Final draft – Published in Synthese (2015), 192(11), 3731–3755.

WHY THERE ISN'T INTER-LEVEL CAUSATION IN MECHANISMS

Felipe Romero[1]

Abstract

The experimental interventions that provide evidence of causal relations are notably similar to those that provide evidence of constitutive relevance relations. In the first two sections, I show that this similarity creates a tension: there is an inconsistent triad between (1) Woodward's popular interventionist theory of causation, (2) Craver's mutual manipulability account of constitutive relevance in mechanisms, and a variety of arguments for (3) the incoherence of inter-level causation. I argue for an interpretation of the views in which the tension is merely apparent. I propose to explain inter-level relations without inter-level causation by appealing to the notion of fat-handed interventions, and an argument against inter-level causation which dissolves the problem.

1. INTRODUCTION

On mechanistic views of explanation (Machamer, Darden, & Craver, 2000), good explanations in the biological sciences characterize mechanisms: entities with specific roles, composed by other entities that interact causally. In this paper I discuss a problem that arises frequently in discussions about mechanisms. I articulate the problem as a tension between three plausible views:

(1) The *interventionist view of causation* (Woodward, 2003), according to which ideal interventions, i.e., manipulations that change the value of one variable X and produce a change in another variable Y, controlling for other influences, are necessary and sufficient for X to be a direct cause of Y.

(2) The *mutual manipulability account of constitutive relevance* (Craver, 2007), according to which we can know whether a component is constitutively relevant in a mechanism by finding an experimental intervention on the component that produces a change in the mechanism, and *also* the other way around.

(3) The *incoherence of inter-level causation*, the idea that causal relations between entities at different levels, such as a mechanism and its components, produce problematic redundant and cyclic causal structures.

My aim is not to argue independently for (1)-(3). My interest is to offer an account of how they fit together. I have two main reasons. First, the three views have reached a prominent status in the mechanisms literature. Appeals to inter-level causation in explanations often raise concerns for philosophers of science, and Craver's mutual manipulability account and Woodward's interventionism are mandatory starting points for discussions about explanation in the life sciences, even when critics disagree with them (for instance, Waskan, 2011; Fagan,

[1]Department of Philosophy, Washington University in St. Louis; cfromero@wustl.edu

I wish to thank Carl Craver, John Heil, Frederick Eberhardt, Lena Kästner, Lauren Olin, Isaac Wiegman, Mark Povich, and two anonymous reviewers for *Synthese* for comments on previous drafts. I also received helpful comments when this paper was presented at the 2012 Models and Mechanisms Conference, Tilburg University; and the 2013 St. Louis Area Philosophy of Science Association Meeting, University of Missouri in St.Louis.

2012; Leuridan, 2012; Franklin-Hall, in press). Hence, it is useful to explore their relations, and whether they are consistent, to help move the discussion forward.

The second reason is more substantive: the three views jointly offer an account of experimental practices that it is worth preserving. In the life sciences, scientists often try to offer causal explanations (as opposed to mere correlations) and often perform experiments to identify components of systems. These practices are so common that it would be good, to say the least, if philosophers of science could offer a story to make sense of them and/or assess them critically. Not having such a story would lead to an undesirable gap, which would suggest that the practices lack proper philosophical foundations, or that philosophers of science have no concern for what scientists actually do. We can interpret these three views as a contribution to fill that gap. And even though there are certainly alternatives to each view (and I spend some time discussing some of them in section 4), it is not clear whether such alternatives serve this epistemological purpose.

The remainder of the paper is organized as follows. In section 2, I unpack (1), (2), and (3) and explain the tension in detail, showing how the views seem to form an inconsistent triad, in the sense that accepting any two entails the negation of the third. At the end of the section, I discuss their relation with the basic mechanistic view. In section 3 I attempt to dissolve the tension. I show an alternative way of understanding the relation between mechanisms and their components using the notion of "fat-handed" intervention. I argue that the interventions that reveal constitutive relevance are fat-handed interventions. From this claim two things follow. On the one hand, it follows that (1), (2) and (3) are not mutually inconsistent. On the other hand, (1), (2), and the claim about fat-handedness entail that there is no inter-level causation within a mechanism. Finally, in section 4, I review the consequences and unwanted costs of rejecting (1), (2), or (3) to avoid the inconsistency.

2. Mutual Manipulability, Interventionism and Inter-Level Causation

Mechanistic views of explanation have earned an important place in philosophy of biology and neuroscience. We can synthesize the insights of this view in three main claims. First, *mechanisms are bundles of structure and activity*. For example, Machamer, Darden, and Craver (2000, p.3) define mechanisms as "entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions". This way of understanding explanations integrates behaviors with physical structures and temporal dynamics. In addition, *mechanisms are causal structures*. Their underlying behavior is produced by a complex system of parts interacting causally (Glennan, 1996, p.52). Finally, *mechanisms are multi-level structures*: "entities and activities in mechanisms are organized together spatially, temporally, causally, and hierarchically" (Craver, 2007, pp.5-6).

The aforementioned mechanists would subscribe to versions of these three claims, and henceforth I will refer to mechanistic explanation as a view that embraces the three of them. Nonetheless, it is worth mentioning that different mechanists use the notion of mechanism for different (and not obviously consistent) purposes in metaphysics, explanation, and methodology (see Levy, 2013,

for a discussion of these different purposes). In addition, there are differences between mechanistic views. For instance, in Glennan's view, parts of mechanisms interact according to "direct causal laws" (i.e., generalizations that support counterfactuals) (Glennan, 1996, p.52), but Machamer, Darden, and Craver explicitly reject this commitment (2000, p.4). On the other hand, Machamer, Darden, and Craver (2000) explain causation in terms of "activities" (which is akin to productivity accounts of causation), whereas Craver (2007) subscribes to Woodward's interventionism (more on this on Section 2.2). I will begin characterizing the interventionist view.

2.1. The Interventionist View of Causation

One popular and widely endorsed way of understanding causation (at least in philosophy of science) is Woodward's interventionist view (also called the "manipulationist view"), which defines causation in terms of interventions. The intuition is that if one "surgically" manipulates a variable X and that produces a change in Y, it is only because there is a causal relationship between X and Y. More precisely:

(M) A necessary and sufficient condition for X to be a (type-level) direct cause of Y with respect to a variable set V is that there be a possible intervention on X that will change Y or the probability distribution of Y when one holds fixed at some value all other variables Z_i in V. A necessary and sufficient condition for X to be a (type-level) contributing cause of Y with respect to variable set V is that (i) there be a directed path from X to Y such that each link in this path is a direct causal relationship; [...], and that (ii) there be some intervention on X that will change Y when all other variables in V that are not on this path are fixed at some value. (Woodward, 2003, p.59)

Woodward's idea is that given two variables X and Y, if, after fixing all other variables, all we need to do to change Y is to wiggle X, then it means that X is a cause of Y. It is worth clarifying that the condition of holding fixed other variables means that there is at least one assignment of values for all other Z_i in V such that some intervention on X will change Y. The condition does not require that the change should occur under all assignments of values of Z_i (Woodward, 2003, p.53). This hypothetical scenario represents experimental conditions in which (ideally) other causal influences on Y are stable, allowing us to distinguish the causal influence that an intervention on X has on Y. Now, what does it mean to intervene on X? Woodward gives the following definition:

(M*) I is an intervention variable for X with respect to Y iff

1. I causes X.
2. I acts as a switch for all other variables that cause X. That is, certain values of I are such that when I attains those values, X ceases to depend on the values of other variables that cause X and instead depends only on the value taken by I.
3. Any directed path from I to Y goes through X. That is, I does not directly cause Y and is not a cause of any causes of Y that are distinct from X except, of course, for those causes of Y, if any, that are built into the $I - X - Y$ connection itself; that is, except for (a) any causes of

Y that are effects of X (i.e., variables that are causally between X and Y) and (b) any causes of Y that are between I and X and have no effect on Y independently of X.

4. I is (statistically) independent of any variable Z that causes Y and that is on a directed path that does not go through X. (Woodward, 2003, p. 98)

To get an intuitive idea of these conditions, consider the following reconstruction of one important case of causal discovery in medicine. In the 17th century, Ignaz Semmelweis was puzzled about the large difference in mortality rates due to puerperal fever in women who were treated at two very similar maternity clinics in Vienna. After ruling out explanations such as climate, he realized that at the clinic with the highest mortality rate, women were treated by medical students who were also performing autopsies. This led him to hypothesize that cadaveric material was involved in causing the fever. And then, he established the policy that students had to wash their hands after performing autopsies using a chlorinated lime solution, which he found good to get rid of cadaveric smells. This was a novel insight at the time, because the idea that germs could cause diseases had not yet been developed. As a result, the mortality rate decreased dramatically in the first clinic, and eventually was similar to that of the second clinic (Semmelweis, 1983). Figures 1(a) and 1(b) represent this situation.

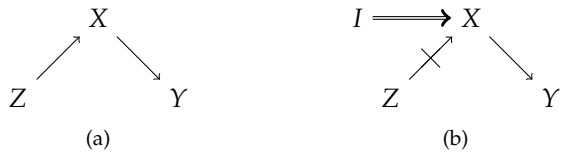

Figure 1: Intervention and Causation

Suppose all the variables in Figure 1(a) are binary. Z represents whether the medical students perform autopsies at the clinic; X represents whether the students' hands are contaminated with cadaveric material when they treat patients; Y represents whether patients gets puerperal fever. In a graph representation, an intervention can be seen as a variable within the system that alters the value of the variable intervened upon (Pearl, 2000, p.70). This is shown in Figure 1(b). I is the intervention of making students wash their hands in the chlorinated lime solution. The double arrow represents the fact that the intervention's causal influence is extrinsic to the system. Now, there is an observed correlation between Z and Y. An ideal intervention would break the causal relation from Z to X (represented by the crossed out arrow), making it irrelevant whether patients are treated by students who do autopsies or not; and would also change the probability distribution of Y, which would mean that there is a causal relation between X and Y.

An intervention is not ideal if it violates any of the conditions in (M*). Figures 2(a), (b), and (c) show violations of conditions (2), (3), and (4) respectively. Suppose, for example, that the policy of washing hands only applies to half of the students. In such a case, I would not break the arrow from Z to X (i.e, some students with contaminated hands would still treat patients), so the intervention is not ideal. This is represented in Figure 2(a). On the other hand, Figure 2(b)

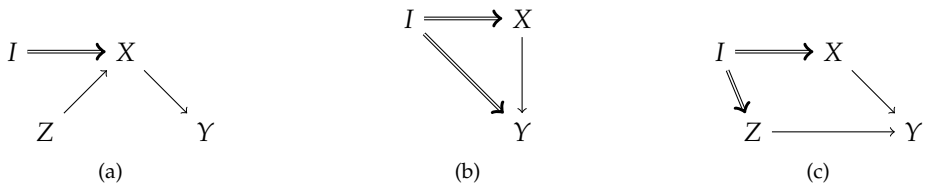

Figure 2: Violations of Conditions (M*)

represents a case in which the intervention changes both the purported cause and the effect. For example, suspending all operations at the clinic would trivially (but effectively) prevent students with contaminated hands from treating patients, and patients from getting fever at the clinic. And Figure 2(c) could represent a similar situation, but with another causal intermediate between I and Y. For a more recent discussion about these conditions, see Woodward (2014, pp. 12-14). And for a more formally precise characterization of interventions in causal graphs, see Eberhardt and Scheines (2007).

Whether interventionism counts as a complete story about the nature of causation is a contentious issue. Often, discussions about causation distinguish two different (and competing) accounts: "difference-making" accounts (including interventionism), on the one hand, which support the idea that causes make a difference to their effects; and "geometrical/mechanical" or "production" accounts, on the other, which understand causation as a "connecting process" between events (see Woodward, 2011, pp.411–413). For the purposes of this paper, I interpret interventionism primarily as a theory about the *semantics* of causal claims, which is the way in which Woodward presents it in *Making Things Happen* (2003, p.28). There are other interpretations, and I discuss them in Section 4.1. But as a semantic theory, interventionism provides "truth conditions for claims about total, direct, contributing, and actual causal relationships"(2003, p.95), and has implications about the epistemology of causation, without further commitments about the nature of causation.

One reason the interventionist view is popular among philosophers of science is because it establishes a conceptual relation between causation and experimental practices. In particular, in the context of discussions about explanatory models in the biological sciences, it attempts to analyze the connection between the results of controlled experiments and causal claims.

Specifying a mechanism requires identifying its components. It is desirable, therefore, for an account of mechanistic explanation, to provide an account of when a given entity is a component in a mechanism and when it is not; that is, an account of *constitutive relevance.* Craver's account, which I explain next, is the *mutual manipulability* account, and it was developed in the context of interventionism. This account is the second view in the inconsistent triad.

2.2. The Mutual Manipulability Account of Constitutive Relevance

Craver's mutual manipulability account offers a sufficient condition for constitutive relevance. He says "a part is a component in a mechanism if one can change the behavior of the mechanism as a whole by intervening to change the compo-

nent and one can change the behavior of the component by intervening to change the behavior of the mechanism as a whole" (Craver, 2007, p.141). I will unpack this view using an example and some more precise definitions.

Different subfields of cognitive science approach the same phenomena from different perspectives. According to Craver, we can describe multi-level mechanisms of such phenomena integrating evidence from these subfields. Given some phenomenon, the different subfields may provide evidence of how the mechanism that produces it works at higher and lower levels. Craver uses evidence about spatial memory encoding and retrieval to illustrate these points (Craver, 2007, p.165). Here I take the case of the well-studied phenomenon of face recognition in humans:

1. *Psychological studies of performance in face recognition*. Behavioral experiments in psychology provide evidence that people are particularly sensitive to faces (Bruce & Young, 1986). We are able to distinguish and identify hundreds of different faces, but are rarely so good at identifying other sorts of objects. We can describe these experiments as taking place at a high mechanistic level –the level of the whole organism.

2. *Studies that identify brain structures involved in face recognition*. At a lower mechanistic level, imaging studies have consistently shown that subjects presented with facial images have major activation of the fusiform gyrus, and that activation is greater than observed in subjects presented with other images (Meng, Cherian, Singal, & Sinha, 2012). Additionally, classic studies (Damasio, Damasio, & Van Hoesen, 1982) show that patients with damage in the fusiform gyrus exhibit prosopagnosia, an inability to recognize familiar faces. Other imaging studies (Ishai, Schmidt, & Boesiger, 2005) find activation within specific networks that include the fusiform gyrus when subjects are asked to identify facial stimuli.

3. *Studies about the activity of particular cells in face recognition tasks*. At an even lower level, vision scientists study the activity of particular cells recruited in face recognition. For example, using fMRI, Tsao et al. (2006) identified a face-selective region in macaques which they regard as topographically homologous to the human fusiform face area (located in the fusiform gyrus in humans). Then, they did single-unit recordings in that area and were able to identify cells with a 19:1 and 14:1 ratio of face-to-nonface object response.

Here are some features of Craver's notion of "level". First, levels in a mechanism are not just levels of objects that compose other objects (e.g., atoms, molecules, cells, systems, organisms, and societies). Although entities at lower levels are smaller than entities at higher levels, they also need to be relevant to the higher levels in terms of the activities they perform (more on this shortly). Second, Craver's view of levels is different from the view that levels are monolithic strata identifiable in terms of size. On such a view, reality is a hierarchical structure that exhibits regularities at different size scales (e.g., regularities at the atomic level, molecular level, etc.). The problem for this view is that there are no well-defined strata in cognitive science. Hence, levels in a mechanism have a more *local* significance: what each field contributes depends on the explanatory context, and the relevance of entities at different levels must be determined case by case.

Craver defines the relation between levels and components as follows. Let x be an entity, for example, a cell in the fusiform gyrus, and ϕ be the activity of firing. Craver calls "acting entity" the performing of the activity by the entity, which we can denote by ϕ_x. That is, ϕ_x is a firing neuron. I simplify Craver's notation: he uses "X's ϕ-ing" to refer to what I call here ϕ_x (a notation that has the virtue of privileging activities over entities). Also, let s be a higher-level mechanism, whose behavior ψ_s is (at least partially) associated with ϕ_x. For example s could be the subject in a face recognition experiment and ψ the activity of recognizing a face. Thus we have:

(L) ϕ_x is at a lower mechanistic level than ψ_s if and only if ϕ_x is a component in the mechanism for ψ_s. (Craver, 2007, p.189)

This notion of composition is thick: describing a component requires specifying not only a part-whole relation between x and s, but also the activities ϕ and ψ that make x relevant to s. Merely being a cell in the fusiform gyrus, for instance, is not sufficient to be a component in the mechanism for face recognition. The cell has to be active when the subject performs the activity of face recognition to be considered a component. In this notation, the problem of giving an account of constitutive relevance is to establish conditions under which ϕ_x is a component in the mechanism for ψ_s.

In practice, myriad cases in neuroscience illustrate how scientists discover mechanisms by identifying their lower-level components in inter-level experiments. For example, the imaging experiments that provide evidence that the fusiform gyrus is involved in face recognition (Ishai et al., 2005) are *top-down* experiments: there is a higher-level intervention (i.e., the subject is presented with facial stimuli) and a low-level detection technique (i.e., fMRI). Additionally, there are *bottom-up* experiments. Lesions studies, such as Damasio's (Damasio et al., 1982), provide a set of examples: there is a low-level intervention (i.e., a lesion that damages the fusiform gyrus) and a higher-level detection technique (i.e., a behavioral experiment that reveals prosopagnosia). Consider also the case of spatial memory discussed by Craver. It is known that cells in the hippocampus are active at sustained rates when rats move inside a specific place field. These studies are *top-down*, in the sense that there is a top-level intervention (i.e., the rat is moving over the specified field) and a low-level detection technique (i.e., cell recording). Additionally, there are *bottom-up* experiments in which lesioning a rat's dorsal hippocampus impairs its navigation. Specifically, lesions in the dorsal hippocampus cause spatial memory impairment in rats navigating the radial arm maze. Scientists' conclusion is that the hippocampus plays a role in a mechanism for spatial memory.

Drawing on the spatial memory example, Craver states that ϕ_x and ψ_s are *mutually manipulable* when it is possible to perform both bottom-up and top-down interventions. Craver says "My working account of constitutive relevance is as follows: a component is relevant to the behavior of a mechanism as a whole when one can wiggle the behavior of the whole by wiggling the behavior of the component and one can wiggle the behavior of the component by wiggling the behavior as a whole. The two are related as part to whole and they are mutually manipulable" (2007, p. 153). And also "$[\phi_x]$ is constitutively relevant to $[\psi_s]$ if the two are related as part to whole and the relata are mutually manipulable. There should be some ideal intervention on ϕ under which ψ changes, and there should

be some ideal intervention on ψ under which ϕ changes" (2007, p. 154). I state this view and its relation to the notion of component as follows:

(MM1) ϕ_x and ψ_s are mutually manipulable if and only if there is an ideal bottom-up experimental intervention on ϕ_x that produces a detectable change in ψ_s; and there is an ideal top-down experimental intervention on ψ_s that produces a detectable change in ϕ_x.

(MM2) If ϕ_x and ψ_s are mutually manipulable, and x is a part of s, then ϕ_x is a component in the mechanism for ψ_s.

Condition (MM2) is important because there are cases in which we can have mutual manipulability of two variables but not a componential relation, such as when there are causal loops. The parthood condition rules out this possibility. On the other hand, as mentioned before, being a part is not sufficient for being a component. For example, the hubcaps are parts of a car, but they are not a component in the mechanism that makes the car run (Craver, 2007, p.140).

The definition of ideal intervention that Craver offers is as follows:

(I1c) The intervention I does not change ψ_s directly;

(I2c) I does not change the value of some other variable ϕ'_x that changes the value of ψ_s except via the change introduced into ϕ_x;

(I3c) that I is not correlated with some other variable M that is causally independent of I and also a cause of ψ_s; and

(I4c) that I fixes the value of ϕ_x in such a way as to screen off the contribution of ϕ_x's other causes to the value of ϕ_x (Craver, 2007, p.154).

For Craver, mutual manipulability adds an important piece to mechanistic accounts of explanation: it shows how to identify components in a mechanism, a project that he regards as a "regulative ideal" in neuroscience. In his view, scientists (should) aim for constitutive explanations that "describe all and only the component entities, activities, properties, and organizational features that are relevant to the multifaceted phenomenon to be explained" (Craver, 2007, p.111). Now, the fact that mutual manipulability relies heavily on interventionism can be seen as a problem, as I show next.

2.3. The Incoherence of Inter-Level Causation

It is easy to notice that interventionism and the mutual manipulability view rely on similar notions of ideal intervention. Leuridan (2012), for instance, argues that constitutive relevance in Craver's account threatens to imply causal relevance in Woodward's framework, by showing that conditions (I1c)-(I4c) imply conditions 1-4 in (M*). To see the similarly more generally, consider the way in which the two notions are formulated: first, as definition (MM1) in section 2.2 shows, mutual manipulability uses a notion of intervention that captures the fact that inter-level experiments are controlled (i.e., all the non-intervened on variables should be kept constant); and second, as (M) in section 2.1 shows, interventionism uses a notion of intervention according to which interventions manipulate surgically one variable to observe changes in another while holding all other variables constant (i.e., interventions are controlled experiments).

If the two notions are essentially the same, then we have:

(MM*) ϕ_x and ψ_s are mutually manipulable if and only if ϕ_x is a contributing cause of ψ_s, and ψ_s is a contributing cause of ϕ_x.[2]

Then, (MM*) and (MM2) and (L) entail the following inter-level causation thesis:

(ILC) If ϕ_x and ψ_s are mutually manipulable, and x is a part of s, then ϕ_x is at a lower mechanistic level than ψ_s, ϕ_x is a contributing cause of ψ_s, and ψ_s is a contributing cause of ϕ_x.

Mechanists in this debate don't like inter-level causation. Craver's critics take it to be a problem for his mutual manipulability account. And Craver and Bechtel offer arguments to avoid it. Craver says "many philosophers have held that causes and effects must be logically independent. If one endorses this restriction on causal relations, then one should balk at positing a causal relationship between constitutively related properties" (Craver, 2007, p.153) and "If one is committed to the idea that causes must precede their effects, then constitutive relationships are not causal relationships" (Craver, 2007, p.154). On the other hand, Craver and Bechtel (Craver & Bechtel, 2007) attempt to explain (away) inter-level causal relations proposing the notion of "mechanistically mediated effects". Such effects are "hybrids of causal and constitutive relations, where causal relations are exclusively intralevel"(Craver & Bechtel, 2007, p.547), and inter-level relations are explained in terms of constitution, which metaphysically excludes causation.

Why is inter-level causation so undesirable? In the context of mechanisms, I can articulate two unwanted consequences:

1. *Redundancy Problem.* Suppose physical events are fully caused by other physical events. This physicalist assumption is intuitive when we think of the lower levels of reality (e.g., particles exerting forces, colliding, and transferring quantities). Suppose specifically, that some ϕ_x has a sufficient same-level set of causes, represented by ρ_y. Now, suppose we find that ϕ_x and ψ_s are mutually manipulable, and x is a part of s. Then, by (ILC), it follows that ϕ_x is at a lower mechanistic level than ψ_s, and ψ_s is a contributing cause of ϕ_x. However, given that ρ_y sufficiently causes ϕ_x (by assumption), ϕ_x has redundant causes. For example, if all neural activity is caused by cellular activity (e.g., neurons' firing), then higher-level structures (e.g., cerebellum, hippocampus, cortex, etc.) would have redundant causal influences (e.g., it would be redundant to say that the cortex has some regulatory function on other structures). Causal redundancy itself is not a problem in general. Many events are overdetermined (e.g., one bullet might be sufficient cause of Jones' death, and all the other bullets overdetermine his death). The problem is that if (ILC) is true, then *all* the activities of a mechanism's components are redundant, and that sort of relation might be different from more intuitive cases of overdetermination. This is an argument in the vicinity of the *exclusion problem* in philosophy of mind (Kim, 1989). The problem is whether supervening entities (e.g., minds) can have top-down causal influence on lower level physical entities. Consider the *closure principle*: for every physical property P, there is a set of sufficient physical causes $P*$. Suppose M is a mental event that supervenes on

[2]Given that all direct causes are contributing causes, I write the conclusion in terms of the latter.

(is not reducible to) P. It follows from the closure principle that either M overdetermines P or that M is not a contributing cause of P. For recent discussions about Kim's exclusion problem in the context of interventionism see Baumgartner (2009) and Woodward (2014).

2. *Cyclicity Problem.* Suppose, again, that we find that ϕ_x and ψ_s are mutually manipulable, and x is a part of s. Then, by (ILC), it follows that ϕ_x is a contributing cause of ψ_s, and ψ_s is a contributing cause of ϕ_x. One plausible interpretation is that causal cycles represent feedback relations in the mechanism. If we add a temporal variable, the cycle means that at t_1, ϕ_x is a contributing cause of ψ_s, and t_2, ψ_s is a contributing cause of ϕ_x. There are cases of such feedback relations in biological organisms. One clear example is when organisms exhibit negative feedback loops that allow them to regulate the levels of certain values (e.g., temperature) around one target set point, as in homeostasis (Bechtel, 2011). However, the temporal order required to make such an interpretation plausible is not available in Craver's understanding of levels, because the activities that constitute higher-order behavior are supposed to be synchronic with the behavior. Hence, it is unlikely that the relation between a mechanism and *all* its components is feedback. The other alternative is to interpret such causal cycles synchronically, but there is also a problem: if inter-level relations between ϕ_x and ψ_s are causal, then s could not ψ at t, unless x could ϕ at t; and it is puzzling to say that x was *also* caused to ϕ at the same time t by s's being able to ψ. The reason for puzzlement comes from the intuitive principle that an entity cannot be caused to have a causal power at t and also exercise that causal power at t. This form of argument comes from Kim (1999, pp.28-29).

I have presented one argument: mutual manipulability and interventionism (in the way in which I stated them) entail that there are inter-level causes. Negating the conclusion of this argument (as I contend we should), leads to an inconsistent set of claims. The other two directions of the inconsistency are now easier to see: if we accept mutual manipulability and reject inter-level causes, then we need an alternative theory of causation in which the interventions that reveal a mutual manipulability relation between two entities are not sufficient for there to be a causal relation between them; and if we accept interventionism without inter-level causes, then we have to reject the mutual manipulability account. These alternatives are discussed in more detail in sections 4.1 and 4.2 respectively.

2.4. The Inconsistent Triad in Relation to the Basic Mechanistic View

Let me say how these three views relate to the three features of the mechanistic view presented at the beginning of this section. I said mechanisms are bundles of structure and function, and they are also causal structures. Most mechanists can agree about such tenets. Now, the details of what it means for two components of a mechanism to be in a causal relation is an additional issue, which could be addressed with different theories of causation. Some mechanists appeal to interventionism, but it is not necessary. Indeed, the seminal papers in the literature (Glennan, 1996; Machamer et al., 2000) don't appeal to it. Nonetheless, in addition to providing semantics for causal claims, interventionism can be used to tell

a story about the discovery of causal relations that strengthens the epistemology of mechanisms. This advantage does not come for free, because the interpretation of interventionism also as a metaphysical theory is controversial (I discuss this issue in section 4.1). However, if we displace interventionism, even if we could perhaps use more metaphysically robust accounts of causation, we would also have to think about how to fill the epistemological gap.

The other feature I mentioned is that mechanisms are multi-level structures. We can understand this characteristic as an inference to the best explanation: the best explanation for the success of practices such as interference and detection experiments described in section 2.2 is that phenomena in the life sciences take place at different levels. That is, if we assume that mechanisms are multi-level, then such experimental practices can be naturally described as helping to break down the different levels of a system. Additionally, viewing mechanisms as multi-level is useful to understand how the different life sciences relate to each other. In cognitive science, for example, different disciplines study the same subject matter broadly speaking (i.e., cognition) using different methodological approaches, which according to the aforementioned multi-level assumption, take place at different levels (e.g., cognitive psychology studies mental processes at the organism level, whereas neuroscience studies phenomena at the nervous system level). Mutual manipulability tries to offer a general account of the discovery process of such levels by finding components and subcomponents, and takes some steps in suggesting how the results from different disciplines could be integrated to construct more robust explanations.

Finally, the idea that there are no inter-level causes is central to the mechanistic view. In particular, there is a strong way of understanding inter-level causes that makes them anti-mechanistic. In such a sense, if there are non-redundant top-down causes, then it means that they are produced by higher-level entities that are causally independent from their parts. Mechanists try to describe the world while avoiding that. This is true not only for the so called "new mechanists", but also for a more traditional view that regards nature as a machine that we can decompose and whose behavior can be fully explained in terms of its inner workings. Allowing for top-down causes in that strong sense in our explanations implies that some aspects of systems escape a characterization in terms of their structure and organization. Hence, when the problem arises, mechanists try to find ways of explaining away apparent top-down causes, making them less suspicious.

In short, interventionism and mutual manipulability tell a story about the epistemology of mechanisms qua causal multi-level structures. And a view of reality in which we can explain phenomena without appealing to mysterious inter-level causes is central to the mechanist project. The reader might disagree about how important these commitments are or should be. But I hope I have made a clear case for their tension. Also, stating the problem in this way can help to elucidate where the reader's commitments lie, even in disagreement with my proposal to dissolve the tension.

In the next section I present the proposal.

3. An Attempt to Dissolve the Inconsistency

In this section, I offer an interpretation of the relation between mechanisms and their components in which there is no inter-level causation. My strategy is as

follows. I show that in the interventionist view defined by (M), for there to be a causal relation from one mechanistic level to other, there has to be at least an ideal intervention (on one level with observable changes at the other level) that satisfies condition (M*). However, I argue that all interventions that provide evidence of constitutive relevance relations are not ideal interventions in the sense of (M*), because they are "fat-handed" interventions. I begin giving some details about how to model mechanisms in causal graphs (section 3.1). Then I move on to an explanation of inter-level relations appealing to fat-handedness (section 3.2). After that, I can present straightforwardly an argument that shows that the inter-level causation thesis (ILC) is false (section 3.3).

3.1. Representing Mechanisms in Causal Graphs

Recall that in section 2.2 I introduced the notation $\textbf{activity}_{\textbf{entity}}$ (e.g., ϕ_x) to refer to particular activities that components of a mechanism can perform. A mechanism is constituted by several entities $E = \{s, x, y, ...\}$, each of which has an associated set of activities $\{\psi, \alpha, \beta, ...\}$ (e.g., synthesize proteins, transmit action potentials, carry out long-term potentiation). There are several cases in which parts that are clearly demarcated from a physiological perspective are involved in several different activities. For example, the fusiform gyrus is associated with face recognition, but also with performance in word recognition, color identification, or within-category identification. Another example is the basal ganglia, a group of nuclei connected to several other brain regions which are associated with a variety of activities: one of them, the caudate nucleus, has been associated with language comprehension, emotions, OCD, and memory.

We can represent some aspects of mechanisms in causal graphs. My aim here is to provide a graphic representation to clarify my arguments, more than to provide a complete general framework to model mechanisms formally. First, I explain how to represent activities, entities, and part-whole relations. And in the next section I explain how to represent constitutive relevance.

First, I make one simplifying assumption: I take each $\textbf{activity}_{\textbf{entity}}$ as a *variable*, values of which refer to quantities of aspects of the activity. For example, if we are representing a neuron's firing, the variable could be its firing rate. Or if we are representing a higher-level capacity, such as face recognition, the variable could be a measure of performance in discriminating seen and unseen faces. Thus, a mechanism can be represented as a graph $G = (A_E, R)$, where A_E is the set of nodes (e.g., if x is an entity that performs two different activities ϕ and α, then ϕ_x and α_x would be different nodes in the graph), and R is a set of edges between nodes that represent causal relations. Figure 3 shows a three-level mechanism. The differences in vertical location of the variables represent different part-whole relations.

Let me make two important clarifications. First, in biological systems the relation between the variables cannot be so strict as to define a completely hierarchical structure (e.g., a tree in which every node has exactly one parent). Here is an example. Neural pathways are composed of bundles of neurons and connect different regions in the nervous system. In terms of a part-whole relation, neurons are at a lower level than pathways. Now, consider that the same neuron (e.g., p) can be connected to thousands of presynaptic and postsynaptic neurons. Hence, a single neuron can be part of different pathways (e.g., x and y) that are

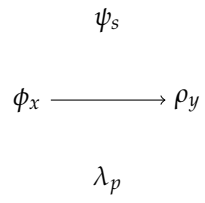

Figure 3: Mechanism in Causal Graph

not hierarchically related.

Second, someone might be worried that the assumption of taking an $\textbf{activity}_{\textbf{entity}}$ to be a variable sweeps a metaphysical problem under the carpet, because it says nothing about the ontology of the entities. I think that while this might be true for the interventionist (who defines causation as a relation between variables), it is not true for the mechanist. The mechanist, I take it, is committed to the assumption that there is a fact of the matter about what constitutes a part in a biological system. Such systems have extension and location in space, and therefore they are physical entities, not mental, abstract, or functional. This is implicit in the representation, because we can derive a part-whole structure from the graph of a mechanism. For example, take the graph in Figure 3. Let the subscripts defining the entities be the nodes (i.e. remove the greek letters defining activities), remove causal connections, and connect each node to their immediate upper level. Let the edges of this new structure represent an irreflexive, asymmetric, and transitive relation. The resulting graph would represent the part-whole relationships between the entities of the mechanism (basic mereology uses reflexivity instead of irreflexivity as a primitive notion, but in Biology it is unusual to say that something is a part of itself). The relation defines a strict partial order, which implies that one node (e.g., p) could be descendant of two other nodes (e.g., x and y). Two clarifications: first part-whole relations are one condition for (and shouldn't be identified with) constitutive relevance relations (see MM2); second, the fact that the variables can be organized as a strict partial order is not sufficient to conclude that they stand for physical entities, but it is necessary constraint.

Nonetheless, being committed to the existence of parts does not thereby commit the mechanist to a particular view on the organization of neural systems. For example, some philosophers and scientists argue that the way in which we currently classify some brain areas as 'parts' is mistaken, and therefore we should focus on networks rather than localized structures. If they are right, however, the mechanistic view would not be necessarily threatened. In this scenario, mechanisms would have to be decomposed in different units (e.g., clusters of structures such as neurons), but would still be subject to part-whole relations.

A mechanist can represent results of her enterprise as a graph G. Procedures (e.g., staining tissue) can help to identify entities and part-whole relations (strictly speaking, these procedures are not interventions on the mechanism as defined by (MM1), because they are intended to identify structure and not manipulate activities). Other entities are observable with (or inferred based on) the same procedures that lead to the discovery of the relations between activities (e.g., the visual system contains multiple pathways that go from the optic nerve to visual cortex in the back of the brain, which are identifiable by dissociating

their functions). The graph is constructed by intervening experimentally in A_E and observing dependencies. Consider the example in Figure 4. Suppose x, y, and z are parts of s. Intervention I manipulates the value of ϕ_x. The intervention blocks the influence from γ_z to ϕ_x (satisfying condition 2 in (M*)). This allows us to infer whether ϕ_x has a causal influence on ρ_y ruling out spurious correlations between ϕ_x and ρ_y due to γ_z.

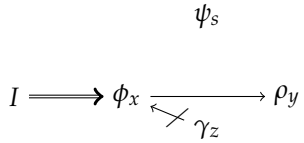

Figure 4: Interventions

I turn now to the question of how to represent constitutive relevance.

3.2. Inter-level Relations and Fat-handedness

We can represent constitutive relevance (section 2.2) as follows. Let I_1 be some experimental intervention on ϕ_x, and suppose we observe that the value of ψ_s changes to ${\psi_s}^*$. Also, let I_2 be an experimental intervention on ψ_s, and suppose we observe that the value of ϕ_x changes to ${\phi_x}^*$. This is shown in Figures 5(a) and 5(b) respectively.

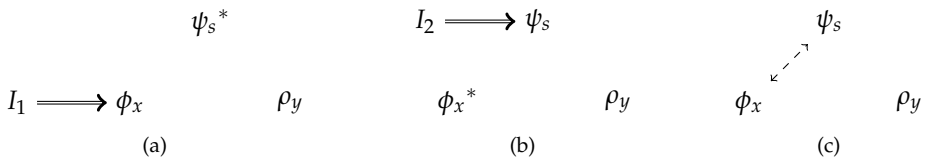

Figure 5: Mutual manipulability

A prima facie conclusion from each of these experiments is that there are inter-level causal relationships (which should be depicted by causal arrows, as in Figure 4). This is, however, the problem that we want to avoid. I proceed in two steps. I first look at what we are required to account for if we want to avoid such relations, which leads to some changes in the graph representation and definitions. And second, I argue that we are justified in making such changes on metaphysical grounds.

Suppose we deny that the interventions I_1 and I_2 and corresponding changes in values for ψ_s and ϕ_x mean that there are inter-level causal relations between ψ_s and ϕ_x. Something that we cannot deny, however, is that these experiments succeed in providing correlational data between the two variables. To represent correlations in causal graphs, a standard is to use dotted bidirectional arrows. This is shown in Figure 5(c). I represent inter-level relations with correlational arrows (but of course, not all correlational arrows stand for inter-level relations).

We have to explain this correlation, because in a causal system there is no correlation without causation (i.e., if there is no causation between the correlated

events, then there are hidden factors exerting causal influences). This idea is sometimes referred to as Reichenbach's *common-cause assumption*. In the example the assumption is as follows:

(R) If there is a correlation between ϕ_x and ψ_s, then (a) ϕ_x is a cause of ψ_s, or (b) ψ_s is a cause of ϕ_x, or (c) there is a common cause for ϕ_x and ψ_s.

According to (R), a causal system that accounts for the correlation of ϕ_x and ψ_s has to satisfy *at least* one of the three disjuncts. Disjuncts (a) and (b) should be rejected given that we want to avoid inter-level causation. Hence, the only option is to accept disjunct (c).

A common cause for two events can be understood as a "fat-handed" intervention. Woodward defines fat-handed interventions as those "affecting not just X and other variables lying on the route from I to X to Y, but also other variables that are not on this route and that affect Y" (Woodward, 2008a, p.209). In the context of the quotation, he does not intend to offer a formally precise definition, but I interpret his idea in terms of the definitions given in section 2.1 as follows: if I is a fat-handed intervention, then I does not satisfy either condition 3 or condition 4, but satisfies condition 1 and condition 2 in (M*).

Some straightforward examples of fat-handed interventions are those in which the intervention does not have a localized effect on one single variable (i.e., it changes several simultaneously) because the intervening instrument is not precise enough. In medicine, treatments with side effects have that characteristic: chemotherapy drugs, for example, may kill cancer cells effectively, but they also destroy healthy cells. Another example is antidepressants that have an effect on serotonin levels, but also lead to weight gain. In these cases, it is possible to imagine interventions that could have a more localized effect on a target variable, avoiding the unwanted side effects.

Fat-handed interventions deviate from the normative ideal, but this does not mean that they don't provide useful information. Often they are done in exploratory stages, and give insights to develop more refined experiments. And sometimes they change variables that are not relevant to the current study. Hence, as Woodward acknowledges (Woodward, 2004), it would be too strong to say that prediction is only attainable through ideal interventions, because accumulating evidence from several fat-handed interventions could also provide it.

Now, it is possible to quantify the same physical event in different ways, using different variables. In such contexts, an intervention that manipulates one of such variables *necessarily* manipulates the other. This, I think, is the right way of thinking about interventions on variables that quantify different aspects of a part-whole relationship. That is, in the part-whole case, interventions are necessarily fat-handed: the intervention on the variable related to a part is also necessarily an intervention on the variable related to the whole, and vice versa. We should not conceive the intervention as being on one variable or the other, but on both at once. The difference with the examples mentioned before (i.e., treatments with side-effects), is that in the part-whole case (and therefore in the mechanistic case), it is not possible that a more precise instrument would make the intervention not fat-handed. Both cases are structurally the same, however, in the sense that there are two variables that change simultaneously.

More precisely, in the part-whole case, condition 3 in (M*) is violated. An intervention I causes ϕ_x (condition 1), and suppress other causal influences on ϕ_x

(condition 2), but does not satisfy condition 3, because it directly causes ψ_s (more on this below). This is sufficient to render the intervention non-ideal. There is another case of a non-ideal intervention in the vicinity. Imagine a case such as the one depicted in Figure 2(c), but in which X is a higher-level variable and Z and Y are lower level variables. In such a case, condition 4 is violated because I and Z are correlated (I causes Z), and therefore the intervention is non-ideal. The important case for the present discussion, however, is when condition 3 is violated.

In short, I think the concept of fat-handed intervention is not only a logical alternative available to understand inter-level relations in an interventionist framework. Also, it is one that we are justified in adopting, as I argue next.

I propose the following working definition of the relation between constitutive relevance and interventions:

(F) If ϕ_x is a component in ψ_s, then (i) any intervention I on ϕ_x is also necessarily an intervention on ψ_s, and (ii) any intervention I on ψ_s is also necessarily an intervention on ϕ_x.

This is shown in Figure 6.

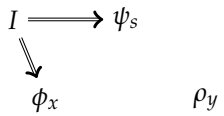

Figure 6: Components and Fat-Handedness

(F) is a consequence of what I take the right metaphysical picture of mechanisms to be: a mechanism's activity and its components arranged and working in the right way are the same physical event. As such, an intervention on a component changes both the component and the whole. And an intervention on the whole changes both the whole and *at least one* of its components. Nonetheless, even though the intervention counts as a single physical event, we can measure its effects either on the whole mechanism or the components using different techniques (i.e., techniques that have different resolutions), and register them in changes in variables at different levels. Thus, ψ_s is a higher-level variable because some lower-resolution technique measures the state of the whole mechanism, and ϕ_x is a lower-level variable because some other higher-resolution technique measures the state of one component.

This does not necessarily mean that detection techniques are the only criterion to individuate mechanistic levels. Composition is also required: if ϕ_x and ρ_y are components in the same mechanism, and we can measure changes in their values with the same technique (or techniques with the same resolution), then they are at the same mechanistic level (and the converse is not necessarily true).

I think previous arguments by (Craver, 2007) and Craver and Bechtel (Craver & Bechtel, 2007) against inter-level causation (see section 2.3) require the same metaphysical picture, but the authors don't make that commitment explicit. Furthermore, their arguments are incomplete as a solution to the problem because they don't explain how the underlying metaphysical picture relates with the interventionist framework in a way in which the problems in section 2.3 don't arise.

How does this metaphysical picture fit with the interventionist view of causation? In the interventionist's semantics, causation is not a relation between entities but between variables (this is also true in the causal graph representation). Hence, to represent a change in a (variable for a) mechanism and one (variable for one) of its components that results from the same event, we need to use two simultaneous causal arrows on the two variables from the intervention. And this is precisely what (F) expresses. If (F) is right, then (R) is true, because (F) makes disjunct (c) true. In other words, the metaphysical picture implies that the intervention changes the two variables at the same time. In the interventionist semantics, this means that the relation between the two variables cannot be causal, because saying that it is causal means that there is an ideal intervention on one of them that has an effect on the other, but this cannot be the case, because fat-handed interventions are not ideal.

Let me illustrate this idea with two examples. The first shows the general idea. The second has a more mechanistic and biological feel.

Example 1. This is an example based on an example from Sperry (1980). You push a tire (that is the intervention I), the tire moves (a change in ψ_s measured in terms of distance) and its rubber molecules also move (a change in ϕ_x measured in terms of rotation of the molecules). You observe a correlation between the two variables. However, there is no way of making the intervention without producing both effects. The intervention itself causes both. The marcrophysical properties of the tire depend on the properties of its rubber molecules, but this does not imply that the high-level variable is reducible to the lower-level variable, since rolling is a property of the tire as a whole, not of individual molecules.

Example 2. Consider a rough description of how aspirin works. You take an aspirin (I) and it relieves your pain (a change in ψ_s measured perhaps by a subjective report about the intensity of your pain on a scale). Aspirin suppresses the production of prostaglandins, which are associated with transmission of pain information to the brain. More specifically, aspirin inhibits cyclooxygenases COX-1 and COX-2, two enzymes required for the synthesis of prostaglandins (suppose this is a change in ϕ_x). Now, it is not possible for the aspirin to relieve your pain (higher-level variable) *without* suppressing the production of prostaglandins (lower-level variable), or the other way around (controlling for other influences). Your intervention at the upper level (i.e., you taking the aspirin) is also necessarily an intervention at the lower level. For a longer discussion about this mechanism see (Woodward, 2011, p.418–419).

Some clarifications. First, mutual manipulability between two variables alone does not entail that any intervention on one of them is a fat-handed intervention: two variables in a feedback relation might be mutually manipulable and they could still be subject to ideal interventions.

Second, (F) provides a necessary, but not sufficient condition. Any time we have a fat-handed intervention between two variables we cannot infer that we are talking about a constitutive relevance relation in a mechanism. As mentioned before when discussing treatments with side-effects, some fat-handed interventions are so because they aren't refined enough to have a localized effect in one variable only. Other fat-handed interventions, however, are so because they target variables for both a mechanism and its components. This latter type is the one that explains (away) inter-level causal relations.

Third, fat-handed interventions such as I in Figure 6 may affect several vari-

ables (more than two) at two levels simultaneously. For instance, since ρ_y is also a component in ψ_s, by (F), I is also an intervention on ρ_y. I don't show the arrow that represents that causal relation for simplicity. One practical limitation of the representation is that the diagrams get cumbersome even with a small number of variables.

At this point, the reader might have an idea of how the argument against the inter-level causation thesis looks. In the next section I make it explicit.

3.3. Argument Against Inter-Level Causation in a Mechanism and Incoherence Revisited

Using (F) it is possible to formulate an argument against inter-level mechanistic causation straightforwardly. Let ψ_s and ϕ_x be two variables such that ϕ_x is at a lower mechanistic level than ψ_s. Suppose for reductio that ψ_s causes ϕ_x (i.e., top-down causation). Hence, according to (M), there is an ideal intervention I on ψ_s that will change the value of ϕ_x when all variables other than ϕ_x and ψ_s are fixed at some value. By condition 1 in (M*), I causes ψ_s. And by condition 3 in (M*), I does not directly cause ϕ_x. Since ϕ_x is at a lower mechanistic level than ψ_s, ϕ_x is a component in ψ_s by (L). Therefore, by (F), I is also necessarily an intervention on ϕ_x. Hence, I causes ϕ_x. *Contradiction*. Hence, ψ_s does not cause ϕ_x, which means that there is no top-down causation. A similar argument inverting the order of the variables shows that there is no bottom-up causation. Therefore, there is no inter-level causation between ϕ_x and ψ_s.

It is worth stressing that the argument not only shows that interventions do not provide evidence of inter-level causal relations, but also that they cannot provide such evidence. All the interventions that discover constitutive relevance relations are not ideal interventions in the sense of (M*).

Let me discuss briefly the implications of this argument for the problems in section 2.3. If my argument is right, then thesis (ILC) is false, so the problems of cyclicity and redundancy don't get off the ground.

1. *Redundancy Problem Revisited*. In my proposal there are no violations of the closure principle (i.e., the idea that for every physical property P, there is a set of sufficient physical causes P*) given that all active entities are physical. Therefore, all causes and effects are physical. Questions such as "what is the most fundamental level in causal terms?" or "what is the level in which causation is really taking place?" are misleading. Assuming interventionism, causation is well-defined at different levels. Causal relations occur at different levels because at different levels we can find some changes in variables that produce changes in other variables. On the other hand, since there is no top-down causation there are no problems of higher-level entities exerting redundant causal influences on lower-level entities.

2. *Cyclicity Problem Revisited*. While there are cycles of correlations between levels, there are no causal cycles between a mechanism and its components. Take ϕ_x and ψ_s such that there is causal cycle between them. It follows that ϕ_x and ψ_s are correlated, ϕ_x causes ψ_s, and ψ_s causes ϕ_x. Given that ϕ_x causes ψ_s, then by (M) there is an ideal intervention I on ϕ_x that will change the value of ψ_s when all variables other than ϕ_x and ψ_s are fixed at some value. Since I is ideal, it does not cause ψ_s. Hence, by the contrapositive of

(F), ϕ_x is not a component in ψ_s. This case could be interpreted as a case of feedback between two different mechanisms.

Now I will discuss four remaining potential worries about my proposal.

1. Does the proposal really preserve Craver's mutual manipulability theory? My proposal calls for a reinterpretation of Craver's formulation of mutual manipulability. He defines mutual manipulability in terms of *ideal* interventions (Craver, 2007, p. 154), an aspect of the theory that I stated in (MM1) in section 2.2. But if (F) is correct, then interventions in inter-level experiments are never ideal in the interventionist sense. Does the mutual manipulability theorist have to be committed to the interventionist notion of ideal intervention? I don't think so. To preserve the insight of the theory, the mutual manipulability theorist requires a notion of intervention that does not include (I1c) in section 2.2. Such a notion is weaker than the interventionist notion. It is beyond the scope of this paper, however, to provide a fully worked-out alternative.

2. Does the argument prove that there is not inter-level causation at all? One can be worried that the argument I have provided proves too much, ruling out inter-level causation in all cases. In response, I think it is worth stressing that in order for the argument to work, we need the two entities in question to be in a constitutive relevance (i.e., compositional) relation. If that is not the case, then it is not possible to appeal to definition (F). Hence, the argument concerns the issue of inter-level causation within a single mechanism, and does not say anything about the inter-level causation between components of different mechanisms. The puzzling issue is, however, the former. Inter-level causation is problematic in cases in which there is a compositional relation between entities in the levels that we consider, such as within a mechanism. To address the latter issue, we would need a notion of levels different from the notion used here (i.e., levels have local significance, and the predicate "x is at a lower level than y" is defined only within a mechanism, as discussed in section 2.2).

3. Can the proposal account for undetectable changes? Recall that lesion experiments suggest that the fusiform gyrus is relevant in the mechanism for face recognition. But presumably, if you destroy one single neuron in the fusiform gyrus, you won't observe a change in face recognition performance. In such a case, it seems that we have a case of composition, and a fat-handed intervention that does not bring about effects in both high-level and low-level variables. I think, however, that this is not a case of composition. The fact that we can create a sorites paradox, because at some point destroying n neurons would result in an observable change, suggests that we can treat "being constitutively relevant" as a vague predicate. We have clear cases at both ends, and borderline cases in between. The activation of an individual neuron is not constitutively relevant for face recognition, but the variable that measures the activation of, say, n neurons is. Mutual manipulations try to identify the clear cases, and I think (F) works for those.

This objection is also avoided by (MM2) being a sufficient but not a necessary condition for constitutive relevance.

4. Does the distinction between fat-handed and ideal interventions depend on what we are interested in? In my proposal, whether an intervention is fat-handed or not depends on the context, and this could seem worrisome from a metaphysical perspective. Here is an example of the worry: suppose we are working with

variables for activities at one level (e.g., cellular), and we identify an ideal intervention I on variable ϕ_x with respect to ρ_y. Now, suppose we recognize a lower level (e.g. molecular), and we observe that a variable λ_p at such level is a component in ϕ_x. This implies that I is also a cause of λ_p, which raises the question, is I really an ideal intervention or really a fat-handed intervention? I don't think that this sort of context dependence has problematic metaphysical implications. It would be worrisome if, given a set of variables to work with, it could turn out that I could be seen as ideal or fat-handed (i.e., because the question of whether there is a causal relation or not would not be settled). But once we fix the set of variables, interventions are of one kind only (i.e, ideal or fat-handed). This response is partial, however, because it is fair to say that it pushes the problem back to another one: how do we identify the appropriate set of variables to study a phenomenon? This is an important open problem that has a wider reach than my proposal: there is no complete theory about what being well-defined variable amounts to. And interventionism, the mutual manipulability theory, and my proposal presuppose a set of well-defined variables to work with.

As I suggested, giving up the triad has theoretical costs, because the three views integrate important concepts in the mechanisms literature, such as level, constitutive relevance, and causation. At this point, however, some readers might not be entirely persuaded by my dissolution attempt, and could still think that the triad is indeed inconsistent. Hence, before concluding, I will briefly discuss the alternatives available to try to solve the inconsistency by rejecting one of the three views.

4. Solving the Inconsistency by Rejecting one of the Three Views

4.1. Weaken or Reject the Interventionist View

One way around the triad is to weaken interventionism. The theory, however, is open to several interpretations, so how we can weaken it would depend on that. In *Making Things Happen*, Woodward stresses emphatically that his project is "semantic" (2003, p.28). This, however, does not mean that he intends to offer a conceptual analysis of "cause" and its uses in ordinary and scientific contexts, uses that he regards as sometimes unclear and ambiguous. That is, even though his project attempts to account for familiar intuitions about causal explanations, it is not merely descriptive; it also has a normative component: it makes "recommendations about what one ought to mean by various causal and explanatory terms" (Woodward, 2003, p. 7).

Nonetheless, it is possible to read interventionism as a theory that has broader epistemological and metaphysical implications. The epistemological reading I think is natural, because (M) establishes a connection between controlled experiments and the concept of causation. Nonetheless, interventionism has to be distinguished (as Woodward himself acknowledges) from the project of inferring and discovering causal relations from statistical data (Pearl, 2000), a project that can also be characterized as epistemological.

Now, whether we can read interventionism as a metaphysical theory is a more contentious issue. Woodward tries not commit himself to such a reading in *Making Things Happen*. He says "I leave it to the reader to decide whether [interventionism] counts as discovering 'what causation is' " (Woodward, 2003, p. 7). More recently, however, his stance is less neutral, and he as defended interven-

tionism as "a contribution to methodology rather than a set of theses about the ontology or metaphysics of causation" (Woodward, 2014).

I think the fact that the original formulation of the theory is silent about metaphysics, leaves open the possibility to use the theory in contexts that are traditionally regarded as the domain of metaphysics, something that he and others have done. In particular, interventionism has been applied in discussions about the causal exclusion arguments advanced by Kim (see Woodward, 2008b; Raatikainen, 2010). Such a metaphysical reading raises worries for philosophers interested in the metaphysics of causation. Glennan, for example, says that interventionism makes "an important point about the epistemology of causation" because "experimental manipulations can provide evidence that variables are connected, even in the absence of mechanical knowledge of how they are connected" but "this epistemologically important point does not legitimate [interventionism] as a metaphysical account of causation"(Glennan, 2009, p. 318). I think the root of such metaphysical worries is the fact that interventionism explains causation as a relation between variables, rather than a relation between objects. For example, in some metaphysical pictures, causation is a relation between entities at the fundamental physical level of reality (Heil, 2003), which excludes many of the relations that the interventionist talks about.

Having these metaphysical worries in mind, one might be inclined to weaken interventionism arguing that the existence of an ideal intervention is not a sufficient condition for a causal relation. That is, we could weaken interventionism by breaking the biconditional expressed by (M) in section 2.1, accepting only the necessity claim. Roughly: if X is a direct cause of Y, then there is (in principle) an ideal intervention on X that produces a change in Y (while keeping everything else constant), but not the converse. This solution, however, has an epistemological cost, because it implies that the relationship between X and Y might not be causal, even if there is an ideal intervention on X that has an effect on Y. In other words, even if one doubts interventionism as a complete story about the nature of causation, from an epistemological point of view, it is the sufficiency part of (M) that matters the most.

A second alternative is to simply reject the semantic and metaphysical readings of interventionism. To do this, one could draw a distinction between *causation* and *causal explanation*, and argue that interventionism is not a theory of causation but a theory of causal explanation, where causal explanation is strictly an epistemic notion. This would require replacing all the instances of "cause" with "causal explanation" in the formulation of interventionism. Roughly, for instance, (M) would be: X is a causal explanation of Y iff there is (in principle) an ideal intervention on X that produces a change in Y. I think such a change would eliminate the tension with Craver's mutual manipulability view, because the sort of inter-level relations implied by the modified theory would not be causal, and therefore would not have to be considered a metaphysical problem. Nonetheless, if we separate the two projects, we would then have to adopt (or develop) a theory about the nature of causation, and after having such a theory, we would still have to answer the question about the relation between causal explanations and genuine causal relations. This alternative might not seem so problematic for supporters of other theories of causation, but we would have to assess whether such theories preserve the virtues of interventionism in other contexts.

If we don't reject interventionism, another alternative is to weaken or reject

mutual manipulability.[3]

4.2. Weaken or Reject the Mutual Manipulability Account

Craver's critics (Leuridan, 2012) have suggested rejecting mutual manipulability on the grounds that, given its use of Woodward's interventionism, it seems to entail that inter-level relations are causal, a consequence that we have independent reasons to reject. This creates a problem, for we would need an account of constitutive relevance to replace mutual manipulability. However, one might ask, why would we want an account of constitutive relevance in the first place? Someone, perhaps with reductionist leanings, could argue that there is something wrong in the very idea of constitutive relevance. I won't explore such avenue here, because I'm assuming the tenet of mechanistic philosophy, according to which mechanisms are multi-level structures, whose description involves identifying components. If one accepts this, I think dismissing mutual manipulability on the grounds of the aforementioned tension would be too quick. As I said before, I think it is important to preserve the fact that mutual manipulability articulates a connection between very common kinds of experiments in the biological sciences (i.e., top-down and bottom-up experiments) and the task of describing mechanisms.

I think the insight of mutual manipulability is primarily epistemic. The reason is that mutual manipulability, as stated in (MM2), does not offer a necessary condition for constitutive relevance, which suggests that it is not a theory about the nature of the constitutive relevance relation. Having that in mind, and the fact that (MM2) is a sufficient condition, we can congenially understand the theory as an attempt to characterize practices that are sufficient for scientists to discover constitutive relevance relations, and is also silent about the nature of such relations.

There are other alternative accounts, which perhaps address more directly the question about the nature of the constitutive relevance relation. I don't intend to offer full criticisms of them, but I will mention them briefly. The first alternative is to explain constitutive relevance in terms of density of relations between components (Haugeland, 1998, p. 215). According to this alternative, given a mechanism ψ_s, the interactions between a component ϕ_x and the rest of components of ψ_s are greater than the interactions between ϕ_x and other ρ_y that are not components in ψ_s. This alternative, however, does not allow us to distinguish genuine components from mere correlates, because the latter might be as densely connected to the mechanism as the former (Craver, 2007, p.142–144).

More recently, Couch (2011) proposed another alternative account of constitutive relevance appealing to Mackie's notion of *INUS* conditions. That is, to "define a relevant part as an insufficient but nonredundant part of an unnecessary but sufficient mechanism that serves as the realization of some capacity." However, this proposal, as Couch himself acknowledges, "provides an explanation of what the relevance relation is, and not of the evidence we have for it" (Couch, 2011), so if it goes through, we would still need a further story about the epistemology of this relation.

Another recent alternative, along with other reasons to reject mutual manipulability, is proposed by Fagan (2012). She regards mechanistic explanations as

[3]Thanks to an anonymous reviewer for *Synthese* for pushing me to clarify this section.

bottom-up, and appeals to a notion of "jointness", which emphasizes the "interdependence among causally active components of a mechanism", rather than modularity of components. In her view, the joint activity of a complex of components is the explanans of higher-level phenomena (Fagan, 2012, pp.463–468).

4.3. Weaken or Reject the Incoherence of Inter-Level Causation

Finally, we could preserve mutual manipulability and interventionism in its semantic reading, at the cost of rejecting the incoherence of inter-level causes. I think there are two ways of doing this. First, one could argue that there is no fundamental distinction between causation and composition, perhaps on the grounds that both relations can be ultimately modeled using the same formal tools (e.g., structural equations). Schaffer, for instance, says that "causation links the world across time, [and] grounding [composition] links the world across levels" (Schaffer, 2012). However, most philosophers would be uncomfortable with such a conclusion.

A second, perhaps less controversial way, is to embrace inter-level causation. One recent attempt to do this in the context of mechanisms, precisely to address problems in Craver's mutual manipulability theory is (Harinen, in press). Strong emergence theorists, for instance, might find this solution appealing. If higher-level mechanisms can have properties that are non-reducible to the organization of their components, it is possible to conceive them as causes of the behavior of such components.

I won't provide here a more detailed analysis of these options, but I want to stress that solving the problem by giving up any view in the triad has important epistemological and metaphysical costs. In this paper, I have proposed a way of dissolving the problem, rather than solving it. And if I'm right, the costs of rejecting any one view can be avoided.

5. Conclusion

Woodward's interventionist theory of causation gives necessary and sufficient conditions for there to be causal relations. I have shown that such a view, supplemented by the mutual manipulability account of constitutive relevance, implies problematic theses about inter-level causation. Prima facie, we can avoid this problem either by rejecting interventionism or mutual manipulability, or by embracing inter-level causation. I have tried to articulate, however, one plausible interpretation of interventionism, mutual manipulability and the incoherence of inter-level causation on which there is no tension between them. I have shown a way of explaining inter-level mechanistic relations without appealing to inter-level causal claims. On this view, fat-handedness provides a metaphysically sound way of thinking about interventions in mechanisms. The correlations between what takes place at different levels are explained by the fact that the mechanism and its components (organized and working in the right way) are the same entity. And as such, causal influences are exerted on the the different levels simultaneously. In this way, there are no inter-level causal relations between mechanisms and components. Furthermore, there can't be such relations, because there are not ideal interventions that account for them, in the way that the interventionist view requires.

References

Baumgartner, M. (2009). Interventionist Causal Exclusion and Non-Reductive Physicalism. *International Studies in the Philosophy of Science*, *23*(2), 161–178.

Bechtel, W. (2011). Mechanism and Biological Explanation. *Philosophy of Science*, *78*(4), 533–557.

Bruce, V., & Young, A. (1986). Understanding Face Recognition. *British Journal of Psychology*, *77 (Pt 3)*, 305–327.

Couch, M. B. (2011). Mechanisms and Constitutive Relevance. *Synthese*, *183*(3), 375–388.

Craver, C. F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience.* Clarendon Press: Oxford.

Craver, C. F., & Bechtel, W. (2007). Top-Down Causation Without Top-Down Causes. *Biology and Philosophy*, *22*(4).

Damasio, A. R., Damasio, H., & Van Hoesen, G. W. (1982). Prosopagnosia. *Neurology*, *32*(4), 331.

Eberhardt, F., & Scheines, R. (2007). Interventions and Causal Inference. *Philosophy of Science*, *74*(5), 981–995.

Fagan, M. B. (2012). The Joint Account of Mechanistic Explanation. *Philosophy of Science*, *79*(4), 448-472.

Franklin-Hall, L. R. (in press). High-Level Explanation and the Interventionist's 'Variables Problem'. *British Journal for the Philosophy of Science*.

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

Glennan, S. (2009). Mechanisms. In *The Oxford Handbook of Causation* (pp. 315–325). Oxford University Press.

Harinen, T. (in press). Mutual Manipulability and Causal Inbetweenness. *Synthese*, 1-20.

Haugeland, J. (1998). *Having Thought: Essays in the Metaphysics of Mind*. Harvard University Press.

Heil, J. (2003). *From an Ontological Point of View*. Oxford University Press.

Ishai, A., Schmidt, C. F., & Boesiger, P. (2005). Face Perception is Mediated by a Distributed Cortical Network. *Brain Research Bulletin*, *67*(1–2), 87 - 93.

Kim, J. (1989). Mechanism, Purpose, and Explanatory Exclusion. *Philosophical Perspectives*, *3*, 77–108.

Kim, J. (1999). Making sense of emergence. *Philosophical Studies*, *95*(1-2), 3–36.

Leuridan, B. (2012). Three Problems for the Mutual Manipulability Account of Constitutive Relevance in Mechanisms. *British Journal for the Philosophy of Science*, *63*(2), 399–427.

Levy, A. (2013). Three Kinds of New Mechanism. *Biology and Philosophy*, *28*(1), 99–114.

Machamer, P. K., Darden, L., & Craver, C. F. (2000). Thinking About Mechanisms. *Philosophy Of Science*, *67*(1), 1–25.

Meng, M., Cherian, T., Singal, G., & Sinha, P. (2012). Lateralization of face processing in the human brain. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1735), 2052-61.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Raatikainen, P. (2010). Causation, Exclusion, and the Special Sciences. *Erkenntnis*, *73*(3), 349–363.

Schaffer, J. (2012). Grounding, Transitivity, and Contrastivity. In Correia & Schnieder (Eds.), *Grounding and explanation* (pp. 128–138). Cambridge.

Semmelweis, I. (1983). *The Etiology, Concept, and Prophylaxis of Childbed Fever*. (Translated by Carter, K.C.) University of Wisconsin Press.

Sperry, R. W. (1980). Mind-brain interaction: Mentalism yes, dualism no. *Neuroscience*, *5*(2), 195–206.

Tsao, D. Y., Freiwald, W. A., Tootell, R. B., & Livingstone, M. S. (2006, 3). A cortical region consisting entirely of face-selective cells. *Science*, *311*(5761), 670–674.

Waskan, J. (2011). Mechanistic explanation at the limit. *Synthese*, *183*(3), 389–408.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanations*. Oxford University Press.

Woodward, J. (2004). Counterfactuals and Causal Explanation. *International Studies in the Philosophy of Science*, *18*(1), 41–72.

Woodward, J. (2008a). Invariance, Modularity, and all That. In S. Hartman, C. Hoefer, & L. Bovens (Eds.), *Nancy Cartwright's Philosophy of Science* (pp. 198–237). Taylor & Francis.

Woodward, J. (2008b). Mental Causation and Neural Mechanisms. In *Being reduced: New essays on reduction, explanation, and causation.* Oxford University Press.

Woodward, J. (2011). Mechanisms Revisited. *Synthese*, *183*(3), 409–427.

Woodward, J. (2014). Interventionism and Causal Exclusion. *Philosophy and Phenomenological Research*, 1–45.