Rendering Interventionism and Non-Reductive Physicalism Compatible

Michael BAUMGARTNER*

ABSTRACT

In recent years, the debate on the problem of causal exclusion has seen an 'interventionist turn'. Numerous non-reductive physicalists (e.g. Shapiro and Sober 2007) have argued that Woodward's (2003) interventionist theory of causation provides a means to empirically establish the existence of non-reducible mental-to-physical causation. By contrast, Baumgartner (2010) has presented an interventionist exclusion argument showing that interventionism is in fact incompatible with non-reductive physicalism. In response, a number of revised versions of interventionism have been suggested that are compatible with non-reductive physicalism. The first part of this paper reconstructs the definitional details of these modified interventionist theories. The second part investigates whether the modification proposed in Woodward (2011) is not only compatible with, but moreover supports non-reductive physicalism. In particular, it is examined whether that newest variant of interventionism allows for empirically resolving the problem of causal exclusion as envisaged by Shapiro, Sober and others.

1 Introduction

Arguments of causal exclusion, as most famously advanced by Kim (1989; 2003; 2005), aim to expose a tension within the position of non-reductive physicalists who endorse the following claims: (i) the domain of the physical is causally complete, (ii) biological, mental, and other macro properties non-reductively supervene on physical properties, and (iii) biological, mental, and other macro properties may be causes of physical effects of their own supervenience bases. In a nutshell, the tension exposed by exclusion arguments exists between the first two claims and the third: it is unclear how something that is not reducible to the physical domain, i.e. something ultimately nonphysical, could causally affect the domain of the physical, even though the latter is causally complete.

In recent years, the debate on the conclusions to be drawn from arguments of causal exclusion has seen an 'interventionist turn'. On the one hand, numerous non-reductive physicalists have argued that an interventionist theory of causation—to which I shall henceforth refer as *interventionism*—as most exhaustively developed in Woodward (2003) immunizes non-reductive physicalism against exclusion

*Institute of Cognitive Science, University of Osnabrück, Albrechtstrasse 28, 49076 Osnabrück Germany; Email: michael.baumgartner@uni-osnabrueck.de

arguments. Authors as Shapiro and Sober (2007), Menzies (2008), Shapiro (2010; 2012), or Raatikainen (2010) claim that interventionism accounts for downward causal dependencies among supervening biological or mental properties and effects of their supervenience bases, in spite of the causal completeness of the physical and the non-reducibility of corresponding macro properties. Moreover, they contend that the interventionist framework provides a means to empirically identify the physical effects of biological or mental properties. Hence, the idea is that interventionism—the theory of causation that, unlike any other account currently available, establishes a tight connection between causation and experimental scientific practice—allows for an evidence-based resolution of the exclusion problem, for which solutions have traditionally been searched on the basis of metaphysical (armchair) argumentation only.[1]

On the other hand, in Baumgartner (2009; 2010) I have taken issue with these claims (cf. also Marcellesi 2010). I argue that, far from securing non-reductive physicalism against the threat posed by exclusion arguments, the interventionist theory of causation gives rise to a self-contained *interventionist* exclusion argument, which even rests on weaker premises than Kim's arguments. More specifically, I show that tenets (i) and (ii) of non-reductive physicalism in combination with an interventionist understanding of causation entail the negation of (iii), i.e. that interventionism and non-reductive physicalism are incompatible on conceptual grounds.

This, in turn, has provoked reactions from proponents of interventionism who, even if they do not themselves subscribe to non-reductive physicalism, hold that interventionism should not exclude non-reductive physicalism on purely conceptual grounds. Woodward (2011), Eronen (2012), or Weslake (unpublished) make different suggestions for adaptations of the definitional core of interventionism that render interventionism and non-reductive physicalism compatible. These newest variants of interventionism no longer permit the inference from causal completeness and non-reductive supervenience, i.e. from (i) and (ii), to the impossibility of downward causation, i.e. to the negation of (iii). However, non-reductive physicalists who advance interventionist solutions to the problem of causal exclusion—as Shapiro, Sober, Menzies, or Raatikainen—do not merely want to settle for a variant of interventionism that does not entail the negation of (iii). Rather, they require a variant that theoretically supports (iii) and relative to which it becomes possible to produce empirical evidence for the existence of downward causation.

Accordingly, this paper investigates how well those recently proposed adaptations of the definitional core of interventionism fare with respect to theoretically grounding the possibility of non-reductive downward causation and to empirically

[1] Note that Woodward himself has never explicitly subscribed to non-reductive physicalism, nor has he ever claimed that interventionism allows for an evidence-based resolution of the problem of causal exclusion. Furthermore, apart from Shapiro, Sober, Menzies, and Raatikainen, there are numerous other authors who work under the assumption that the possibility of non-reductive downward causation can be established on interventionist grounds without explicitly arguing the point themselves, cf. e.g. Reisman and Forber (2005) or Campbell (2007).

solving the problem of causal exclusion. The aim is neither to defend nor to reject non-reductive physicalism or interventionism, but rather to lay open the relationship between these two theoretical frameworks, and thereby to examine the prospects of successfully backing up non-reductive physicalism by interventionist means. As will become apparent quickly, definitional details will be crucial for our discussion. In fact, I suspect that one of the main reasons why so many authors believe that interventionism can be put to use when it comes to grounding the possibility of non-reductive downward causation is that they rely on a merely intuitive understanding of the basic ideas behind interventionism without taking the definitional foundation of that theory at face value. For that reason, I will discuss all the relevant definitions in detail and extract their pertinent consequences—even though some of these definitions are frequently cited in the literature.

The paper is organized as follows. Section 2 exhibits the core tenets of non-reductive physicalism and interventionism and briefly reviews my incompatibility argument (Baumgartner 2009; 2010). Section 3 reconstructs the definitional details of the modified interventionist theories proposed in the literature and analyzes what these theories conceptually entail for the possibility of downward causation. Finally, in section 4, I assess the relationship between the newest version of interventionism advanced in Woodward (2011) and non-reductive physicalism, and I investigate whether that theory indeed paves the way for an evidence-based resolution of the problem of causal exclusion.

2 *The incompatibility of non-reductive physicalism and interventionism*

Even though non-reductive physicalism is nowadays often seen as a doctrine about the relation and interplay between nonphysical and physical properties in general, it is commonly most vigorously defended for the special case of mental properties and their physical supervenience bases (Eronen 2012, §5). Thus, in order to avoid unnecessary complications, I tailor the subsequent discussion to that special case from the beginning.

The relevant tenets of non-reductive physicalism as regards the relation of the mental and the physical are as follows:

- (NR_1) Every physical state/property that has a cause has a (complete/sufficient) physical cause.
- (NR_2) Mental states/properties supervene on physical states/properties without being reducible to the latter.
- (NR_3) Mental states/properties may cause physical effects of their own supervenience bases.

(NR_1) simply states that the domain of the physical is causally complete, which is the characteristic feature of physicalism. Accordingly, reductive physicalists—the main opponents of non-reductive physicalists—also subscribe to the causal

completeness of the physical. (NR_1) is thus the uncontroversial element of non-reductive physicalism.

(NR_2) characterizes the relationship between mental and physical properties in terms of non-reductive supervenience. The exact meaning and implications of (NR_2), of course, hinge on the presupposed notion of supervenience. As is well known, supervenience has been cashed out in a number of very different ways (cf. e.g. McLaughlin 1995 or Bennett 2004). However, there are at least two features shared by all of these notions: first, supervenience is a non-causal relation and, second, every change in the supervening property is necessarily accompanied by a change in the supervenience base. Thus, (NR_2) entails that mental properties and their physical supervenience bases are non-causally correlated. Moreover, the non-reducibility of the mental implies that mental properties and their supervenience bases are non-identical (notably because of multiple realizability). That is, mental properties differ ontologically from physical properties (Walter and Eronen 2011, 142).

Finally, (NR_3) states that mental properties can have a causal impact on physical properties, in particular, on physical effects of their own supervenience bases.[2] In other words, there possibly exists mental-to-physical downward causation, where causation is understood as a type-level relation that takes properties or event types as relata (which are standardly modeled by means of random variables).[3]

Many non-reductive physicalists additionally assume something along the lines of the *Eleatic Principle* (cf. e.g. Armstrong 1997; Oddie 1982) according to which everything that exists (in space and time) has causal powers. From the fact that mental properties constitute an ontological category that is distinct and non-reducible to the domain of the physical, as induced by (NR_2), they then infer that mental properties have their own non-reducible causal powers. That is, mental properties do not have causal powers merely by virtue of being physically realized but by their own right, i.e. they have genuinely mental causal powers. This is known as the principle of the *Causal Autonomy of the Mental* (CAM) (cf. e.g. Fodor 1989, Lowe 1993, Antony 2007; Menzies and List 2010). Here is the version of the principle advanced in Menzies and List (2010, 111):

(CAM) For some mental property M and physical property P, where an instance of property M is realized by an instance of property P, the causal powers of the M-instance are not a subset of those of the P-instance.

If (NR_3) is combined with (CAM) it follows that mental properties can have a non-reducible, genuinely mental causal impact on physical effects of their super-

[2] Of course, apart from mental-to-physical causation non-reductive physicalists usually also endorse the possibility of mental-to-mental causation. This paper will only be concerned with the former type of mental causation.

[3] Non-reductive physicalists endorse the possibility of downward mental causation not only on the type but also on the token level. Yet, as the main controversies among non-reductive and reductive physicalists concern the type level, the subsequent discussion will be focusing on type causation.

venience bases. While Antony (2007) takes the autonomy of the mental to be a defining feature of non-reductive physicalism, not all authors that identify themselves as non-reductive physicalists endorse (CAM) as explicitly as e.g. Antony (2007), Lowe (1993), or Menzies and List (2010). It seems fair to say that (CAM) is the most controversial element of non-reductive physicalism. On the face of it, it might even be thought that (CAM) conflicts with the causal completeness of the physical (NR_1): if some mental properties have causal powers influencing the realm of the physical that are not shared by their physical supervenience bases, it seems to follow that the domain of the physical is not causally complete. However, (CAM) must not be read to entail that mental properties can cause physical effects that no physical properties have the power to cause. Rather, it entails that some physical effects are due both to a physical and a mental cause, i.e. that even though the physical causal powers are sufficient to account for all physical effects, there also exist mental causal powers. Of course, this then induces an overdetermination of physical effects of mental causes, which non-reductive physicalists take great pains to render plausible (cf. e.g. van Gulick 1993; Marcus 2001; Loewer 2007; Harbecke 2008, ch. 4).

As this paper is not concerned with defending or rejecting non-reductive physicalism but with the relationship between non-reductive physicalism and interventionism, I abstain from further substantiating (CAM) or any other tenet of non-reductive physicalism at this point. Rather, my focus will be on the *interventionist movement* that has recently begun to emerge among non-reductive physicalists. Numerous non-reductive physicalists—e.g. Shapiro and Sober (2007), Menzies (2008), Shapiro (2010; 2012), or Raatikainen (2010) (cf. also Macdonald 2007)—claim that Woodward's (2003) interventionist theory of causation paves the way for a non-reductionist account of downward mental causation. Furthermore, Shapiro and Sober (2007, 247) contend that interventionism "provides a means by which to test which causal powers a macroproperty has". More specifically, Shapiro (2010, 600-601) conceives of the following test for downward mental causation: hold fixed all common causes of a mental variable M_1 and of a physical variable P_2 such that P_2 is an effect of M_1's supervenience base, and intervene on M_1; if such interventions are correlated with changes in P_2, this constitutes empirical evidence that M_1 is a cause of P_2. Thereby, the existence of downward causation is claimed to be establishable on evidence-based grounds, that is, outside of the arena of armchair metaphysics (cf. Raatikainen 2010, 351).[4] The remainder of the paper will thus be concerned with the question to what degree interventionism indeed lends theoretical and/or empirical support to (NR_3) and (CAM).

Woodward's interventionist theory of causation turns on two (frequently cited) definitions:

[4]This thread in the literature must be distinguished from another thread which aims to resolve the problem of causal exclusion by drawing on theories that define causation in terms of contrasts or proportionality (cf. e.g. List and Menzies 2009). From the interventionist perspective, contrasts and proportionality are important for causal explanations, but these notions do not enter the definiens of causation.

(M) A necessary and sufficient condition for X to be a (type-level) direct cause of Y with respect to a variable set $\mathbf{V}$ 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 $\mathbf{V}$. A necessary and sufficient condition for X to be a (type-level) *contributing cause* of Y with respect to variable set $\mathbf{V}$ 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 $\mathbf{V}$ that are not on this path are fixed at some value. (Woodward 2003, 59)

(IV) I is an intervention variable for X with respect to Y iff (i) I causes X; (ii) I acts as a switch for all the other variables that cause X; (iii) any directed path from I to Y goes through X; (iv) 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, 98)

Relative to the notion of an intervention variable, an intervention on X with respect to Y is then straightforwardly spelled out in terms of an intervention variable I for X with respect to Y taking on some value z_i such that $I = z_i$ causes X to take on some determinate value z_j (Woodward 2003, 98).

Before turning to the relationship between interventionism and non-reductive physicalism, two things need to be made explicit about a theory of causation built on (M) and (IV)—things that are too often neglected in the literature. First, the notion of causation provided by (M) is relativized to a set of analyzed variables $\mathbf{V}$, but, as Woodward emphasizes in (2008b, 202), the notion of an intervention variable defined by (IV) is not relativized in that manner.[5] That means a variable X is a direct or contributing cause of a variable Y only relative to some variable set $\mathbf{V}$, while the interventionist nature of a variable I does not depend on which variables are being causally modeled. In that context, it must also be noted that the notion of causation that appears in (IV) is not the relativized notion defined in (M), i.e. not the ternary relation "X causes Y with respect to $\mathbf{V}$". Rather, (IV) draws on de-relativized causation or *causation simpliciter* which Woodward (2008b, 209) defines via existential generalization of (M): a variable X is a cause of Y iff there exists at least one set $\mathbf{V}$ with respect to which X is either a direct or a contributing cause of Y as defined by (M).

Second, (M) and (IV) establish a tight conceptual connection between manipulability (or counterfactual manipulations), difference-making in context, and causality, which Woodward (2003, 61) sums up in the following slogan: No causal difference without a difference in manipulability relations, and no difference in manipulability relations without a causal difference. In particular, the analysans of causation supplied by (M) stipulates that if X is a (type-level) cause of Y, then

[5] If (IV) were relativized like (M), interventionism could not distinguish between difference-making relations that stem from causal dependencies and difference-making relations that are due to common causes (for details cf. section 3 below).

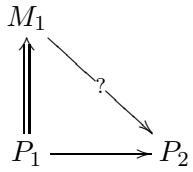

Figure 1: M_1 represents a mental event type, P_1 its physical supervenience base, and P_2 a physical effect of P_1. "$\Longrightarrow$" represents supervenience, "$\longrightarrow$" stands for causation.

there exists a possible intervention $I = z_i$ on X with respect to Y. That is, (M) determines that the manipulability of X is necessary for X to cause Y. Moreover, the conjunction of (IV.i) and (IV.iv) entails that an intervention variable for X with respect to Y causes changes in the values (or the probability distribution) of X and is (statistically) independent of all causes of Y that are not located on a path that goes through X. If there does not possibly exist a variable I that meets this *independence requirement*, there does not exist a possible intervention on X with respect to Y which implies that the manipulability of X is violated which, again, implies that X does not cause Y.

We have now assembled all elements and consequences of non-reductive physicalism and interventionism that are needed to investigate their relationship. As anticipated in the previous section, I have shown in Baumgartner (2009; 2010) that the conjunction (NR_1) $\wedge$ (NR_2) $\wedge$ (M) $\wedge$ (IV) entails the negation of (NR_3). To see this, consider the diagram in figure 1 which is standardly discussed in the literature on Kimian exclusion arguments. The variable P_2 represents some physical event type, say, some type of action. (NR_1) entails that there exists a physical cause of P_2, for instance, some brain process, which in figure 1 is represented by P_1. Moreover, there shall be a mental event type represented by M_1, which, according to (NR_2), non-reductively supervenes on P_1. For example, M_1 can be taken to stand for a decision or choice to perform the P_2-action. One crucial question now is whether it is possible for M_1 to be a cause of P_2 as stipulated by (NR_3).

That (NR_1)$\wedge$(NR_2)$\wedge$(M)$\wedge$(IV) entails a negative answer to that question can be shown as follows. As we have seen above, (NR_2) implies, first, that M_1 and P_1 are not causally related, second, that they represent different event types or properties, and third, that all changes in M_1 are necessarily accompanied by changes in P_1. From (NR_1) it follows that P_1 is on a causal path to P_2, which, in light of what we have just said about the consequences of (NR_2), must be a causal path that does not include M_1. This, in combination with (IV), implies that intervention variables for M_1 with respect to P_2 have to be independent of changes in P_1. However, subject to (NR_2), it is impossible to induce changes in M_1 independently of P_1. Therefore, there cannot possibly exist an (IV)-defined intervention variable for M_1 with respect to P_2. If we now apply (M) to the variable set $\{M_1, P_1, P_2\}$, we can infer that M_1 does not cause P_2 relative to $\{M_1, P_1, P_2\}$. Finally, since the notion of an intervention variable provided by (IV) is not relativized to variable

sets, every possible intervention on M_1 must be independent of P_1—irrespective of whether P_1 is contained in an analyzed set or not. It follows that there does not exist a variable set with respect to which there possibly exists an (IV)-defined intervention variable for M_1 with respect to P_2, which entails that M_1 does not cause P_2 *simpliciter*.

As the validity of the above argument in no way hinges on which concrete properties or event types the variables in figure 1 stand for, the conclusion as to the causal irrelevance of M_1 to P_2 can be generalized for all mental properties. The argument reveals that the conjunction of (NR_1), (NR_2), (M), and (IV) rules out the possibility of mental properties causing physical effects of their own supervenience bases. That is, downward mental causation which is claimed to be possible by (NR_3) is discarded as impossible by the variant of interventionism presented in Woodward (2003), i.e. (NR_1)$\wedge$(NR_2)$\wedge$(M)$\wedge$(IV) entails $\neg$(NR_3). Interventionism and non-reductive physicalism are hence *incompatible*.

3 *Adaptations of interventionism*

Independently of whether one has sympathies for non-reductive physicalism or not, the question as to the possible existence of cases of non-reductive downward causation does not appear to be of purely conceptual nature. Thus, there seems to be something wrong with a theory of causation that excludes that possibility on *a priori* conceptual grounds. Accordingly, Woodward (2011), Eronen (2012), or Weslake (unpublished) suggest weakened variants of interventionism that no longer conflict with non-reductive physicalism.[6]

While these weakened variants of interventionism differ in details, they agree on the basic strategy to render interventionism compatible with non-reductive physicalism: Woodward (2011), Eronen (2012), and Weslake (unpublished) contend that interventionism is embedded in a tradition of causal modeling (e.g. Spirtes et al. 2000; Pearl 2000; Woodward and Hitchcock 2003; Halpern and Pearl 2005) where it is generally presupposed that analyzed sets of variables meet specific suitability conditions; and these conditions are violated by sets featuring supervenience relations (or relations of definitional dependence).[7] It is thus not adequate to apply interventionist definitions to such sets, as I do in my incompatibility argument. For instance, Eronen (2012, 228) points out that the causal modeling tradition axiomat-

[6] Woodward (2011) moreover insists that he intended the weak reading of the interventionist definitions already in Woodward (2003). I sidestep this exegetic issue in this paper. Note also that Woodward (2011) does not weaken interventionism in order to argue in favor of non-reductive physicalism on evidence-based grounds in the vein of Sober, Shapiro et al. Rather, he primarily intends to render the two theories compatible.

[7] In fact, there is a whole array of metaphysical and conceptual relations that give rise to violations of these suitability conditions imposed on modeled variable sets, e.g. supervenience, emergence, constitution, logical dependence, mereological dependence, definitional dependence etc. In this paper, I simply use supervenience and definitional dependence as proxies for this whole array.

ically stipulates that every modeled variable set $\mathbf{V}$ satisfies the *Causal Markov Condition* (e.g. Spirtes et al. 2000, 29):

(CMC) If $\mathbf{V}$ is causally sufficient, then every variable in $\mathbf{V}$ is (probabilistically) independent of all its non-effects in $\mathbf{V}$ conditional on its direct causes in $\mathbf{V}$.

A set $\mathbf{V}$ is said to be *causally sufficient* iff any common cause C of two variables X and Y in $\mathbf{V}$ either belongs to $\mathbf{V}$ or has a cause that belongs to $\mathbf{V}$ or an effect that is located on all directed paths from C to X and from C to Y and that belongs to $\mathbf{V}$. If we assume that the set $\{M_1, P_1, P_2\}$ of figure 1 is causally sufficient, it follows from (CMC)—in combination with the fact that P_1 and M_1 have no direct causes in figure 1—that P_1 and M_1 are unconditionally independent, which however, due to supervenience, they are not. Hence, $\{M_1, P_1, P_2\}$ violates (CMC). Accordingly, the set $\{M_1, P_1, P_2\}$ is unsuitable for causal modeling, in general, and for interventionist causal modeling, in particular.

Comparably, Woodward (2011) and Weslake (unpublished) maintain that variable sets that are suitable for causal modeling must satisfy what Woodward (2011, 12) calls *Independent Fixability*:[8]

(IF) A set of variables $\mathbf{V}$ satisfies independent fixability of values iff for each value it is possible for a variable to take individually, it is possible (that is, "possible" in terms of their assumed definitional, logical, mathematical, or mereological relations or "metaphysically possible") to set the variable to that value via an intervention, concurrently with each of the other variables in $\mathbf{V}$ also being set to any of its individually possible values by independent interventions.

While in the causal modeling tradition (CMC) is standardly and very explicitly assumed, (IF)—or something of that kind—might indeed be in the back of many authors' minds, but it is rarely imposed explicitly.[9] Clearly though, the set $\{M_1, P_1, P_2\}$ of figure 1 violates (IF), for, in virtue of M_1 supervening on P_1, it is (metaphysically) impossible to independently set M_1 and P_1 to any of their values. It follows, as in case of a violation of (CMC), that the set $\{M_1, P_1, P_2\}$ cannot be causally modeled by interventionist means.

(CMC) and (IF) are not logically independent, but their exact logical relation is intricate, and I shall not attempt to clarify it here. It is clear, however, that while (IF) imposes constraints on all variable sets, (CMC) only has ramifications for causally sufficient sets. $\{M_1, P_1, P_2\}$ can be said to unconditionally violate (IF), whereas that set violates (CMC) only if it is causally sufficient; and whether that is indeed the case is far from obvious. In light of the unconditional nature of (IF),

[8] Weslake (unpublished) introduces an analogous condition that he calls *Independent Manipulability*.

[9] An exception is Halpern and Hitchcock (2010), who explicitly insist on independent manipulability.

imposing (IF) is the more straightforward way of rendering variable sets involving supervenience relations unsuitable for causal modeling. I shall hence not pursue the proposal to attain the same goal by imposing (CMC) here.

Instead, let us now turn to what follows for my incompatibility argument from stipulating that variable sets that are suitable for causal modeling must satisfy (IF). The fact that $\{M_1, P_1, P_2\}$ is unsuitable because it violates (IF) entails that interventionist definitions do not apply. Accordingly, interventionism cannot be used to infer that M_1 does not cause P_2 in the structure of figure 1, i.e. with respect to the set $\{M_1, P_1, P_2\}$. Thereby my incompatibility argument is blocked. This finding, however, raises the follow-up question what the inapplicability of interventionist definitions to the structure in figure 1 means for the possibility of downward mental causation. From the mere fact that interventionism can no longer be said to entail that M_1 does not cause P_2 with respect to $\{M_1, P_1, P_2\}$ it obviously cannot be inferred that interventionism entails that it is possible for M_1 to cause P_2 (simpliciter); yet, the latter is what non-reductive physicalists need in order to back up their claim (NR_3).

To determine what restricting the modeling suitability of variable sets to (IF)-compatible sets entails for interventionist downward causation, it must be clarified where and how exactly (IF) enters the interventionist framework. If (IF) is introduced as a precondition for the applicability of interventionist definitions, interventionism simply remains silent about causal dependencies among variables in (IF)-violating sets. More concretely, if (IF) is a precondition for applying (M) and (IV), the latter cannot be applied to $\{M_1, P_1, P_2\}$. Accordingly, these definitions neither imply that M_1 causes P_2 nor that M_1 does not cause P_2 with respect to $\{M_1, P_1, P_2\}$. It could then be said that the downward causal claim "M_1 causes P_2 with respect to $\{M_1, P_1, P_2\}$" is *ill-defined* (from the perspective of interventionism).[10] However, non-reductive physicalists, of course, do not want to maintain that downward causal claims are ill-defined. Rather, they insist that such claims are well-defined and, moreover, often true.

Yet, notwithstanding the fact that M_1 cannot be said to cause P_2 with respect to $\{M_1, P_1, P_2\}$, it might still be possible to establish that M_1 causes P_2 simpliciter on interventionist grounds. There might exist *another variable set* that complies with (IF) and relative to which (M) and (IV) yield that M_1 is a cause of P_2. That is, (IF) could be used as a sort of maxim for selecting variable sets that are causally analyzable by interventionist means: if a variable set $\mathbf{V}$ violates (IF) due to supervenience relationships among some of its elements, construct a set $\mathbf{V}'$ from $\mathbf{V}$ by eliminating—depending on given modeling purposes—the variables representing supervening properties or the ones representing corresponding supervenience bases. (M) and (IV) should then be applicable to $\mathbf{V}'$. This is essentially how Ero-

[10] This is the picture Woodward has in mind in Woodward (2008a) where he assumes "that if a candidate causal claim is associated with interventions that are impossible for (or lack any clear sense because of) logical, conceptual or perhaps metaphysical reasons, then that causal claim is itself illegitimate or ill-defined" (Woodward 2008a, 224).

nen (2012) suggests to put (CMC) to use in the course of selecting a variable set that is tractable by interventionist techniques.

Let us apply that idea to the set $\{M_1, P_1, P_2\}$ of figure 1. We want to determine what interventionism implies for the downward causal claim "M_1 is a cause of P_2". To answer that question we need to construct a set from $\{M_1, P_1, P_2\}$ that complies with (IF) and contains both M_1 and P_2. Hence, we eliminate P_1 and are left with $\{M_1, P_2\}$. This set no longer features any supervenience relations and satisfies (IF). Eronen (2012) argues that, based on (M) and (IV), M_1 can now be shown to cause P_2 with respect to $\{M_1, P_2\}$. That, however, is not the case without further adaptations of interventionist definitions.[11] To see this, note again that (IV) is *not relativized* to variable sets. (IV) requires intervention variables on M_1 with respect to P_2 to be independent of all other causes of P_2 that are located on a path to P_2 that does not include M_1. Accordingly, if there does not possibly exist an intervention variable for M_1 with respect to P_2 relative to $\{M_1, P_1, P_2\}$, there does not possibly exist such an intervention variable relative to $\{M_1, P_2\}$ either. Hence, as manipulations of M_1 will necessarily be correlated with changes in P_1, which according to non-reductive physicalism is a cause of P_2 not located on an M_1-path, there does not possibly exist an intervention variable for M_1 with respect to P_2 simpliciter. (M) hence rules that M_1 does not cause P_2 with respect to $\{M_1, P_2\}$ either.

That is, if the non-reductive physicalist wants an interventionist account of causation that renders the claim "M_1 is a cause of P_2 with respect to $\{M_1, P_2\}$" true, she must weaken (IV). One such weakening comes easily to mind: (IV) might simply be relativized to a variable set $\mathbf{V}$ in close analogy to (M). More specifically, (IV) might be replaced by (IV$_{rel}$):

(IV$_{rel}$) I is an intervention variable for X with respect to Y relative to $\mathbf{V}$ iff (i) I causes X with respect to $\mathbf{V}$; (ii) I acts as a switch for all the other variables that cause X in $\mathbf{V}$; (iii) any directed path from I to Y in $\mathbf{V}$ goes through X; (iv) I is statistically independent of any variable Z in $\mathbf{V}$ that causes Y and that is on a directed path that does not go through X.

Now, eliminating P_1 from $\{M_1, P_1, P_2\}$ makes a difference to whether M_1 is (IV$_{rel}$)-manipulable with respect to P_2. While there does not possibly exist an (IV$_{rel}$)-defined intervention variable for M_1 with respect to P_2 relative to $\{M_1, P_1, P_2\}$, there possibly exists such a variable relative to $\{M_1, P_2\}$. (IV$_{rel}$.iv) does not require intervention variables for M_1 to be independent of all other causes of P_2 not located on a path from M_1 to P_2, but only of those other causes contained in $\{M_1, P_2\}$; and $\{M_1, P_2\}$ does not contain any other causes of P_2 at all. Hence, (IV$_{rel}$) does not require intervention variables for M_1 with respect to P_2 to be independent of P_1. It is thus possible to (IV$_{rel}$)-manipulate M_1. Moreover, if (IV$_{rel}$)-wiggling M_1 is accompanied by changes in P_2, (M) entails what the non-reductive

[11] Similarly, Yang (forthcoming) does not see that simply rendering analyzed variable sets (IF)-compatible does not itself guarantee that (M) and, in particular, (IV) are applicable.

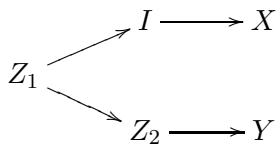

Figure 2: An ordinary common cause structure over an (IF)-compatible variable set.

physicalist wants, *viz.* that M_1 causes P_2 with respect to $\{M_1, P_2\}$, which, in turn, yields that M_1 causes P_2 simpliciter.

Nonetheless, non-reductive physicalists would be ill-advised to settle for a variant of interventionism that turns on (M) and (IV$_{rel}$), for the latter do not provide an adequate analysis of causation. To see this, consider the causal structure depicted in figure 2, which is an ordinary common cause structure without supervenience relationships. Hence, any set of variables involved in that structure satisfies (IF). Suppose, we select the set $\{I, X, Y\}$. Relative to $\{I, X, Y\}$, I passes as an (IV$_{rel}$)-intervention variable for X with respect to Y: I is a cause of X that is independent of all other causes of Y in $\{I, X, Y\}$ and I is not connected to Y in $\{I, X, Y\}$ along a path that does not go through X. Furthermore, when all other variables in $\{I, X, Y\}$ are fixed, changes induced on X via I are correlated with changes in Y. Accordingly, (M) rules that X is a direct cause of Y with respect to $\{I, X, Y\}$. Obviously, this is false. In the structure of figure 2, X does not cause Y simpliciter, i.e. with respect to any set. That means a theory of causation that is built on (M) and (IV$_{rel}$) cannot distinguish between difference-making relations that stem from causal dependencies and difference-making relations that stem from common causes.

Indeed, neither Woodward (2011) nor Weslake (unpublished) advocate a weakening of (IV) in the vein of (IV$_{rel}$). Rather, Woodward (2011, 27) weakens (IV) by introducing exemption clauses for supervenience relationships (and definitional dependencies) into (IV.iii) and (IV.iv). An intervention on a supervening variable X necessarily also changes X's supervenience base $SB(X)$. That means interventions on X are tantamount to interventions on $SB(X)$. Accordingly, (IV.iii) and (IV.iv) must be weakened in such a way that they neither prohibit that an intervention variable I for X with respect to Y is connected to Y via causal paths through $SB(X)$ nor that I is correlated with variables on such paths. In that light, Woodward proposes to replace (IV) by (IV*):[12]

(IV*) I is an intervention variable for X with respect to Y iff I satisfies (IV.i), (IV.ii), (IV.iii*), and (IV.iv*):

[12]Weslake's (unpublished) weakening of (IV) is somewhat different. His discussion is, from the outset, centered on a version of interventionism that is defined on the basis of the auxiliary notion of a *causal model*, which Woodward's original version of interventionism is not. To stick to a more standard variant of interventionism, I focus on Woodward's (2011) discussion here.

(IV.iii*) any directed path from I to Y goes through X *or* through a variable Z which is related to X in terms of supervenience (or definition);[13]

(IV.iv*) I is (statistically) independent of every cause of Y which is neither located on a path through X *nor* on a path through a variable Z which is related to X in terms of supervenience (or definition).

(IV*) permits intervention variables for X with respect to Y to be causes of, and correlated with, other causes of Y that are not located on an X-path, provided that those other causes are on a path through $SB(X)$ (or through a variable that is definitionally related to X). Against that background, non-reductively supervening macro variables become manipulable. In the case of figure 1, (IV*)-intervention variables for M_1 with respect to P_2 are allowed to induce changes on P_1, even though P_1 is a cause of P_2 that is not located on an M_1-path, because P_1 represents the supervenience base of M_1. M_1 is thus (IV*)-manipulable both relative to $\{M_1, P_1, P_2\}$ and relative to $\{M_1, P_2\}$.

Furthermore, it turns out that a variant of interventionism that is based on (M) and (IV*) implies that M_1 is a cause of P_2 with respect to $\{M_1, P_2\}$, provided that (IV*)-wiggling M_1 is accompanied by changes in P_2. As such a correlation of M_1 and P_2 under (IV*)-manipulations of M_1 is certainly possible, (M) and (IV*) yield the possibility of mental downward causation—just as advocated by non-reductive physicalists. Hence, implementing (IF) as a maxim for selecting suitable variable sets and weakening (IV) in terms of (IV*) results in a variant of interventionism that not only fails to entail $\neg$(NR$_3$), but that additionally entails the possibility of downward causation, i.e. that entails (NR$_3$).

Nonetheless, Woodward (2011) does not opt for a weakening of his original theory that is built on (M) and (IV*) and that uses (IF) as a maxim for selecting suitable variable sets. The reason is that he takes (M) to be overly restrictive. (M) not only requires X to be manipulable with respect to Y in order for X to cause Y, but also that all variables not located on a path from X to Y, i.e. all *off-path variables*, are fixed while X is manipulated. That is, (M) determines that, in addition to the manipulability of X, also the *fixability of all off-path variables* in $\mathbf{V}$ is a necessary condition for X to cause Y. According to Woodward (2011, 27), however, this is asking too much:

> In assessing whether X is a direct cause of Y, the "other variables" in $\mathbf{V}$ that we should hold fixed independently of the intervention on X (i) should *not* include the supervenience base for X *and* (ii) should *not* include the supervenience base for Y. (...) In order to assess whether X is a direct cause of Y, we *should* hold fixed via an IV*-intervention any other variables V_i in $\mathbf{V}$ that stand in ordinary causal or correlational relations to X and Y.

[13] This is a deliberately abbreviated way of expression. It is, of course, not variables that are related in terms of supervenience, but the properties represented by those variables.

As a consequence, Woodward not only builds exemption clauses for supervenience relations into (IV) but also into (M). More specifically, he suggests to replace (M) by something along the lines of (M*):[14]

(M*) X is a (type-level) *direct* cause of Y with respect to the variable set $\mathbf{V}$ iff there possibly exists an (IV*)-defined intervention on X with respect to Y such that all other variables in $\mathbf{V}$ that are not related in terms of supervenience (or definition) to X or Y are held fixed, and the value or the probability distribution of Y changes.

X is a (type-level) *contributing* cause of Y with respect to the variable set $\mathbf{V}$ iff (i) there is a directed path from X to Y such that each link on this path is a direct causal relationship and (ii) there possibly exists an (IV*)-defined intervention on X with respect to Y such that all other variables in $\mathbf{V}$ that are not located on a causal path from X to Y or on a path from a variable Z to Y, such that Z is related in terms of supervenience (or definition) to X or Y, are held fixed and the value or the probability distribution of Y changes.

By appending exemption clauses to (M) and (IV) for (IF)-violating sets, (IF) can be dispensed with, for, contrary to (M)-(IV*)-(IF)-interventionism, (M*)-(IV*)-interventionism is applicable to all variable sets, even to (IF)-violating sets. Moreover, (M*)-(IV*)-interventionism also entails that downward causation is possible. To see this, take once more the set $\{M_1, P_1, P_2\}$ of figure 1. (M*) does not require P_1—the supervenience base of M_1—to be fixed while M_1 is (IV*)-manipulated. Thus, (M*) determines that M_1 is a cause of P_2 with respect to $\{M_1, P_1, P_2\}$ if, and only if, (IV*)-manipulations of M_1 are accompanied by changes in P_2. Plainly, that (IV*)-wiggling M_1 makes a difference to P_2 is possible, whereby downward mental causation is rendered possible.

The reason why Woodward prefers (M*) over (M) is that he takes it to be a methodological mistake to control for supervenience bases when testing the causal efficacy of variables representing mental properties (cf. Woodward 2011, 29). Fixing off-path variables when testing the causal relation between M_1 and P_2 serves the purpose of controlling for variables that could possibly confound corresponding empirical data. According to Woodward, however, P_1 could not possibly confound data on the causal interplay of M_1 and P_2. He offers two rationales to back up that assessment: one based on an analogy with definitionally related variables and one based on scientific practice. Let me take these rationales in turn.

The analogy with definitionally related variables turns on a well-known example from Spirtes and Scheines (2004). There are two sorts of cholesterol: high-density cholesterol (HC) and low-density cholesterol (LC). LC promotes heart disease (D), whereas HC prevents heart disease. Suppose, we now introduce a further (coarse-grained) variable representing total cholesterol (TC) which is simply defined as the arithmetic sum of HC and LC, i.e. $TC = HC + LC$. It is

[14] Woodward (2011) does not state (M*) explicitly, but merely indicates its relevant features. Thus, (M*), as given here, is my reconstruction of Woodward's suggestion.

uncontroversial that, in order to determine what the causal impact of TC is on D, it would be a mistake to attempt to intervene on TC while holding HC and LC fixed, even though the latter are causally relevant to D. Weslake (unpublished, 24) uses Yablo's (1992, 257) prominent example of a pigeon's pecking habits to the same effect. Suppose, a pigeon pecks (P) if, any only if, presented with red things (R). It follows that presenting it with scarlet things (S) is also sufficient for the pigeon to peck. Against that background, it would be a mistake to argue that R does not cause P, because it is impossible to manipulate the value of R while holding S fixed. The suggestion then is that what holds for these variables also holds for M_1 and P_1 in figure 1. In general, just as variables that are dependent due to the definiendum-to-definiens relation or the determinable-to-determinate relation must not be mutually controlled for in contexts of causal discovery, so the downward causal relevance of variables representing supervening mental properties must be tested without controlling for corresponding physical supervenience bases.

Correspondingly, controlling for supervenience bases is at variance with the usual scientific practice (Woodward 2011, 31). Suppose a researcher wants to investigate whether administering a drug Φ causes recovery from a disease (C). Administering Φ can be modeled by a coarse-grained variable that represents the macroscopic properties of Φ, call that variable A, but it can likewise be modeled by a fine-grained variable that represents the chemical supervenience base of Φ's macroscopic properties, call it B. Now, it is obviously impossible to manipulate the value of A while holding the value of B fixed. Nonetheless, no sane researcher would conclude from this that A is causally irrelevant to C. On the contrary, scientific practice allows for changes in the values of B while testing the causal relevance of A, even if B is causally relevant to C.

In light of observations of this sort, Woodward (2011) thus advances a version of interventionism whose definitional core is constituted by (M*) and (IV*). This section has shown that (M*)-(IV*)-interventionism is not only compatible with non-reductive physicalism, but moreover entails the possibility of downward mental causation as stated in (NR_3). Yet, does (M*)-(IV*)-interventionism also theoretically support (NR_3)? If, say, (M*)-(IV*)-interventionism should turn out to be an inadequate theory of causation, its entailing (NR_3) would be of no avail to the non-reductive physicalist. And what does (M*)-(IV*)-interventionism entail for (CAM)? Moreover, does (M*)-(IV*)-interventionism allow for empirically validating the existence of downward causation? These questions are going to be addressed in the next section.

4 Assessing (M)-(IV*)-interventionism*

Let us begin by considering what follows from the fact that definitionally related variables and variables that stand in the determinable-to-determinate relation do not need to be mutually fixed when their causal effects are tested by interventionist means. In order for this to have ramifications for the case of mental-to-

physical causation as conceived of be the non-reductive physicalist, the relations of definiendum-to-definiens and determinable-to-determinate would have to be shown to be relevantly similar to the non-reductive supervenience of the mental on the physical. However, it seems that the former two relations differ from the latter in respects that are not negligible for the possibility of data-confounding.

To bring out these differences let us first consider the definiendum-to-definiens relation. Definition comes with reducibility, that is, a definiendum is reducible to its definiens. Total cholesterol (TC) is nothing but the arithmetic sum of low-density (LC) and high-density cholesterol (HC). Any of these properties is expressible as a function of the other two. TC and $LC + HC$ do not represent different properties. Accordingly, TC cannot possibly have causal powers that are non-reducible to the causal powers of $LC + HC$, and vice versa. If TC causes D, that holds in virtue of $LC + HC$ causing D. The claims "TC is a cause of D" and "LC and HC are causes of D" describe the same physiological mechanism on different levels of specification. Therefore, LC and HC cannot confound intervention tests conducted on TC, and thus, fixing LC or HC when manipulating TC, or vice versa, is not called for.

The case of the determinable-to-determinate relation is a bit more intricate, but in the end something analogous holds here as well. Many authors have pointed out that mental properties cannot be seen as determinables and their supervenience bases as corresponding determinates (cf. Ehring 1996; Funkhouser 2006; Walter 2007; Menzies 2008). According to Funkhousers' (2006) analysis of the determinable-to-determinate relation—which I take to be the most elaborate analysis currently available—, determinables span a property space with a limited number of so-called *determination dimensions*, and determinates simply come out as points (or sets of points) in that space. For example, the determinable *Red* spans a 3-dimensional space with hue, brightness, and saturation as determination dimensions and the (super-)determinate *Coca-Cola-Red* is one point in that space. The essential difference between the determinable-to-determinate relation and the mental-to-brain relation, according to Funkhouser (2006, 563), is that determinables and determinates necessarily have exactly the same determination dimensions, whereas mental properties and their physical supervenience bases may have different determination dimensions. Against the background of this account, it can easily be explained why determinates do not have to be held fixed when testing for causal effects of corresponding determinables. Determinables correspond to the n-dimensional property spaces they span, and the points in these spaces correspond to their determinates. Hence, determinables are nothing over and above the sets of their determinates. Determinables are thus reducible to their determinates, which yields that determinables have causal effects *in virtue of* their determinates.[15] Therefore, determinates cannot possibly confound intervention tests conducted on their determinables, and vice versa.

[15] Gillett and Rives (2005) even advance the eliminability of determinables.

Yet, something analogous does not hold for the relation in terms of which the non-reductive physicalist conceives of mental properties and their physical supervenience bases. The non-reductive physicalist endorses a property dualism: mental properties are *not reducible* to their supervenience bases and, subject to (CAM), some of them have causal powers their supervenience bases lack. She denies that mental properties cause effects of their own supervenience bases in virtue of being physically realized. Hence, while variables that are definitionally related or related in terms of the determinable-to-determinate relation are innocuous as regards mutual data-confounding, it is far from clear that the same holds for variables representing mental properties and variables representing corresponding supervenience bases. If two causally independent variables X and Y represent mutually non-reducible properties that are *not* related in terms of supervenience, we would decidedly require that the effects of X be tested while Y is held fixed (and vice versa), because if Y is allowed to change its values when X is manipulated, observed differences in investigated effects would not be guaranteed to be due to X. On the face of it, thus, it is not at all clear why the variable P_1 in figure 1 could not equally confound intervention tests conducted on M_1.

By contrast, non-reductive physicalists as Shapiro and Sober (2007), Shapiro (2010; 2012), Raatikainen (2010), or Menzies (2008) who advocate an evidence-based resolution of the problem of causal exclusion based on interventionist techniques seem to hold that the fact that M_1 non-reductively supervenes on P_1 averts the danger of such data-confounding, which is why e.g. Shapiro and Sober (2007) and Shapiro (2010) decidedly insist that fixing P_1 while manipulating M_1 is the wrong test to perform. They implicitly subscribe to a methodological principle along the following lines:

(†) If two causally unrelated variables X and Y cannot be independently manipulated because X *non-reductively* supervenes on Y (or vice versa), then the correlation of X and Y cannot give rise to data-confounding.

While correlations of variables that are related in terms of reductive variants of supervenience are indeed likely not to give rise to data-confounding, it is far from clear whether the same holds for variables related in terms of non-reductive supervenience. In any case, examples involving definitionally related variables or variables representing determinables and determinates provide no rationale for the validity of (†) because such variables do not represent mutually non-reducible properties.

Similarly, the fact that actual scientific practice does not require that a fine-grained variable B, which represents the chemical features of a drug Φ, has to be fixed when investigating the causal effects of a coarse-grained variable A, which represents the macroscopic features of Φ, has no immediate bearing on variables that are related in terms of non-reductive supervenience. A researcher who models the causal effects of Φ on a coarse-grained level using the variable A does not work under the assumption that A could be causally autonomous of the fine-grained variable B. Rather, the biomedical researcher takes it for granted that Φ's macro

properties are reducible to its underlying micro properties. If Φ actually cures the disease (C), it does so *in virtue of* its microscopic features. Hence, the reason why that researcher does not attempt to hold B fixed when manipulating A is that she takes it to be excluded that there could be *two* non-reducible causal routes that could confound her data: one from Φ's macro properties to C and another one from Φ's micro properties to C. Rather, she presumes that there is at most one causal process that leads from drug intake to recovery—a process, though, that can be modeled on multiple levels of specification (cf. Wimsatt 2007, ch. 11; Walter and Eronen 2011).[16] That actual scientific practice does not call for fixing supervenience bases when manipulating macro variables simply shows that the notion of supervenience (implicitly) adopted in these contexts is one that allows for reduction, and hence provides no rationale for the validity of principle (†) either.

At the same time, though, the above considerations do not demonstrate the invalidity of (†) either. All we have seen so far is that non-reductive physicalists who want to put (M*)-(IV*)-interventionism to use when it comes to backing up the possibility of non-reductive mental downward causation on evidence-based grounds must endorse (†), that is, a methodological principle whose validity is unclear and would need argumentative backing. I shall not, however, pursue the question of (†)'s validity here. Rather, let us now turn to the question whether (M*)-(IV*)-interventionism, apart from conceptually entailing the truth of (NR$_3$), also suits the other purposes of non-reductive physicalists. To this end, it will be instructive to lay open some of the relevant consequences and presuppositions of (M*)-(IV*)-interventionism.

First, it follows from (M*) and (IV*) that every cause of a mental variable M_i is a common cause of M_i and its supervenience base $SB(M_i)$. To see this, suppose that Z is a cause of M_i. Subject to (M*), this entails that it is possible to (IV*)-wiggle Z such that M_i changes. As every change in M_i is necessarily accompanied by a change in $SB(M_i)$, it follows that it is possible to (IV*)-wiggle Z such that $SB(M_i)$ changes, in light of which (M*), in turn, rules that Z is a cause of $SB(M_i)$. Now, the relationship between M_i and $SB(M_i)$ is non-causal, hence, Z cannot be connected to M_i and $SB(M_i)$ via one single causal path. Therefore, there must be two causal paths: one from Z to M_i and another one from Z to $SB(M_i)$. In particular, this yields that every (IV*)-intervention on M_i—which subject to (IV.i) is a cause of M_i—is a common cause of M_i and $SB(M_i)$. Or differently, M_i can only be (IV*)-intervened upon via a common cause of M_i and $SB(M_i)$.

Next, consider the structure in figure 3, which is an extension of figure 1. In addition to M_1 which represents the decision to perform an action P_2 and M_1's physical supervenience base P_1, figure 3 also features a causal intermediary P' between P_1 and P_2, which e.g. represents the neural activity in the motor cortex

[16] This reductionist paradigm is not only adopted in biomedical disciplines but also in neuroscience and many other disciplines: "Reductionism is the dominant methodology of 'big science' today, percolating widely through the sciences" (Wimsatt 2007, 4).

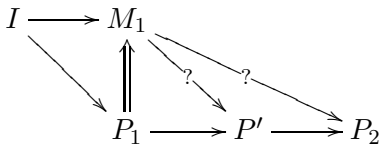

Figure 3: An extension of the structure in figure 1 over the variable set $\mathbf{V_1}$.

that is necessary for the human body to perform P_2. Furthermore, we introduce an (IV*)-intervention variable I for M_1 with respect to P' and P_2, which *eo ipso* is a common cause of M_1 and P_1. It remains to be clarified what (M*)-(IV*)-interventionism entails for the downward paths from M_1 to P' and from M_1 to P_2. We have already seen in the previous section that, in order to determine whether M_1 is an (M*)-defined cause of P_2, it is not necessary to hold P_1 fixed, because P_1 corresponds to the supervenience base of M_1. Likewise, it is not necessary to hold P' fixed, because P' is on a path from $SB(M_1)$ to P_2. Hence, M_1 is an (M*)-cause of P_2 with respect to the set $\mathbf{V_1} = \{I, M_1, P_1, P', P_2\}$ if, and only if, manipulations of M_1 are associated with changes in P_2. If M_1 and P_2 are indeed correlated under changes induced on M_1 via I, wiggling M_1 by means of I is also accompanied by changes in P'. Hence, if M_1 is an (M*)-cause of P_2, M_1 is also an (M*)-cause of P'. Now, suppose that M_1, P', and P_2 are indeed correlated under changes induced on M_1 via I. That means M_1 causes both P' and P_2 with respect to $\mathbf{V_1}$. Moreover, M_1 is a direct cause of P', because M_1 and P' are correlated when all other variables in $\mathbf{V_1}$ except $SB(M_1)$ are held fixed. Yet, M_1 is *not* a direct cause of P_2 in $\mathbf{V_1}$, because when P' is held fixed wiggling M_1 (and P_1) is no longer associated with changes in P_2.[17] Overall, (M*) and (IV*) issue the causal dependencies depicted in figure 4(a). Thus, the decision M_1 causes the action P_2 by virtue of directly causing the neural activity P' in the motor cortex. By contrast, the structure depicted in figure 4(b) is incompatible with

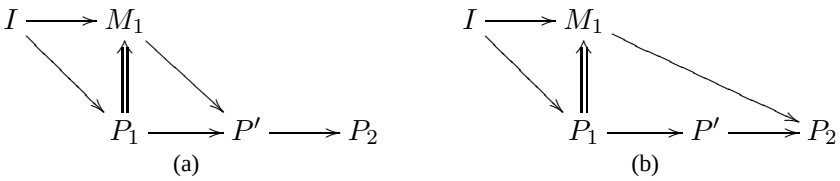

Figure 4: (a) represents a case of downward mental causation that is compatible with (M*)-(IV*)-interventionism. By contrast, (b) is incompatible with (M*)-(IV*)-interventionism.

[17] It might be objected that if P' were not a necessary factor for P_2, M_1 might still make a difference to P_2 even if P' is held fixed. In that case, however, P_1 would also make a difference to P_2 when P' is held fixed. This, in turn, means that there would have to be a directed edge from P_1 to P_2 in figure 3. By contrast, given that the causal relations among the physical variables are as depicted in 3 or 4(a) or 4(b), (M*) rules that M_1 is not a direct cause of P_2.

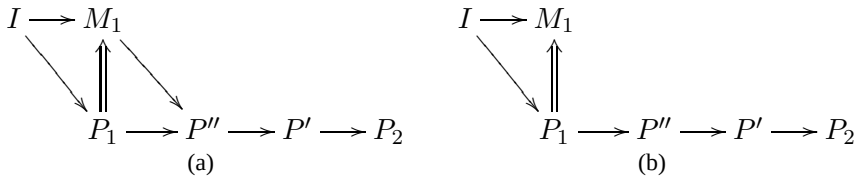

Figure 5: (a) represents a case of downward mental causation that is compatible with (M*)-(IV*)-interventionism. (b) is an alternative epiphenomenalist structure.

(M*)-(IV*)-interventionism. If the causal dependencies among P_1, P', and P_2 are as depicted in figure 3, (M*)-(IV*)-interventionism excludes that M_1 could have an 'autonomous' causal impact on P_2 which is not mediated via P'.

The genuinely mental causal impact of M_1 on P_2 can be even further removed from P_2 by introducing a causal intermediary between P_1 and P' as in figure 5(a). Say, P'' represents the signal transmission from the frontal lobe, where the decision M_1 is realized, to the motor cortex, where neural activity P' coordinates action P_2. If (IV*)-intervening on M_1 is correlated with changes in P_2, it is also correlated with changes in P''. Hence, subject to (M*), M_1 is a direct cause of P'' and a contributing (indirect) cause of P' and P_2 in $\mathbf{V_2} = \{I, M_1, P_1, P', P'', P_2\}$. That is, if a mental event type causes physical effects of its own supervenience base, it directly causes the first physical event type outside of its own supervenience base. From there on, the causal impact of the mental properties *coincides* with the causal impact of its supervenience bases. Or differently, the causal impact of a mental property collapses onto the causal impact of its supervenience base after the first link on a corresponding causal chain. Thus, the 'causal autonomy' of a mental property that (M*)-(IV*)-interventionism allows for is restricted to the first physical link on a causal chain out of that mental property's supervenience base. Undoubtedly, that is a consequence of (M*)-(IV*)-interventionism that does not square nicely with the requirements of non-reductive physicalists who subscribe to (CAM). (M*)-(IV*)-interventionism only leaves room for a very limited sort of causal autonomy of the mental.

Finally, consider structure 5(b), which represents an epiphenomenalist alternative to 5(a), i.e. a structure in which the mental property has no causal relevance to effects of its supervenience base. Just as structure 4(b), 5(b) is incompatible with (M*)-(IV*)-interventionism. According to (M*), the path $I \longrightarrow P_1 \longrightarrow P''$ in 5(b) implies that changes induced on P_1 via I are associated with changes in P''. Yet, these changes are also associated with changes in M_1, when all variables apart from $SB(M_1)$ are fixed. (M*) thus implies that M_1 is a direct cause of P''. That is, (M*)-(IV*)-interventionism entails that whenever an (IV*)-intervention on a mental variable M_i is associated with changes in physical effects of $SB(M_i)$, M_i is also a cause of those effects. However, the epiphenomenalist structure 5(b) generates the exact same difference-making relations or correlations under possible interventions as 5(a). To see this, note that the sole difference between graphs

5(a) and 5(b) is the arrow from M_1 to P'' in 5(a) which is missing in 5(b). The only test scenario in which structures 5(a) and 5(b) behave differently is one where P_1 is held fixed while M_1 is manipulated: relative to this scenario, 5(a) entails that P'' changes, whereas 5(b) entails that P'' does not change. But, of course, this test is just what the presence of the supervenience relationship between M_1 and P_1 renders metaphysically impossible. Thus, in all possible (IV*)-intervention tests structures 5(a) and 5(b) behave exactly alike. The two structures are *empirically indistinguishable*.

Since the causal effects of supervening mental variables can never be tested independently of their supervenience bases, this result generalizes: to every causal structure $\mathcal{S}_1$ featuring downward causation there exists a causal structure $\mathcal{S}_2$ not featuring downward causation such that $\mathcal{S}_1$ and $\mathcal{S}_2$ are equivalent with respect to all possible (IV*)-intervention tests, i.e. such that $\mathcal{S}_1$ and $\mathcal{S}_2$ are empirically indistinguishable. Hence, all empirical data that result from (IV*)-interventions and that could stem from mental-to-physical causation might just as well stem from a structure that only features physical-to-physical causation. Nonetheless though, (M*)-(IV*)-interventionism *always prefers* the structure featuring downward causation over the epiphenomenalist alternative. Even in the absence of any empirical evidence (M*)-(IV*)-interventionism has it that epiphenomenalist structures can be discarded. More specifically, (M*) implies:

(‡) Whenever an (IV*)-intervention $I = x_1$ on a mental variable M_i is associated with changes in a physical effect Z of $SB(M_i)$, M_i is a downward cause of Z, notwithstanding the fact that I is connected to Z on a path that does not go through M_i but through $SB(M_i)$.

(‡) can be given a metaphysical and—as we shall see shortly—also a non-metaphysical reading. The metaphysical reading essentially amounts to the claim that *epiphenomenalist structures do not exist*. When read in that way, (‡) is a very bold claim, which can only be maintained if it can be excluded on *a priori* grounds that there exists even one single epiphenomenalist structure of type 5(b), i.e. if it can be excluded that there exists a case such that the correlation of M_i and Z under (IV*)-interventions on M_i is exclusively due to the path through $SB(M_i)$. Such a reading of (‡) would obviously require considerable argumentative backing, which, so far, has not been provided in the literature. And indeed, I do not see even the beginning of an argument as to why correlations of M_i and Z under (IV*)-interventions on M_i should never be due to the path through $SB(M_i)$ only.

However, there is also a non-metaphysical reading of (‡). Woodward (2011, 35) argues that the question whether graphs as the one in figure 5(a) or the one in 5(b) truly represent causal relations in the world is ill-posed from the interventionist perspective. As there does not exist a difference in (IV*)-manipulability relations between 5(a) and 5(b), the interventionist maxim "no causal difference without a difference in manipulability relations" (cf. p. 6 above) yields that there is no causal difference between 5(a) and 5(b) either. According to interventionism, thus, 5(a)

and 5(b) represent *one and the same* causal structure. In consequence, the only remaining question is whether that causal structure is best represented by graph 5(a) or by graph 5(b). That is, (‡) can be read as a principle about the *proper representation* of causal relations. As such, it stipulates that whenever (IV*)-interventions on a mental variable M_i are associated with changes in an effect Z of $SB(M_i)$, representations of the underlying causal structure must feature a path from M_i to Z. Under this purely representational reading, (‡) does not have to be extensively justified, but can simply be introduced as an innocuous *representational convention*. Of course, one might just as well introduce the convention that a causal structure should always be represented by a minimal graph, i.e. a graph with the least amount of edges, which adequately reproduces the empirical behavior of that structure. Such a convention would then universally favor epiphenomenalist representations.

Neither the metaphysical nor the representational reading of (‡) suits the purposes of non-reductive physicalists as Shapiro and Sober (2007), Menzies (2008), Shapiro (2010; 2012), or Raatikainen (2010), who would like to see interventionism as a means to produce empirical evidence in favor of (NR$_3$) and to empirically disprove epiphenomenalism and who take (NR$_3$) to be a claim about causal powers in the world and not about proper representations of causal structures. On the one hand, the metaphysical reading of (‡) is unnecessarily strong for the non-reductive physicalist. The latter just wants to maintain that there exist correlations of mental variables and effects of their supervenience bases that are not exclusively due to physical-to-physical causation. Read metaphysically, however, (‡) yields that such correlations are *never* exclusively due to physical-to-physical causation, but always involve mental-to-physical causation. According to the metaphysical reading, it holds that whoever accepts (M*)-(IV*)-interventionism as an adequate theory of causation, thereby presupposes the nonexistence of epiphenomenalist structures. Against such a background, (M*)-(IV*)-interventionism obviously does not argumentatively resolve the debate between the non-reductive physicalist and the epiphenomenalist, rather, it downright begs the question against the epiphenomenalist. The epiphenomenalist who holds that there exist epiphenomenalist structures of type 5(b), has every reason to simply reject (M*)-(IV*)-interventionism as an inadequate theory of causation.

On the other hand, if (‡) is read representationally it follows that the debate between the non-reductive physicalist and the epiphenomenalist is a mere debate about the proper representation of causal structures: the non-reductive physicalist holds that a proper representation of the structure generating a correlation of M_i and an effect Z of $SB(M_i)$ should feature a path from M_i to Z, whereas the epiphenomenalist maintains that such a path is not necessary. Under this reading of (‡), (M*)-(IV*)-interventionism only supports (NR$_3$) insofar as (NR$_3$) is a claim about proper representations of causal structures. Plainly though, non-reductive physicalism is commonly considered a theory about causal dependencies and powers in the world. Under a representational reading of (‡), (M*)-(IV*)-

interventionism provides no support for a metaphysical reading of (NR_3) whatsoever.

Independently of whether (‡) is read metaphysically or representationally, it is clear that the reason why (M*)-(IV*)-interventionism always prefers structures featuring downward causation as 5(a) over epiphenomenalist structures as 5(b) is entirely non-empirical. Either the discarding of epiphenomenalist structures is based on some *a priori* metaphysical stipulation to the effect that epiphenomenalist structures are inexistent or it is based on a representational convention. In neither case does (M*)-(IV*)-interventionism establish the existence of non-reducible downward causation on *empirical* grounds.

5 *Conclusion*

This paper has shown that it is very difficult to get interventionist support for non-reductive physicalism. The original version of interventionism developed in Woodward (2003) is incompatible with non-reductive physicalism. And even though the newest version presented in Woodward (2011), i.e. (M*)-(IV*)-interventionism, is not only compatible with, but in fact entails (NR_3), the support non-reductive physicalists as Shapiro and Sober (2007), Menzies (2008), Shapiro (2010; 2012), or Raatikainen (2010) can at best hope to receive from (M*)-(IV*)-interventionism is extremely slim.

Putting (M*)-(IV*)-interventionism to use when it comes to backing up the possibility of non-reductive mental downward causation on evidence-based grounds presupposes a methodological principle, *viz*. (†), whose validity is questionable. Also, (M*)-(IV*)-interventionism restricts the 'autonomous' downward causal power of a mental variable M_i to the first physical variable on a causal chain out of $SB(M_i)$. Furthermore, under a metaphysical reading of (‡), the adequacy of (M*)-(IV*)-interventionism *qua* theory of causation presupposes that epiphenomenalist structures can be ruled out on *a priori* grounds, which in all likelihood they cannot. By contrast, under a representational reading of (‡), (M*)-(IV*)-interventionism supplies a representational convention that supports (NR_3) as a maxim about proper representations of causal structures. The fact that a theory of causation that is either inadequate or introduces downward causal paths as mere representational conventions entails (NR_3) is of little help to the non-reductive physicalist. Most certainly, a representational reading of (‡) provides no support for a metaphysical reading of (NR_3) or (CAM) whatsoever.

It must be emphasized again that the results of this paper do not concern the truth or adequacy or plausibility of either (M*)-(IV*)-interventionism or non-reductive physicalism. Rather, they concern the relationship between these two theoretical frameworks and, in particular, the prospects of successfully backing up non-reductive physicalism by interventionist means. In recent years, there has been an interventionist movement in the literature on non-reductive physicalism. Numerous authors have proposed interventionist and evidence-based solutions of

the problem of causal exclusion. The main result of this paper is that the project behind these proposals is bound to fail.*

References

ANTONY, L. M. 2007, 'Everybody has got it: A defense of non-reductive materialism', in B. P. McLaughlin and J. D. Cohen, eds., *Contemporary Debates in Philosophy of Mind*, Malden, MA: Blackwell, pp. 143–159.

ARMSTRONG, D. M. 1997, *A World of States of Affairs*, Cambridge: Cambridge University Press.

BAUMGARTNER, M. 2009, 'Interventionist causal exclusion and non-reductive physicalism', *International Studies in the Philosophy of Science 23*, pp. 161–178.

BAUMGARTNER, M. 2010, 'Interventionism and epiphenomenalism', *Canadian Journal of Philosophy 40*, pp. 359–384.

BENNETT, K. 2004, 'Global supervenience and dependence', *Philosophy and Phenomenological Research 68*, pp. 510–529.

CAMPBELL, J. 2007, 'An interventionist approach to causation in psychology', in A. Gopnik and L. Schulz, eds., *Causal Learning. Psychology, Philosophy, and Computation*, Oxford: Oxford University Press, pp. 58–66.

EHRING, D. E. 1996, 'Mental causation, determinables, and property instances', *Noûs 30*, pp. 461–480.

ERONEN, M. 2012, 'Pluralistic physicalism and the causal exclusion argument', *European Journal for the Philosophy of Science 2*, pp. 219–232.

FODOR, J. 1989, 'Making mind matter more', *Philosophical Topics 17*, pp. 59–79.

FUNKHOUSER, E. 2006, 'The determinable-determinate relation', *Noûs 40*, pp. 548–569.

GILLETT, C. and RIVES, B. 2005, 'The nonexistence of determinables: Or, a world of absolute determinates as default hypothesis', *Noûs 39*, pp. 483–504.

HALPERN, J. Y. and HITCHCOCK, C. 2010, 'Actual causation and the art of modelling', in R. Dechter, H. Geffner, and J. Y. Halpern, eds., *Heuristics, Probability, and Causality*, London: College Publications, pp. 383–406.

HALPERN, J. Y. and PEARL, J. 2005, 'Causes and explanations: A structural-model approach. Part I: Causes', *British Journal for the Philosophy of Science 56*, pp. 843–887.

HARBECKE, J. 2008, *Mental Causation. Investigating the Mind's Powers in a Natural World*, Frankfurt (Main): Ontos.

KIM, J. 1989, 'Mechanism, purpose, and explanatory exclusion', *Philosophical Perspectives 3*, pp. 77–108.

KIM, J. 2003, 'Blocking causal drainage and other maintenance chores with mental causation', *Philosophy and Phenomenological Research 67*, pp. 151–176.

KIM, J. 2005, *Physicalism or Something Near Enough*. Princeton: Princeton University Press.

LIST, C. and MENZIES, P. 2009, 'Non-reductive physicalism and the limits of the exclusion principle', *The Journal of Philosophy 106*, pp. 475–502.

LOEWER, B. M. 2007, 'Mental causation, or something near enough', in B. P. McLaughlin and J. D. Cohen, eds., *Contemporary Debates in Philosophy of Mind*, Malden, MA: Blackwell, pp. 243–264.

LOWE, E. J. 1993, 'The causal autonomy of the mental', *Mind 102*, pp. 629–644.

MACDONALD, G. 2007, 'Emergence and causal powers', *Erkenntnis 67*, pp. 239–253.

MARCELLESI, A. 2010, 'L'interventionnisme permet-il la causalité "descendante"'?' *Igitur 2*, pp. 1–11.

*I am grateful to David Danks, Markus Eronen, Jens Harbecke, Brad Weslake, and Jim Woodward for very helpful comments and/or discussions, and to the anonymous referees of *dialectica* for their valuable comments on earlier versions of this paper. Moreover, I am indebted to the Deutsche Forschungsgemeinschaft (DFG) for generous support of this work (project CAUSAPROBA).

MARCUS, E. 2001, 'Mental causation: Unnaturalized but not unnatural', *Philosophy and Phenomenological Research 63*, pp. 57–83.

MCLAUGHLIN, B. P. 1995, 'Varieties of supervenience', in E. Savellos and U. Yalcin, eds., *Supervenience: New Essays*, Cambridge: Cambridge University Press, pp. 16–59.

MENZIES, P. 2008, 'The exclusion problem, the determination relation, and contrastive causation', in J. Hohwy and J. Kallestrup, eds., *Being Reduced. New Essays on Reduction, Explanation, and Causation*, Oxford: Oxford University Press, pp. 196–217.

MENZIES, P. and LIST, C. 2010, 'The causal autonomy of the special sciences', in C. Mcdonald and G. Mcdonald, eds., *Emergence in Mind*, Oxford: Oxford University Press, pp. 108–128.

ODDIE, G. 1982, 'Armstrong on the Eleatic principle and abstract entities', *Philosophical Studies 41*, pp. 285–295.

PEARL, J. 2000, *Causality. Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

RAATIKAINEN, P. 2010, 'Causation, exclusion, and the special sciences', *Erkenntnis 73*, pp. 349–363.

REISMAN, K. and FORBER, P. 2005, 'Manipulation and the causes of evolution', *Philosophy of Science 72*, pp. 1113–1123.

SHAPIRO, L. 2010, 'Lessons from causal exclusion', *Philosophy and Phenomenological Research 81*, pp. 594–604.

SHAPIRO, L. 2012, 'Mental manipulations and the problem of causal exclusion', *Australasian Journal of Philosophy 90*, pp. 507–524.

SHAPIRO, L. and SOBER, E. 2007, 'Epiphenomenalism. The dos and don'ts', in G. Wolters and P. Machamer, eds., *Thinking about Causes: From Greek Philosophy to Modern Physics*, Pittsburgh: University of Pittsburgh Press, pp. 235–264.

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. 2000, *Causation, Prediction, and Search* (2 ed.). Cambridge: MIT Press.

SPIRTES, P. and SCHEINES, R. 2004, 'Causal inference of ambiguous manipulations', *Philosophy of Science 71*, pp. 833–845.

VAN GULICK, R. 1993, 'Who's in charge here? And who's doing all the work?' in J. Heil and A. R. Mele, eds., *Mental Causation*, Oxford: Oxford University Press, pp. 233–256.

WALTER, S. 2007, 'Determinables, determinates, and causal relevance', *Canadian Journal of Philosophy 37*, pp. 217–244.

WALTER, S. and ERONEN, M. I. 2011, 'Reductionism, multiple realizability, and levels of reality' in S. French and J. Saatsi, eds., *Continuum Companion to the Philosophy of Science*, London: Continuum, pp. 138–156.

WESLAKE, B. unpublished, 'Exclusion excluded', URL: <*http://philpapers.org/rec/WESEE*>.

WIMSATT, W. C. 2007, *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Cambridge: Harvard University Press.

WOODWARD, J. 2003, *Making Things Happen*. Oxford: Oxford University Press.

WOODWARD, J. 2008a, 'Mental causation and neural mechanisms', in J. Hohwy and J. Kallestrup, eds., *Being Reduced: New Essays on Reductive Explanation and Special Science Causation*, Oxford: Oxford University Press, pp. 218–262.

WOODWARD, J. 2008b, 'Response to Strevens', *Philosophy and Phenomenological Research 77*, pp. 193–212.

WOODWARD, J. 2011, 'Interventionism and causal exclusion', *PhilSci-Archive*, URL: <*http://philsci-archive.pitt.edu/8651/*>.

WOODWARD, J. and HITCHCOCK, C. 2003, 'Explanatory generalizations, part I: A counterfactual account', *Noûs 37*, pp. 1–24.

YABLO, S. 1992, 'Mental causation', *Philosophical Review 101*, pp. 245–280.

YANG, E. forthcoming, 'Eliminativism, interventionism and the overdetermination argument', *Philosophical Studies*, pp. 1–20. 10.1007/s11098-012-9856-0.