is unclear what the project would be contributing to the topic of explanation, or why mechanistic explanation has received the amount of attention it has. Moreover, the broadly held thesis that mechanistic explanations are 'inter-level' seems important precisely because it helps to distinguish them from within-level explanations. In any event, should it turn out that the 'mechanical' in 'causal-mechanical' is not doing any conceptual work, this should at least be made clear.

Do mechanisms distinctively explain? My argument for skepticism begins with the premise that if there is a distinctive form of mechanistic explanation in Craver (2007) it is to be found in his account of constitutive relevance. His account of within-level causal relevance is just Woodward's interventionist account. His novel contribution

Footnote 3 continued
'Ψ' as it is defined within Harinen's formal account. The aim of doing so is to determine whether the alleged inter-level relationship plays any role in the account. As I argue in the following section, an account of explanation positing inter-level relationships should elucidate the advantages of modeling inter-level relationships. This requirement is reasonable whether talk of levels is grounded in our interests or in some objective feature of the world.

on the topic of explanation comes in his treatment of the between-level constitutive relevance relationship. Should it turn out that his account of constitutive relevance does not contribute to our understanding of explanation, it would not suffice for the mechanist to respond that at least Craver defines mechanistic *within*-level relevance. We might choose to refer to causal relationships as "mechanistic", but Craver has not provided any distinctive mechanistic account of causation.

To be clear, I am not presupposing that mechanists must offer a non-causal form of explanation in order for mechanisms to distinctively explain.[4] For instance, Leuridan (2011) claims that constitutive relevance is itself causal. Nevertheless–and this is the key point–he still assumes that constitutive relevance involves a different type of explanatory relationship than does within-level causal relevance. It is the assumption that constitutive relevance is distinct from within-level forms of relevance that I am questioning here, and that I will claim is necessary for inter-level explanation.

Craver's focus on inter-level explanation is characteristic of the broader literature. While mechanists [e.g. Craver and Darden (2013)] have emphasized the multiplicity of mechanism types, some of which might not be best represented in terms of levels, there exist few discussions (if any) of the forms of explanation relevant to these other mechanism types. Should there be no inter-level explanatory relationship, it is unclear what distinctive form of explanation the mechanists have to offer.

In evaluating whether mechanists have in fact presented an inter-level form of explanation, I focus exclusively on accounts of constitutive relevance. This requires some justification. One might suppose that the question of whether there is some inter-level explanatory relationship is independent from that of what counts as part of a mechanism. Perhaps simply observing the complex ways that a mechanism must be organized to produce a phenomenon itself justifies talk of levels. Yet, there are limits to how much one can separate constitutive relevance from the question of whether there is an inter-level form of explanation. If there is a form of inter-level explanation, then the nature of this explanation should be evident from accounts of what it means for two entities to stand to one another in this inter-level relationship.[5]

The claim that the nature of the inter-level relationship should be "evident" from an account of constitutive relevance is a bit vague. What matters for my argument is the following, more precise, claim: There are trivial ways of defining constitutive relevance such that, if the definitions provided were adequate, we would conclude that there is no form of inter-level explanation. Imagine a system containing three variables X, Y, and Z such that Xcauses $Y$and Ycauses Z.One could stipulate that we should refer to the

[4] While I do not presuppose that constitutive relevance is not causal, I see little basis for thinking it is. Existing arguments that constitutive relevance is causal derive this conclusion as a consequence of Craver's account, rather than from reconsidering the metaphysics of between-level causation.

[5] Note that I am not saying that the only way mechanists could justify talk of explanatory levels is with an account of constitutive relevance. Rather, I claim that an adequate account of constitutive relevance should enable one to see why it matters that two entities stand to one another in an inter-level relationship. This leaves open the possibility that a general account of how a mechanism's behavior depends on the organized activity of the components could by itself clarify which forms of organization should be described in terms of levels. While the literature on mechanistic discovery [e.g. Bechtel and Richardson, (1993/2010)] provides examples that could be utilized towards this project, the recent literature on mechanistic explanation has focused on constitutive relevance at the expense of organization.

causal chain $X \rightarrow Y \rightarrow Z$ as '**C**' and then describe the relationship between **C** and a particular variable on the chain such as Y. **C** does not cause Y and Y does not cause **C**, and Y is a part of **C**. So the relationship between **C** and Y is clearly distinct from the causal relationships between X,Y, and Z. But simply relabeling $X \rightarrow Y \rightarrow Z$ as **C** would not by itself bring into existence a new explanatory relationship. Of course, no mechanists are trying to define constitutive relevance relationships into existence by mere relabeling. But there is still a lesson to be learned from the toy example. Namely, part-whole relationships are cheap. The possibility of identifying two variables representing entities standing to one another in a part-whole relationship does little to resolve the substantive question of whether there is an inter-level form of explanation.

Finding explanatory part-whole relationships is straightforward enough. The weight of an object is explained by the weight of its parts. But such relations exist even in 'mere aggregates', which Craver distinguishes from functionally organized mechanisms (2007, p. 135). In specifying how a mechanism is explained by the organized activities of its parts, the fact that certain properties unrelated to the mechanism's function–e.g. the mechanism's weight in certain cases–depend in an aggregate way on the properties of its components is a distraction. What matters is whether the part-whole relationship plays a role in explaining how the components come together to produce the mechanistic phenomenon.

To see whether the alleged part-whole relationship matters for explanation, we need to make the explanatory relata precise. This is what Craver attempts to do in his account of constitutive relevance. But it remains unclear that one can fill in the account so that the explanatory relata do stand to one another as part and whole. Perhaps once one spells out the relata, it will emerge that mechanistic explanation takes place at a single level.

This is what happens in Harinen's account. As I have argued, Harinen lacks a basis for describing his variables Ψ_{in} and Ψ_{out} as being at a different level from φ. If one were to relabel the variables in his causal chain $\Psi_{\text{in}} \rightarrow \varphi \rightarrow \Psi_{\text{out}}$ as generic variables $X \rightarrow Y \rightarrow Z$ nothing of explanatory value would be lost from his account. One might insist that Ψ_{in} and Ψ_{out} should receive special labels because they correspond to the input and output of the mechanism that links them. But if all of the explanatory relationships in the account are at a single level, we won't get new explanatory relationships by relabeling. Moreover, recall that Harinen draws conclusions from his account regarding the possibility of inter-level causation. And certainly one cannot derive new *metaphysical* relationships by relabeling.[6]

One cannot evade the criticism that mechanists have not provided an inter-level form of explanation simply by asserting that mechanistic accounts have a different explanandum than standard causal accounts. That is, these accounts seek to explain

[6] The idea that debates about the relationships among mechanism-levels rest on deeper metaphysical issues is often supported by a misguided link to Kim's (1998) causal exclusion argument (Harinen 2014; Romero 2015). Kim argues second-order properties (such as mental properties, according to non-reductive physicalism) cannot have any causal powers over and above the first-order properties that realize them. As Craver (2007, pp. 197–8) correctly notes, Kim's distinction between higher- and lower-order properties is orthogonal to that between higher- and lower- mechanism-levels. Kim grants that wholes have causal properties lacked by their parts (pp. 80–87). His higher- and lower-order properties are properties of a single entity.

the mechanism as a whole in terms of its parts, rather than to explain an event in terms of its cause. But even if the target of mechanistic explanation differs from that of causal explanation, this does not show that mechanists have distinctively explained the target. If the phenomenon turns out to be a re-description of the within-level relationships, we should say they have not. This is true even though scientists do not set out to re-describe a mechanism in terms of its components and in fact characterize the mechanistic phenomena prior to learning about the mechanism's components. What matters here is not how scientists discover the entities responsible for a phenomenon, but whether the proposed inter-level relationship plays a role in explanation.

I have claimed that Harinen's difficulties reflect a more general problem for Craver-inspired accounts. One might respond by denying that Harinen offers a plausible account of constitutive relevance, since his variables do not represent the activities of entities standing to one another in part-whole relationships. In fact, there have been attempts to preserve the part-whole relationship between Craver's variables by altering other features of his account. Romero (2015) proposes altering Craver's notion of an intervention. He notes that if X's φ-ing is genuinely a part of S's Ψ-ing then one cannot ideally intervene on φ by intervening on Ψ, or vice versa. When intervening on X with respect to Y, any influence of the intervention on Y must be *via* X. But if φ is a part of Ψ, then any intervention on the former is necessarily and simultaneously an intervention on the later. And any intervention on Ψ is also an intervention on at least one component. Romero proposes that for establishing constitutive relevance, the appropriate type of intervention is not an ideal intervention, but a "fat-handed" intervention. Fat-handed interventions on X with respect to Y influence Y via causal paths not going through X.

Romero is correct that one cannot (generally) intervene on φ with respect to Ψ (or vice versa). Nevertheless, revising Craver's account to involve fat-handed interventions is not sufficient for clarifying why part-whole relationships matter for explanation. The issue is that any intervention on a part with respect to its whole will be fat handed, but not all part-whole relationships are constitutive relevance relationships. To determine which ones are, Romero still needs the mutual manipulability criterion. But to apply this criterion, we need to disambiguate the various possible interventions on Ψ, as Harinen does. While I have been critical of Harinen's account, his analysis of constitutive relevance as a three-variable affair remains the most straightforward way of accounting for the interventions in the inter-level experiments Craver uses to develop his account.

Attempts to modify interventionism to account for constitutive relevance are becoming increasingly sophisticated (e.g. Baumgartner and Gebharter 2015; Casini 2016). What matters for the present argument is that such attempts do nothing to reveal why talk of higher-level activities is not just a re-description of lower-level activities. In the schematic example above, $\mathbf{C}$ and Y are not distinct in the way that is typically assumed to be necessary for talk of interventions. While it is interesting to think about how to modify interventionism to deal with mereologically related variables such as φ and Ψ (or $\mathbf{C}$ and Y), discussions of how to model such variables do not reveal that we must do so to account for mechanistic behavior.

One might suspect that the difficulties I am describing for Craver's account of constitutive relevance ultimately derive from his use of an interventionist causal account of *within*-level relevance. Perhaps such an account cannot capture the complexities of

mechanistic behavior, and a better account would elucidate how the mechanistic whole is more than the sum of its parts.[7] While it seems plausible that Craver's account of within-level relevance is inadequate,[8] in the context of the present discussion rejecting this account is a step backwards. Without an account of the relationships among entities at a level, it becomes harder to evaluate whether between-level relationships distinctively contribute to explaining mechanistic behavior.

I now summarize my argument. I began this section with the claim that if the mechanistic literature has offered a distinctive form of explanation, it is an inter-level form of explanation, which we should be able to find by considering accounts of constitutive relevance. I then argued that there are ways of spelling out constitutive relevance on which we should deny that there is an inter-level form of explanation and that Harinen's account is an example. So if Harinen's account is correct, then there is no distinctive form of mechanistic explanation. This might incline one to reject his account, but I now want to explore the implications of accepting it and abandoning the idea that there is a distinctive form of mechanistic explanation.

4 Causal mediation techniques

I now explore the possibility that components are causal intermediaries between the variables that are treated as the input and output of a mechanism. I am not, of course, the first person to suggest this possibility. Menzies (2012) explicitly makes this proposal, and if the arguments in Sect. 2 succeed, then Harinen's account reduces to Menzies'. Given that such an account leaves little room for inter-level explanation, it is unsurprising that it remains a minority view. But let's put the issue of levels to the side for a moment, and turn to a more pressing question. Would identifying components with intermediate variables illuminate why scientists study mechanisms? If not, then the proposal is dead on arrival.

Fortunately one can give an account of why scientists are interested in measuring intermediate variables–or *mediators*–and of why doing so matters for explanation. In this section I will present Judea Pearl's (2001, 2012) causal mediation techniques, which provide conceptual resources for identifying the extent to which a cause influences its effect via particular mediators, and draw some implications for explanation. In presenting these techniques, I rely on an example that does not look like standard mechanistic ones. This is purposeful. One of my aims is to challenge the mechanists to say what in their accounts excludes such "non-mechanistic" cases.

Causal mediation techniques apply to models in which there are multiple causal paths linking two variables. In cases where there is just a single mediator M on a single path between two variables X and Y, the relationship between M and the effect of X on Y is not especially interesting: M contributes to the effect of X on Y insofar as when one ideally intervenes on M, the causal relationship between X and Y ceases to

[7] See Fagan (2012) for one attempt to develop an account along these lines.

[8] It is difficult to see how Craver's account would adequately model complex dynamical systems such as those discussed by Bechtel and Abrahamsen (2013). Additionally Roe and Baumgaertner (2016) raise issues with using it to understand complex mechanism-environment interactions.

obtain. The more interesting cases are those in which there are distinct causal paths going through different mediators. In such cases, it is more difficult to characterize the distinct contributions of variables along different paths. But causal mediation techniques provide conceptual resources for measuring the contributions of distinct causal paths between two variables. Pearl's (2001) treatment of mediation differs from earlier attempts (e.g. Baron and Kenny 1986) in that it allows one to model systems in which causes do not contribute additively to their effects.

Causal mediation techniques rely on the formal framework developed by Sprites, Glymour and Scheines (2000) and Pearl (2009). This framework employs directed acyclic graphs (DAGs) to represent the causal relationships among random variables. As in Woodward's account, this framework allows one to explicate causes using ideal interventions. In a DAG, an arrow between variables X and Y denotes that X is a direct cause of Y. X is a *direct cause* of Y just in case there is some ideal intervention on X that changes the value of Y while all other variables in the model are held fixed.[9] A *causal path* between X and Y is a set of connected arrows all going in the same direction from X to Y. A DAG for a set of measured variables corresponds to a set of *structural equations* in which each variable is represented as a function of its direct causes and an error term. The error term represents all unmeasured causes of a variable that are not causes of other variables in the model.

When a variable Z has two modeled causes X and Y, there is an arrow from X to Z and from Y to Z. Since there are two distinct arrows, it is tempting to read the graph as saying that the effects of these variables on Z are additive and separable. This temptation should be avoided. The value of Z is a function of its direct causes, and this function may have any form, including one in which the effect of either cause on Z depends on the value of the other. In such a case X and Y *interact.* The distinct arrows do not indicate the independence of the *causal contributions* of X and Y to Z, but rather the possibility of separately *intervening* on X and Y.

I now present causal mediation techniques using an example. Imagine that scientists develop a drug to reduce cholesterol. The drug has the intended effect, but unfortunately it also increases blood pressure. Worse, the scientists suspect that the drug is more effective at reducing cholesterol in people with higher blood pressures. They consider developing an auxiliary drug to block the effect of the cholesterol drug on blood pressure. To determine whether such a drug would be worthwhile to develop, they seek to learn more about the contribution of the drug to cholesterol *via* blood pressure.

Figure 2 presents two DAGs corresponding to this scenario. These DAGs are compatible, though they differ in that Fig. 2b includes a mediator between linking *drug* to *cholesterol*. In this DAG, we refer to *drug* as the 'treatment', *blood pressure* as the 'mediator', and *cholesterol* as the 'outcome'. The path from the treatment to the outcome via the mediator is the *indirect path*, and the other is the *direct path*. The *total effect* of taking the drug (as opposed to not taking the drug) on cholesterol is the effect going through all paths. It corresponds to the difference in (expected) level of cholesterol in the cases where one does and does not take the drug. Causal mediation

[9] Whether X is a direct cause of Y is always relative to a set of variables. X can be a direct cause of Y relative to X and Y, but not a direct cause of Y relative to a variable set that includes a variable Z that is causally between X and Y.

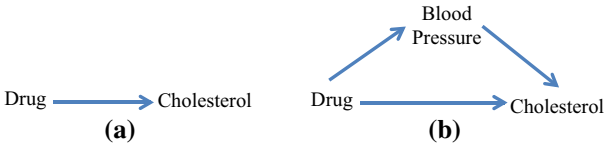

Fig. 2 In **a**, the arrow corresponds to the total effect of the drug on cholesterol. In **b**, the arrow between *drug* and *cholesterol* indicates that *drug* influences *cholesterol* through a path not going through *blood pressure.*

techniques identify the contributions of the direct and indirect paths to the total effect, though as we will see, the question of what a path 'contributes' requires disambiguation.

When the treatment and the mediator interact in producing the outcome, there is a sense in which it is impossible to fully isolate the contribution of each path. Given this interaction, the effect of the drug on cholesterol depends on the level of blood pressure. This leads to complications for initially promising proposals to quantify the influences of the paths. Consider the idea that to evaluate the effect of the drug on cholesterol along the direct path, one should perform an ideal intervention on *blood pressure*. This would have the welcome result of disrupting the indirect path, thus making it the case that (additionally) changing the value of the treatment only influences the outcome via the direct path. But this proposal is incomplete, since it does not specify the *value* to which one should set the mediator. Even in the disrupted system, the effect of the treatment on the outcome depends on the value to which one sets the mediator. Given interaction, the role of the mediator in determining the magnitude of the effect along the direct path cannot be eliminated.

Even in the presence of interaction, there is a way of understanding the contributions of particular paths to a total effect. The first thing to realize is that the contributions of paths must be evaluated relative to particular *changes* in the value of the treatment, e.g., the effect of taking the pill as opposed to not taking the pill. In thinking about this particular effect, it helps to label the scenario in which one does not take the pill as one in which the treatment variable takes on its 'default' state. Since the mediator and outcome depend on the treatment via the model's structural equations, they will also have 'default' values (or distributions) when the treatment is assigned its default state. Correspondingly, we can refer to the scenario of taking the pill as the non-default scenario, and specify the variables' non-default values analogously. Which value of the treatment gets labeled the 'default' is a matter of convention, but these labels allow one to easily talk about the complex counterfactuals involved in changing the treatment from one value to another.[10]

Given this specification of default and non-default values, one can ask: what would happen to the outcome were one to change the value of the treatment to its non-default value, while still holding the mediator at its default value via an intervention? In other words, consider changing the value of the treatment, but only allowing this change to be transmitted via the direct path. The effect measured in this way is called the *natural direct effect* (NDE). It is 'natural' in the sense that although one intervenes on the mediator, one sets it to the value (or distribution) that it would naturally have in the

[10] For a more detailed discussion of defaults and path-specific effects, see Weinberger (forthcoming).

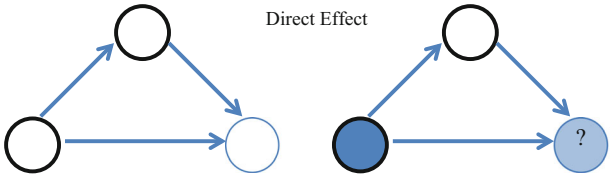

Fig. 3 *White circles* indicate the values that each variable takes on when the treatment takes on its default value and there are no further interventions. *Shaded circles* indicate the corresponding non-default values. The node with the question mark need not take on either the default or non-default value of the outcome. *Black halos* indicate which variables must be intervened upon. The direct effect is derived by subtracting the expected value of the outcome on the left from that on the right.

absence of the treatment. A schematic representation of the relevant interventions is given in Fig. 3.

The NDE of taking the drug is the effect that taking the drug (vs. not) would have on cholesterol, were the individuals who took the drug to have the same blood pressure level as they would have had were they to not take the drug. The scientists in our example should develop an auxiliary blood-pressure-reducing drug only if the NDE is non-zero. An NDE of zero would indicate that the drug is only effective when it raises blood pressure, so the auxiliary drug would cancel out the effect of the original.

One could also treat the case of taking the pill as the default, and then consider the NDE of not taking the pill as opposed to taking it. One might do so if one were considering a population of individuals who had already taken the pill, and wanted to know what the effect of stopping the pill would be if doing so did not reduce blood pressure. The NDE of not taking the pill is an interestingly different quantity than that of taking it. In evaluating the NDE of taking the pill (vs. not) and in evaluating the NDE of *not* taking it (vs. taking it) one intervenes to set the mediator to different default values. Specifically, in the latter case, one intervenes to set the mediator to the value it would take on if one *does* take the pill. One consequence of this is that the magnitude of the NDE of taking the pill is not generally the negative NDE of not taking the pill. These are two distinct quantities that are relevant for answering different questions regarding the contribution of the direct path.

What makes it possible to separate the contribution of the treatment to the mediator along the direct path is that the variables in the model are linked by structural equations indicating the counterfactual responses of variables to changes in their direct causes. This is why both the default and non-default values of the mediator are well defined. One can further exploit these structural relations in order to find the contributions of the indirect path. Our model does not include any variable along the direct path that one might intervene upon, but, surprisingly, it is not necessary to include one. The trick is to keep the treatment variable at its default value, while varying the mediator from its default value to its non-default value (Fig. 4). By varying the value of the mediator, one makes it behave as if it were responding to a change in the treatment. By holding the treatment fixed at its default value, one ensures that the change in the outcome is not the result of change transmitted along the direct path.

The quantity measured in this manner is the *indirect effect*. Treating not taking the pill as the default, the indirect effect of the pill on cholesterol is the difference in

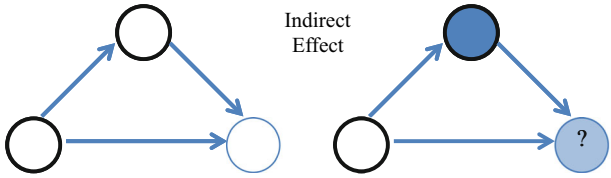

Fig. 4 *White circles* default values. *Filled circles* non-default values. *Question mark* value unknown. *Black halos* interventions. The indirect effect is derived by subtracting the expected value of the outcome on the left from that on the right

(expected) cholesterol level between the case where one does not take the pill, but has the blood pressure as if she did, and the case where one has the default levels of both the treatment and blood pressure.

Counterintuitively, the total effect of taking the drug versus not taking it is not (in general) the sum of the direct and indirect effects for this change in the treatment. This is a consequence of the fact that in systems with interaction, the paths need not contribute additively to their effect. Although it is common to say that in mediation one decomposes the total effect into direct and indirect effects, talk of decomposition may falsely suggest that one can divide the total effect into the additive contributions of the direct and indirect effect (evaluated relative to the same default).[11] This is not what mediation does. Informally, the direct and indirect effects of taking a pill (versus not taking it) are the amounts by which taking the pill would increase the expected value of the outcome, were the effect to be transmitted by either the direct or indirect path alone. Typically, the total effect will not be the sum of the effects of the paths in isolation.

Although I have here defined the *natural* direct effect, in principle any ideal intervention on the mediator will disrupt the indirect path and measure a quantity that we could refer to as a direct effect (though note that there would be as many direct effects as there are values of the mediator). So why focus on the *natural* direct effect? To see this, suppose one were to intervene to set the mediator to some arbitrary value other than the ones it takes on when the treatment has its default or non-default values. Learning about the behavior of the direct path given this intervention on the mediator would not be helpful for understanding the contribution of the direct path in the case where one does *not* intervene on the mediator. In contrast, with the natural direct effect, one intervenes on the mediator to make it behave as if it were still responsive to the treatment (since which value of the mediator counts as the "default" depends on the structural equation linking the treatment to the mediator), but it were not responding to the change in the value of the treatment.

The following household example will help clarify the intuition behind using the specific quantities defined. Suppose one wants to know whether closing the refrigerator

[11] Pearl (2001) provides mathematical decompositions of the total effect of changing the treatment from x to $x\prime$ into the direct effect of going from x to $x\prime$ minus the indirect effect of going from $x\prime$ to x, and into the indirect effect of going from x to $x\prime$ minus the direct effect of going from $x\prime$ to x. These decompositions follow from the definitions of the given effects and do not correspond to the intuitive decomposition of the behavior of the whole into distinct contributions of its parts.

door affects the fridge light via the button that gets pressed when one closes the door. The relevant test is to open the refrigerator and press the button with one's hand to ascertain that the button's being pushed causes the light to go off. The reason that this is the relevant test for evaluating the effect of the door opening on the light via the button is that the pressed state of the button is the state it would be in were one to close the refrigerator. If, hypothetically, the button could take on a state other than pushed or unpushed, then setting it to that state would not be informative about what would happen when the door is opened normally. Although this case is one in which there is only a single path, and in which one is measuring the indirect effect (which here equals the total effect), it illustrates the motivation for defining direct and indirect effects in terms of the counterfactual values that the mediator would take on given particular values of the treatment.

With the concepts of direct and indirect effects defined, we can now turn to the question of why measuring mediators matters for explanation. The answer is straightforward: they enable one to predict how the effect of the treatment on the outcome would change under counterfactual circumstances. Most obviously, they enable one to predict what the effect would be were one of the paths to be disrupted. More generally, it is possible to show that the NDE is invariant across populations with the same default value of the mediator, but which otherwise differ in the structural equation linking the treatment to the mediator (see Weinberger 2015; a similar result is available for the indirect effect.) They thereby enable one to answer a wider range of "w-questions" of the form: What if things had been different? (Woodward 2003 p. 191). Hitchcock and Woodward (2003) plausibly argue the depth of an explanation corresponds to the range of w-questions it is able to address.

The idea that uncovering the components in a mechanism helps one determine the conditions under which the mechanism's activity will be maintained is widespread. Glennan writes:

> Understanding the nature, structure, and functional organization of the parts that make up that mechanism will allow one to determine the range of counterfactual circumstances under which the dependency between X and Y would be maintained–roughly those circumstances in which the mechanism will not break down. (2012, p. 288)

As intuitive as this idea is, there has been little systematic inquiry into how studying mechanisms aids one in extrapolating the functional causal relationship between X and Y across contexts. The one full-length manuscript on the topic (Steel 2008) situates itself in the mechanist literature, although its results appear to be based on principles from graphical causal models rather than any specifically mechanistic commitments. In any event, we can see that mediation techniques do facilitate extrapolation without making any distinctively mechanistic commitments, although this point merits a more comprehensive discussion than I can provide here.

To the extent that components of mechanisms can be characterized as mediators, we can straightforwardly see why discovering components matters for explanation. The explanatory value of these entities in no way depends on whether we refer to them as "components". There is also no obvious need to talk about "levels". True,

the direct and indirect effects are neither causes nor effects of the total effect. But this does not show that there is some new form of explanation in play. Mediation techniques do not supplement causal explanation with another form of explanation. Rather, they show how one can replace one causal explanation with more fine-grained causal explanations.

The cholesterol drug example does not look like standard examples from the mechanisms literature. There is no detailed description of the activities by which the drug brings about higher blood pressure and lower cholesterol and no physiological account of why the drug has its effect. And certainly in some contexts we will want these additional pieces of information. What the mechanists have not shown, however, is that explanations that do and do not include such information are of a different type. For all the talk of the importance of learning about the complex physical organization of mechanisms, details about such organization play no explicit role in the explanatory account. In fact, when one fills in the details of Craver's account with help from Harinen, the process of discovering "lower-level" mechanistic components turns out to be one of providing a more detailed explanation of the mechanism's behavior using causal counterfactuals. This might seem like an unwelcome result. Yet I have argued that counterfactual accounts can elucidate the explanatory value of discovering components, while existing mechanistic accounts leave this opaque.

5 Conclusion

By now there has been an abundance of work on the complications faced by Craver's account of constitutive relevance, and on the question of whether it is causal. In general, this literature presupposes that mechanistic concepts matter for explanation, even if there are details to be filled in about the relationship between a phenomenon and the components underlying it. The modest point made here is that once we fill in these details, we need go back and verify that the accounts provided reveal there to be a distinctive inter-level form of mechanistic explanation. More ambitiously, I argue that when one fills in the details of the dominant account in the literature (Craver's), we reveal there to be no such form of explanation. But perhaps we do not need one. Even if we only consider "within-level" explanatory relationships, we can account for why studying the components in a mechanism enables us to see how a cause brings about its effect and to predict the behavior of the mechanism across counterfactual circumstances. If there are explanatory benefits to talking about the relationships among mechanism levels, mechanists still need to specify what these are.

Acknowledgements I would like to thank Matteo Colombo, Dan Hausman, Shannon Nolen and three anonymous reviewers for extensive feedback. I would also like to thank Erik Nyberg, Felipe Romero, Michael Schon, Felix Elwert and audiences in Tilburg and at the 2016 meeting of the Philosophy of Science Association for helpful discussion.

Funding This work was supported by the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program "New Frameworks of Rationality" (SPP 1516).

References

Baumgartner, M., & Gebharter, A. (2015). Constitutive relevance, mutual manipulability, and fat-handedness. *The British Journal for the Philosophy of Science*, *67*(3), 731–756.

Bechtel, W., & Abrahamsen, A. (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), 421–441.

Bechtel, William, & Abrahamsen, Adele. (2013). Thinking dynamically about biological mechanisms: Networks of coupled oscillators. *Foundations of Science*, *18*(4), 707–723.

Bechtel, W., & Richardson, R. C. (1993). *Discovering complexity: Decomposition and localization as strategies in scientific research*. Princeton: University Press.

Casini, L. (2016). How to model mechanistic hierarchies. *Philosophy of Science*, *83*(5), 946–958.

Craver, C. (2002). Interlevel experiments and multilevel mechanisms in the neuroscience of memory. *Philosophy of Science*, *69*, S83–97.

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

Craver, C. F. (2013). Functions: Selection and mechanisms. *Functions and mechanisms: A perspectivalist view* (pp. 133–158). Netherlands: Springer.

Craver, C., & Bechtel, W. (2007). Top-down causation without top-down causes. *Biology and Philosophy*, *22*(4), 547–563.

Craver, C. F., & Darden, L. (2013). *In search of mechanisms: Discoveries across the life sciences*. Chicago: University of Chicago Press.

Fagan, M. B. (2012). The joint account of mechanistic explanation. *Philosophy of Science*, *79*(4), 448–472.

Franklin-Hall, L. R. (2016). New mechanistic explanation and the need for explanatory constraints. *Scientific composition and metaphysical ground* (pp. 41–74). Basingstoke: Palgrave Macmillan.

Glennan, S. S. (1996). Mechanisms and the nature of causation. *Erkenntnis, 44*(1), 49–71.

Glennan, S. (2012). Mechanisms. In The Oxford (Ed.), *Handbook of causation*. Oxford University Press, Oxford.

Harinen, T. (2014). Mutual manipulability and causal inbetweenness. *Synthese*, 1–20.

Hitchcock, C., & Woodward, J. (2003). Explanatory generalizations, part II: Plumbing explanatory depth. *Noûs*, *37*(2), 181–199.

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

Levy, A. (2013). Three kinds of new mechanism. *Biology & Philosophy*, *28*(1), 99–114.

Leuridan, B. (2011). Three problems for the mutual manipulability account of constitutive relevance in mechanisms. *The British Journal for the Philosophy of Science*, *63*(2), 399–427.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of science*, *67*(1), 1–25.

Menzies, P. (2012). The causal structure of mechanisms. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(4), 796–805.

Pearl, J. (2001). Direct and indirect effects. In: Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Francisco, CA, 411–420.

Pearl, Judea. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge: Cambridge University Press.

Pearl, J. (2012). The causal mediation formula–a guide to the assessment of pathways and mechanisms. *Prevention Science*, *13*(4), 426–436.

Roe, S. and Baumgaertner, B. (2016). Extended mechanistic explanations: Expanding the current mechanistic conception to include more complex biological systems. *Journal for General Philosophy of Science*

Romero, F. (2015). Why there isn't inter-level causation in mechanisms. *Synthese*, *192*(11), 3731–3755.
Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search* (Vol. 81). Cambridge: MIT press.
Steel, D. P. (2008). *Across the boundaries: Extrapolation in biology and social science*. New York: Oxford University Press.
Weinberger, N. (2015). *Causal inference across populations* (Doctoral dissertation, THE UNIVERSITY OF WISCONSIN-MADISON).
Weinberger, N. (forthcoming). Path-specific effects. *British Journal for the Philosophy of Science*
Woodward, James. (2003). *Making things happen*. Oxford: Oxford University Press.