Interventionist Causal Exclusion and Non-reductive Physicalism

Michael Baumgartner

The first part of this paper presents an argument showing that the currently most highly acclaimed interventionist theory of causation, i.e. the one advanced by Woodward, excludes supervening macro properties from having a causal influence on effects of their micro supervenience bases. Moreover, this interventionist exclusion argument is demonstrated to rest on weaker premises than classical exclusion arguments. The second part then discusses a weakening of interventionism that Woodward suggests. This weakened version of interventionism turns out either to be inapplicable to cases of downward causation involving supervening macro properties or to render corresponding causal claims meaningless. In sum, the paper argues that, contrary to what many non-reductive physicalists claim, interventionism does not render non-reductive physicalism immune to the problem of causal exclusion.

1 Introduction

Since Kim (1989), at least, the *exclusion problem* has provoked a large and still growing literature on the question whether supervening macro properties can have a downward causal influence on effects of their supervenience bases.[1] Opponents of the existence of downward causation involving supervening macro and subvening micro properties have willingly used the exclusion problem to argue their case. In a nutshell, these well-known arguments run somewhat along the following lines: The causal closure of the physical implies that any micro property (or variable or factor or event type) P^* has a causally sufficient micro cause—call it P; suppose now that a macro property M supervenes on P such that $P \neq M$ or, more generally, such that M is not reducible to P. It follows from this setting that either P^* is causally overdetermined by P and M or that M does not cause P^*, i.e. that P is the only cause of P^*. Since cases of causal overdetermination are rather rare, it seems implausible that all micro effects whose micro causes feature supervening macro properties are causally overdetermined. Moreover, ordinary overdetermining causes contribute to their common effect on independent causal paths. As the

Michael Baumgartner is at the Department of Philosophy, University of Bern.
Correspondence to: Department of Philosophy, University of Bern, Laenggassstr. 49a, 3012 Bern, Switzerland. E-mail: baumgartner@philo.unibe.ch.

paths from P to P^* and from M to P^*, however, would not be independent due to the supervenience of M on P, it appears safe to additionally assume, as does e.g. Kim (2005, 41–49), that P^* is not overdetermined by P and M. It then follows that M does not cause P^*.

Exclusion arguments of this sort are widely seen to be primarily directed against non-reductive physicalists, who hold that macro properties supervene on micro properties in a non-reducible manner and who, nonetheless, do not want to settle for the causal irrelevance of macro properties to effects of their supervenience bases (cf. e.g. Marras 2007). In order to counter exclusion arguments, non-reductive physicalists normally argue, in one way or another, that the causal sufficiency of P for P^* does not exclude M from having a causal influence on P^* as well.[2] In consequence, the main obstacle to be overcome in order to secure non-reductive physicalism against the threat posed by exclusion arguments is to account for how exactly supervening properties causally contribute to their downward effects *in addition* to the latter's micro causes. That is, a theory of causation is required that allows for and spells out causal dependencies among supervening macro properties and effects of their supervenience bases. Or as Kim puts the challenge:

> To be a cause of P^*, M must somehow ride piggyback on physical causal chains—distinct ones depending on which physical property subserves M on a given occasion, in the same world or in other possible worlds. And we may ask: In virtue of what relation it bears to physical property P does M earn its entitlement to a free ride on the causal chain from P to P^* and to claim this causal chain to be its own? Obviously, the only significant relation M bears to P is supervenience. But why should supervenience confer this right on M? The fact of the matter is that there is only one causal process here, from P to P^*, and M's supposed causal contribution to the production of P^* is totally mysterious. (Kim 2005, 48)

Recently, numerous non-reductive physicalists have attempted to answer this challenge by drawing on interventionist theories of causation, as most thoroughly and prominently presented in Woodward (2003), which lately have been gaining considerable popularity. For instance, Shapiro and Sober (2007), Shapiro (forthcoming), and Raatikainen (unpublished) claim that downward causal dependencies among supervening macro properties and effects of their supervenience bases can be straightforwardly accounted for by drawing on interventionism. They hold that the interventionist framework immunizes non-reductive physicalism against exclusion arguments and, thus, paves the way for a non-reductionist theory of macro-to-micro causation and even "provides a means by which to test which causal powers a macroproperty has" (Shapiro and Sober 2007, 247).

The first part of this paper takes issue with these claims and sets out to show that non-reductive physicalists' enthusiasm about interventionism has been premature, to say the least. We shall find in section 3 that even though interventionism does not in principle exclude that micro effects such as P^* are overdetermined, the notion of causation developed in Woodward (2003) nonetheless excludes causal dependencies among supervening macro properties and effects of their supervenience bases. The conceptual core of Woodward's acclaimed variant of interventionism is so strong that such macro-to-micro causation is ruled out on a priori grounds. Furthermore, it will turn out that the exclusion argument induced by Woodward's theory—contrary to classical exclusion arguments—does not presuppose micro effects to have causally *sufficient* micro causes. Rather, the interventionist exclusion argument merely assumes that micro effects have micro causes (which do not need to be sufficient). Thus, the first part of this paper is going to demonstrate that the variant of interventionism which is currently thought to be most plausible gives rise to an exclusion argument that rests on premises that are significantly weaker than the premises of traditional exclusion arguments.

While Woodward (2003) does not address the issue of causal dependencies involving supervening properties, Woodward (2008) discusses the problem of handling mental-to-physical causation, i.e. a special case of macro-to-micro causation, in the interventionist framework. As we shall see in the second part of this paper, even though Woodward (2008) only considers the classical exclusion problem and does not address the interventionist exclusion argument as presented in section 3, he seems to weaken his original theory in view of supervening mental phenomena bringing about physical effects. However, section 4 will show that, rather than paving the way for a non-reductionist account of downward causation, Woodward (2008) weakens interventionism to such a degree that either it is no longer applicable to causal dependencies among supervening macro properties and effects of their supervenience bases or it renders corresponding causal claims meaningless, because such downward dependencies violate fundamental presuppositions of the analysis presented in Woodward (2008). All in all, therefore, the paper concludes that, contrary to the standard opinion in the literature, existing interventionist theories of causation are of no use to non-reductive physicalists when it comes to answering the challenge the problem of causal exclusion poses for their position.

In section 2, the core of interventionism as presented in Woodward (2003) is briefly reviewed and the logical form of Woodward's definition of causation along with some important implications of his analysis are clarified. Section 3 then introduces the interventionist exclusion argument and proves its validity. Finally, in section 4, Woodward's (2008) weakening of his original theory is discussed and shown to render interventionism inapplicable to non-reductionist downward causation.

2 Interventionism

The by far most thorough and elaborate presentation of an interventionist theory of causation can be found in Woodward (2003), which, consequently, also constitutes the central point of reference for non-reductive physicalists searching for an interventionist path around exclusion arguments. Woodward's theory turns on two core definitions. First, he defines the notions of a direct and of a contributing cause:

(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; $(\ldots)$ 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)

Against this background, a variable X is a cause of Y iff X is either a direct or a contributing cause of Y. Second, Woodward defines the notion of an intervention variable:

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

1. I causes X;
2. I acts as a switch for all the 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, 98)

Finally, relative to the notion of an intervention variable an (actual) *intervention* can be straightforwardly understood 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).

Woodward is very explicit about the fact that he sees (M) and (IV) to be definitions of the respective notions (cf. e.g. Woodward 2003, 55, 60–61, 98). In regard to defining causation in terms of intervention, and vice versa, he writes:

> In other words, once we fix our representational repertory (i.e., once we choose a set of variables to represent the quantities whose causal relationships we are interested in assessing), then two theories will make different claims about causal relationships among these variables if and only if they make different claims about what will happen under some combination of interventions. Putting this in the form of a slogan, we can say that manipulability accounts are committed to the following: *No causal difference without a difference in manipulability relations, and no difference in manipulability relations without a causal difference*. (Woodward 2003, 61)

Before we turn to the implications of this theory for macro-to-micro causation, several things need to be noted about an account of causation that turns on (M) and (IV). First, as Woodward himself points out repeatedly, it is non-reductive insofar as it does not spell out causation in non-causal terms. Second, a theory based on (M) and (IV) differs from traditional reductive interventionist theories—as e.g. professed by Menzies and Price (1993)—in not involving the notion of human action. (IV) yields a notion of an intervention variable that is thoroughly non-anthropocentric. An intervention variable is solely defined in terms of its causal (and statistical) relations to the other variables in a given structure. Third, it is a variant of a counterfactual analysis of causation because the notion of a *possible* intervention contained in (M), according to Woodward, "should be interpreted to mean that there is some intervention on X such that *if it were possible to intervene to manipulate X repeatedly in that way*, Y (or the probability of Y) would change in some reproducible or repeatable way" (Woodward 2003, 70–71).

Finally and most importantly, as definition (M) plays a crucial role for our subsequent discussion of the interventionist exclusion argument, assessing the latter's validity presupposes a clear understanding of (M)'s logical form. The core of (M) is constituted by an analysis of the notion of a direct cause. What a contributing cause amounts to is then spelled out in terms of direct causation. (M) supplies a necessary and sufficient condition for any two variables to stand in a causal relation. Accordingly, the main formal feature of the core of (M) is a universally quantified biconditional. The left-hand side of this biconditional is constituted by the analysandum "X is a direct cause of Y with respect to a variable set $\mathbf{V}$" which—if variables of causal structures are chosen as domain of quantification—can be symbolized by $C_{\mathbf{V}}xy$ with $C_{\mathbf{V}}$ representing the relation "...is a direct cause of...relative to variable set $\mathbf{V}$". The right-hand side of the biconditional, in turn, provides the analysans of $C_{\mathbf{V}}xy$. It first states (*de re*) the existence of a possible intervention on X with respect to Y which, relative to fixed domain semantics of modal logic, amounts to (*de dicto*) stating the possible existence of an intervention on X with respect to Y: $\Diamond\exists iIixy$, with I standing for the ternary relation "...is an intervention on...with respect to...".[3] Informally, the remainder of the right-hand side states that, while the possible intervention is performed on X and all other variables in $\mathbf{V}$ are held fixed, the value or the probability distribution of Y changes. Taken in combination, the two constituents of the right-hand side of the biconditional in (M) allow for three significantly different readings:

(a) There possibly exists an i such that *if* i is an intervention on X with respect to Y and all other variables are held fixed, *then* Y changes its value or its probability distribution.

(b) There possibly exists an i such that i is an intervention on X with respect to Y *and* if all other variables are held fixed, then Y changes its value or its probability distribution.

(c) There possibly exists an i such that i is an intervention on X with respect

> to Y and all other variables are held fixed *and* Y changes its value or its probability distribution.

In reading (a) "if. . . then" is the main operator within the scope of the existential quantifier, in readings (b) and (c) the latter's scope is governed by "and". While reading (b) features two conjuncts—the second conjunct being a conditional—, reading (c) contains three conjuncts. How these readings affect the truth conditions of (M) and, thus, the analysis of causation provided by (M) is most clearly seen if the three complete logical forms of (M) induced by (a), (b), and (c) are contrasted.[4] By introducing the predicates F representing ". . . is held fixed" and H standing for ". . . changes its value or its probability distribution" and by quantifying over the variables in the set $\mathbf{V}$, reading (a) yields (I), reading (b) yields (II), and reading (c) yields (III) as logical form of (M):

$$\forall x, y, z(C_{\mathbf{V}}xy \leftrightarrow \Diamond\exists i(Iixy \wedge (z \neq x \wedge z \neq y \rightarrow Fz) \rightarrow Hy)) \tag{I}$$

$$\forall x, y, z(C_{\mathbf{V}}xy \leftrightarrow \Diamond\exists i(Iixy \wedge ((z \neq x \wedge z \neq y \rightarrow Fz) \rightarrow Hy))) \tag{II}$$

$$\forall x, y, z(C_{\mathbf{V}}xy \leftrightarrow \Diamond\exists i(Iixy \wedge (z \neq x \wedge z \neq y \rightarrow Fz) \wedge Hy)) \tag{III}$$

The right-hand side of the biconditional in (I) turns out to be true when $Iixy$ or $z \neq x \wedge z \neq y \rightarrow Fz$ are not satisfiable, for, in that case, the antecedent of the conditional within the scope of the existential quantifier is false which renders the conditional as a whole true. That is, if (M) were read in terms of (I), it would determine X to directly cause Y if either it is impossible that there exists an intervention on X with respect to Y or the other variables in the structure cannot be held fixed. Obviously, such a reading would have highly unwelcome consequences. For instance, it is impossible to intervene on the first human step on the lunar surface with respect to Neil Armstrong's step off the lunar module Eagle on July 21, 1969, because, as these are identical events, every direct cause of the first is also a direct cause of the second event. Hence, condition (IV.3) cannot be met. According to (I), therefore, the first human step on the moon would have to be identified as a cause of Neil Armstrong's step off the lunar module. More generally, the fact that it is impossible to intervene on events with respect to themselves would, subject to (I), yield that every event trivially causes itself. Or if there is only one variable among the variables other than X and Y which cannot be held fixed, X would automatically be identified as direct cause of Y. Both of these implications of (I) are unacceptable because they run counter to even the most common pre-theoretical causal intuitions. In contrast, if reading (II) is assumed, the possible existence of an intervention on X with respect to Y turns out to be a necessary condition of a direct causal dependency between these two variables. If there is no such possible intervention, the right-hand side of the biconditional in (II) is false, which amounts to the causal irrelevance of X to Y. While, according to (II), X only directly causes Y if there possibly exists an intervention on X with respect to Y, the fixability of the other variables in $\mathbf{V}$ is not necessary for a direct causal dependency between X and Y. For if not all remaining variables in $\mathbf{V}$ can be held fixed, the antecedent of the conditional in the second conjunct within the scope of the existential quantifier in (II) is false, which renders the whole conditional true. Hence, if there possibly exists an intervention on X with respect to Y and at least one of the remaining variables in the structure cannot be held fixed, both conjuncts of (II) are satisfied which, subject to (II), implies that X directly causes Y. Finally, according to reading (III), both the possible existence of an intervention on X with respect to Y and the fixability of all other variables turn out to be necessary conditions of a direct causal dependency between X and Y. (III) requires that in order for Y to be directly causally dependent on X there possibly exists an intervention on X with respect Y such that all other variables in $\mathbf{V}$ are fixed and Y changes its value or its probability distribution.

(III) is the strongest of all possible readings of (M)—it implies both (I) and (II). Clearly, (I) is an inadequate formal representation of (M) because it completely trivializes the interventionist notion of causation, as shown above. (II) is the formalization that comes closest to the grammatical surface of (M).[5] Nonetheless, formally representing (M) in terms of (III), presumably, best captures the intuitions behind an interventionist account of causation which, as quoted above, aims to establish a tight conceptual connection between manipulability, difference-making in context, and causality. If it is impossible to intervene on X with respect to Y while all other variables are held fixed, interventionism yields that there does not exist a direct causal dependency between X and Y. Therefore, even though Woodward does not explicitly discuss the logical form of (M), which, on the face of it, is ambiguous between (I), (II), and (III), the notion of a direct cause contained in (M) shall subsequently be understood along the lines of reading (III). Nothing substantial, however, hinges on this preference of (III) over (II) for our present purposes, as section 3 will show that the validity of the interventionist exclusion argument does not depend on whether the logical form of (M) is taken to be (II) or (III).

The notion of a contributing cause is then spelled out in terms of direct causation by (M). X is a contributing cause of Y relative to $\mathbf{V}$ iff X is linked to Y by a path of direct causal relationships and a possible intervention on X with respect to Y, performed while all variables in $\mathbf{V}$ not located on the path from X to Y are held fixed, is followed by a change in the value of Y. As X is determined to be causally relevant to Y by (M) iff X is either a direct or a contributing cause of Y, our discussion of the logical form of (M) suggests that (M) identifies *manipulability* of X and *fixability* of the variables in $\mathbf{V}$ that are not located on a path from X to Y as two necessary conditions for X to be a cause of Y. These two conditions will be of particular importance to our subsequent discussion of the interventionist exclusion argument. Let us, hence, label them:

(MAN) There possibly exists an intervention $I = z_i$ on X with respect to Y.

(FIX) The possible intervention $I = z_i$ is such that, while it is performed on X, all variables in the pertaining variable set $\mathbf{V}$ that are not located on a causal path from X to Y are held fixed, i.e. the variables in $\mathbf{V}$ that are not located

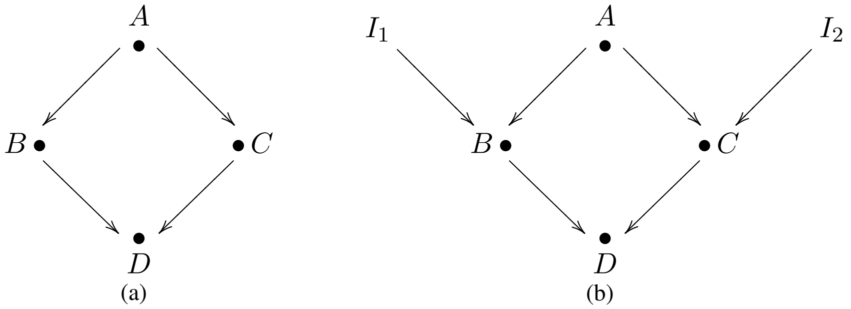

Figure 1: Diagram (a) represents a neuron firing structure with the following properties: If neuron A fires, both B and C fire as well. A firing of B and C, in turn, each is sufficient for neuron D to fire. In diagram (b) two intervention variables for B and C, respectively, are added.

> on a causal path from X to Y *can* be held fixed while $I = z_i$ is performed on X.

If either (MAN) or (FIX) cannot be satisfied by two variables X and Y and the variable set $\mathbf{V}$, X and Y are not causally connected relative to $\mathbf{V}$ according to (M). This reading of (M) shall guide our upcoming considerations with respect to the implications of interventionism for macro-to-micro causation.

Before the next section presents the interventionist exclusion argument, one upshot of (M) and (IV) that is of particular importance for the exclusion problem deserves separate mention. Section 1 has shown that classical exclusion arguments as e.g. presented in Kim (2005) involve an additional assumption to the effect that causal overdetermination is a rare phenomenon which renders it implausible that micro effects are systematically overdetermined by their micro causes and corresponding supervening macro properties. While, for example, standard probabilistic theories of causation indeed exclude systematic overdetermination and, thus, back up Kim's (2005) additional assumption, interventionism is perfectly compatible with certain forms of systematic overdetermination. As we shall see below, the fact that interventionism does not exclude systematic overdetermination in principle might well be one of the primary reasons why this theoretical framework, prima facie, seems so attractive to non-reductive physicalists.

In order to substantiate that interventionism allows for systematic overdetermination, let us consider the causal structure involving neuron firings depicted in diagram (a) of figure 1.[6] This neuron firing structure illustrates the prototypical case of systematic overdetermination. We assume the following dependencies for structure (a): If neuron A fires, both B and C fire as well, and if either B or C fire, neuron D fires too. In such a structure B and C mutually screen each other off from D, for the following conditional independencies hold:

$$p(D|B \wedge C) = p(D|C) = p(D|B) = 1.$$

As is well known, such screening-off relations prevent B and C from being interpreted as causes of D subject to standard probabilistic analyses of causation. In contrast, the interventionist framework does not universally rule out systematic overdetermination. Within this framework, causation is conceptually linked to co-variation under interventions. In order for variable B of our exemplary neuronal structure to be interpretable as a cause of D, it must be assessed whether it is possible to intervene on B while holding C fixed such that D co-varies with corresponding interventions on B—and vice versa for C. We hence need to determine whether there possibly exist two intervention variables I_1 and I_2 for B and C with respect to D such that B can be manipulated by means of I_1 while C is held fixed by means of I_2—and vice versa for C. Structure (b) of figure 1 represents such an expansion of (a) that introduces two suitable intervention variables. If we assume that our exemplary neuronal structure is investigated in a laboratory context, it is very well possible that such intervention variables indeed exist. That is, (MAN) and (FIX) can be satisfied for the structure in figure 1. If D moreover happens to co-vary with corresponding interventions on B and C, the latter variables are determined to cause D subject to (M) and (IV). In that case, firings of B and C *systematically causally overdetermine* firings of D. (M) and (IV), hence, do not rule out systematic overdetermination in principle.[7]

3 An Interventionist Exclusion Argument

Even though interventionism is compatible with certain forms of systematic overdetermination, this section shall show that it nonetheless excludes causal dependencies among supervening macro properties and effects of their supervenience bases. In what follows, I thus present the interventionist exclusion argument.

The argument, roughly, involves the following three premises: (1) causation is to be spelled out in terms of (M), (2) a macro property X supervenes on a physical micro supervenience base $\mathrm{MSB}(X)$ such that $X \neq \mathrm{MSB}(X)$, and (3) $\mathrm{MSB}(X)$ is causally relevant to a micro effect Y. These three premises shall be proven to imply that X does not cause (or is causally irrelevant to) Y. In order to spell out the three premises and the conclusion in more detail, let us draw on the classical case of downward mental causation as illustrated in figure 2. Suppose we are looking at the variable set $\mathbf{V} = \{M, M^*, P, P^*\}$, where M and M^* represent two types of mental phenomena and P and P^* represent the two types of physical phenomena that realize particular values of M and M^*, respectively. M and M^* are taken to supervene on P and P^*. That is, the sets of possible values of P and P^* constitute the physical supervenience bases of M and M^*, i.e. $\mathrm{MSB}(M) = \{P = y_1, P = y_2, \ldots, P = y_n\}$ and $\mathrm{MSB}(M^*) = \{P^* = z_1, P^* = z_2, \ldots, P^* = z_m\}$. It furthermore holds that $M \neq P$ and $M^* \neq P^*$. Finally, P shall be assumed to be causally relevant to P^*. Tailored to this context, the interventionist exclusion argument runs as follows:

$$\begin{tikzcd}
\text{mental} & M \arrow[dr, "?" description] & M^* \\
\text{physical} & P \arrow[u, Rightarrow] \arrow[r] & P^* \arrow[u, Rightarrow]
\end{tikzcd}$$

Figure 2: The classical question of macro-to-micro causation: If the mental phenomenon M supervenes on the physical phenomenon P (symbolized by "$\Longrightarrow$"), which, in turn, is causally relevant to the physical phenomenon P^* (symbolized by "$\longrightarrow$"), in what sense—if at all—can M then be said to be causally relevant to P^*?

(1) M is causally relevant to P^* with respect to the variable set $\mathbf{V} = \{M, M^*, P, P^*\}$ iff there possibly exists an intervention $I_1 = z_1$ on M with respect to P^* such that all other variables in $\mathbf{V}$ that are not located on a causal path from M to P^* are held fixed and the value or the probability distribution of P^* changes.

(2) M supervenes on $\text{MSB}(M) = \{P = y_1, P = y_2, \ldots, P = y_n\}$ without being identical to P.

(3) P is causally relevant to P^* .

$\therefore$ $\neg$(M is causally relevant to P^* with respect to the variable set $\mathbf{V} = \{M, M^*, P, P^*\}$).

(1) is nothing but an adaptation of (M), as spelled out in reading (III), to the variable set $\mathbf{V} = \{M, M^*, P, P^*\}$. In order to determine whether M can be said to be causally relevant to P^* in virtue of (M), we, among other things, need to evaluate whether (MAN) and (FIX) are satisfied in the context at hand. Let us begin with (MAN). Section 2 has shown that an intervention on M with respect to P^* is an intervention variable I_1 for M with respect to P^* taking on some value z_i such that $I_1 = z_i$ causes M to take on some determinate value z_k. A necessary condition for the right-hand side of the biconditional in (1) to be true and, thus, for M to cause P^* is the possible existence of an intervention variable I_1 for M with respect to P^*. In order for a variable I_1 to be such an intervention variable, conditions (IV.1) to (IV.4) must be satisfied. Conditions (IV.1) and (IV.4) are of particular importance for the case at issue here. They require that an intervention variable for M with respect to P^* must cause changes in the values of M while being statistically independent of any further variable that causes P^* and that is located on a directed path to P^* that does not go through M. If it is impossible that there be a variable that satisfies that condition, (MAN) is violated and, hence, M is not causally relevant to P^*. These considerations demonstrate that premise (1)—in which both (M) and (IV) are implicitly contained—implies the following

conditional:

$(1a)$ If M is causally relevant to P^* with respect to the variable set $\mathbf{V} = \{M, M^*, P, P^*\}$, then there possibly exists a variable I_1 that causes changes in the values (or the probability distribution) of M and is statistically independent of any variable Z that causes P^* and that is on a directed path that does not go through M.

The exact meaning and implications of premise (2), of course, hinge on the notion of supervenience presupposed. As is well known, supervenience has been cashed out in a number of very different ways in the literature (cf. e.g. McLaughlin 1995). 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 or variable is necessarily accompanied by a change in the supervenience base. Thus, whatever version of supervenience is taken to be presupposed by (2), the latter implies that M and P differ and that neither M causes P nor vice versa and that changes in M are necessarily accompanied by changes in P. More specifically, (2) entails:

$(2a)$ $M \neq P \wedge \neg(M$ causes $P) \wedge \neg(P$ causes $M)$;

$(2b)$ Every change in the values of M is necessarily accompanied by a change in the values of P.

$(2a)$ and (3) imply:

(4) P is on a causal path to P^* that does not include M.

From the conjunction of $(1a)$ and (4) it follows that, if M is causally relevant to P^*, there possibly exists a variable that causes changes in M while being statistically independent of changes in P. The latter, however, is excluded by $(2b)$, which determines that the values of every variable that induces changes in M will necessarily be correlated with the values of P. Hence, there cannot possibly exist an intervention variable for M with respect to P^*. A straightforward application of modus tollens to $(1a)$ then leads to the conclusion of the interventionist exclusion argument: $\neg(M$ is causally relevant to P^* with respect to the variable set $\mathbf{V} = \{M, M^*, P, P^*\})$ or M is causally irrelevant to P^*, for short. Put differently, M and P^* violate the first necessary condition for M to cause P^* according to reading (III) of (M), *viz.* (MAN).

As can easily be seen from the above considerations, our exemplary case of figure 2 not only violates (MAN) but also (FIX). (4) states that P is located on a causal path to P^* that does not include M, which in virtue of (FIX) requires that P be fixed while M is manipulated. $(2b)$, however, excludes just that fixability, i.e. $(2b)$ excludes that P can possibly be held fixed while M is manipulated. Therefore, M, P, and $\mathbf{V}$ also violate (FIX).

The violations of (MAN) and (FIX) establish the causal irrelevance of M to P^* in virtue of (M). Thus, the interventionist exclusion argument as introduced above is a valid argument. The fact that M, P^*, and $\mathbf{V}$ do not satisfy (MAN) moreover indicates that the validity of this argument does not hinge on whether (M) is spelled out in terms of reading (II) or (III). Subject to both of these readings, (MAN) is a necessary condition for M to cause P^*. In sum, the argument shows that M does not cause P^* in terms of an interventionist notion of causation as advanced in Woodward (2003). Furthermore, since M and M^* represent any properties that supervene on the causally connected variables P and P^*, the conclusion as to the causal irrelevance of M to P^* can be generalized for arbitrary macro properties that supervene on causally connected micro supervenience bases. The argument reveals that (M) rules out the existence of causal dependencies among supervening macro properties, on the one hand, and micro properties that are causally connected to the supervenience bases of the corresponding macro properties, on the other.

This interventionist exclusion argument not only differs from classical exclusion arguments with respect to the fact that it does not involve a premise excluding systematic overdetermination of P^*, it moreover does not presuppose that micro causes are sufficient for their micro effects. The mere causal relevance of P for P^* suffices that P would need to be fixable while M is manipulated in order for the latter to be a cause of P^* in the sense of (M). As we have seen above, such a fixing of P is impossible. In consequence, even though there may well exist countless systematically overdetermined effects and even though micro causes may not fully determine their micro effects, the currently most popular version of interventionism does not allow for any downward causal influence of supervening macro properties.

In order for two variables X and Y to be causally connected, Woodward requires that X has an influence on Y that is *additional* to the influences of all other variables that cause Y. Moreover, it must be possible, at least in principle, to empirically reveal this influence by means of suitable interventions. This condition is not satisfiable for supervening macro properties and effects of their supervenience bases, because it is impossible in principle for any downward causal influence of supervening macro properties to be revealed by means of interventions. In the end, Woodward's interventionism is rooted in very basic empiricist intuitions which require abstaining from causally interpreting dependencies that, for principled reasons, cannot be proven to be of causal nature.

As indicated in the introduction, these findings run counter to the remarkable optimism with which Woodward's interventionism has been received by nonreductive physicalists. Several reasons might be responsible for this overly optimistic reception. First, interventionism does not exclude systematic overdetermination and, thus, prima facie might seem compatible with micro effects being systematically overdetermined by micro and macro causes. Second, as I show in Baumgartner (forthcoming), non-reductive physicalists often read Woodward's interventionism rather loosely. More concretely, Shapiro and Sober (2007), for example, presume that Woodward's theory neither implies (MAN) nor (FIX) to be necessary for causation. I hope this paper has made it clear, though, that subject to any reasonable literal reading of (M) and (IV), (MAN) in fact is necessary for causation—and possibly even (FIX). And third, many of our common intuitions as to the downward causal relevance of macro properties might indeed be rooted in somewhat vague interventionist considerations. For instance, the reason why we are ready to say that Shamus' belief that investing in speculative assets promises high profits causes Shamus' muscles to contract in order to move his arm towards the phone to call his investment banker may well be that, if we intervene—in a pre-theoretical sense of "intervene"—and let Shamus know that the recent credit crunch has seriously hampered the profitability of speculative assets, his muscles relax again. Enlightening Shamus about the credit crunch, however, does not meet the requirements Woodward imposes on interventions for it is necessarily correlated with a change in Shamus' brain state which is another cause of his muscle contractions. Nonetheless, the intuition that informing somebody about a recent economic development is a form of intervention, on the face of it, does not seem too far-fetched.

4 Weakening Interventionism

These strong interventionist intuitions with respect to macro-to-micro causation could be taken to indicate that interventionism does not in principle exclude non-reductionist downward causation. Rather, it might just be Woodward's (2003) variant of the theory that rests on an overly strong conceptual fundament. Hence, non-reductive physicalists with sympathies for interventionism will take the exclusion problem presented in the previous section to demonstrate that Woodward's original version of interventionism must be weakened. While Woodward does not address the issue of causal dependencies involving supervening properties in (2003), in a very recent paper Woodward (2008) discusses the problem of mental-to-physical causation, i.e. of a special kind of macro-to-micro causation, against the background of his interventionist framework. Even though Woodward (2008) only addresses the classical exclusion problem and not the interventionist variant of the problem introduced above and even though his argumentation—as we shall see shortly—is not completely transparent at all points, Woodward (2008) nonetheless seems to weaken his original theory in one respect that is of crucial importance to an interventionist account of downward causation involving supervening properties.[8] In order to determine whether this weakening of interventionism could give non-reductive physicalists a suitable answer to the problem of causal exclusion, let us take a closer look at what Woodward says about mental-to-physical causation in (2008).

After having subscribed to his original notion of an intervention as defined in (IV), Woodward offers the following analysis of causation, which I shall label (M^s) to indicate its condensed and somewhat sketchy nature:[9]

(M^s) X causes Y if and only if there are background circumstances B such that

> if some (single) intervention that changes the value of X (and no other variable) were to occur in B, then Y would change. (Woodward 2008, 222)

On the face of it, "X causes Y" is here not defined in terms of the possible existence of a suitable intervention on X as in (M), but in terms of the counterfactual occurrence of such an intervention. Yet, this is only a notational divergence from (M). As indicated in section 2, Woodward follows standard modal semantics in treating claims about the possible existence of interventions and about their counterfactual occurrence as interchangeable. More substantial than this notational divergence, however, is the fact that the condition as to the counterfactual (or possible) occurrence of an intervention on X unambiguously constitutes the antecedent of an "if…then" clause on the right-hand side of the biconditional in (M^s). That is, Woodward here seems to subscribe to reading (I) of his definition of causation. According to this reading, if—for whatever reason—it is impossible to intervene on X, the counterfactual conditional on the right-hand side of (M^s) is trivially true which renders X a cause of Y. Subject to (M^s), the manipulability of X, i.e. (MAN), is not a necessary condition for X to cause Y, rather, violations of (MAN) turn out to be sufficient for X to cause Y. In section 2 we have seen that such a definition renders interventionism incapable of capturing even the most basic causal intuitions—for instance with respect to the non-ubiquity of self-causation. In order to avoid such a reductio ad absurdum of interventionism, I assume in the following that Woodward does not mean literally what he says here, i.e. that, contrary to the wording of (M^s), he does not really have reading (I) in mind. Unfortunately, though, he does not provide the details of the notion of causation that actually underlies his discussion in (2008).

Woodward's failure to make his intended analysis of causation explicit is even more unfortunate in view of the fact that he apparently does not have any of the remaining readings (II) and (III) of (M) in mind either—as becomes most transparent when he turns to mental-to-physical causation:[10]

> I also assume 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. In other words, I take it to be an implication of (M^s) that a legitimate causal claim should have an intelligible interpretation in terms of counterfactuals [or claims about the possible existence of interventions] the antecedents of which are coherent or make sense. (…) Thus if we have two apparently competing claims, the first contending some mental state is causally inert and the other contending that it causes some outcome, it must be possible to specify some set of (coherent, well-defined) interventions such that the two claims make competing predictions about what would happen under those interventions. If we cannot associate such an interventionist interpretation with one or both of the claims, the claim(s) in question lack a clear sense (…). (Woodward 2008, 224–225)

Plainly, if interventions on X with respect to Y are impossible, none of the readings of (M) (or M^s) discussed in section 2 yields that "X causes Y" is ill-defined or meaningless—contrary to what Woodward claims in this passage. Rather, (I), (II), and (III) assign definite truth-values to causal claims based on, among other things, whether interventions on X are possible or not. If—for whatever reason—it is impossible to intervene on X with respect to Y, (I) and (M^s) render "X causes Y" true, whereas (II) and (III) render "X causes Y" false.

Hence, the quoted passage clearly shows that Woodward (2008) has an account of causation in mind that significantly differs from the theory presented in Woodward (2003). As he does not make his modifications explicit, we are left to reconstruct his weakened version of interventionism on our own. Apparently, Woodward no longer takes (MAN) to be a necessary condition for X to cause Y. If X is not manipulable with respect to Y in the sense of (IV), the claim "X causes Y" shall newly be ill-defined and no longer false.[11] The same holds for (FIX). Woodward (2008, 256) indicates that if it is impossible to intervene on a variable X with respect to Y while holding the other variables in the structure fixed, it does not follow that X is causally irrelevant to Y. Rather, the impossibility to fix the other variables prohibits a coherent or meaningful interpretation of the claim "X causes Y". That means, according to Woodward (2008), the possible existence of an intervention on X with respect to Y and the fixability of the other variables in the structure are *preconditions* of the meaningfulness of claims about causal dependencies among X and Y or about the absence of such dependencies. Rather than being necessary for causation, (MAN) and (FIX) now turn out to be criteria for the well-definedness of interventionist causal claims. These considerations suggest that Woodward (2008) implicitly modifies (M) somewhat along the following lines:

(M') If there possibly exists an intervention $I = z_i$ on X with respect to Y relative to a variable set $\mathbf{V}$ such that $X, Y \in \mathbf{V}$ and such that all other variables in $\mathbf{V}$ that are not located on a path from X to Y are held fixed at some value while $I = z_i$ is performed on X, then X is a (type-level) cause of Y with respect to $\mathbf{V}$ iff Y changes its value or its probability distribution when $I = z_i$ is performed on X.

(M') is weaker than (M), i.e. (M) implies (M') but not vice versa. Contrary to (M), (M') indeed does not assign a truth-value to "X causes Y" when it is impossible to intervene on X with respect to Y or to fix the other variables in the structure. Against the background of (M'), the possibility to intervene on X while the other variables are held fixed can—in the vein of the passage quoted above—be argued to be a precondition of "X causes Y" being a well-defined interventionist causal claim. This consequence of (M') can be given a strong and a weak reading. Subject to the strong reading, whenever (MAN) or (FIX) are violated there is no objective fact of the matter whether "X causes Y" is true or not. In contrast, according to the weak reading, the existence of a causal dependency between X and Y simply cannot be assessed within the interventionist framework in cases of violations of (MAN) or (FIX). That, however, does not exclude that X might be identified as cause of Y within some other theoretical framework in such cases. Or differently, if (MAN) and (FIX) are violated the strong reading of (M') implies that "X causes Y" is meaningless, while the weak reading only implies that the interventionist framework is inapplicable to determining the truth-value of that causal claim.

Woodward seems to favor the strong reading. Still, relative to either of the two readings, the interventionist exclusion argument put forward in the previous section does not go through any longer if (M) is replaced by (M') in premise (1) of the argument. For according to both readings of (M'), neither (MAN) nor (FIX) are necessary conditions for causation. That is, if (MAN) or (FIX) are violated, (M') does not entail that "X causes Y" is false. The validity of the interventionist exclusion argument essentially hinges on the fact that supervening macro properties and effects of their supervenience bases violate (MAN) and (FIX) for conceptual or metaphysical reasons—depending on the notion of supervenience presupposed. If neither (MAN) nor (FIX) are necessary for causation, such violations no longer imply the downward causal inertness of supervening macro properties.

Yet, it is plain that even though Woodward's (2008) weakening of interventionism blocks the exclusion argument presented in section 3, (M') is far from serving the purposes of non-reductive physicalists. The latter want to claim that supervening macro properties are *causally relevant* to effects of their supervenience bases, i.e. they hold that "M causes P^*" is meaningful and moreover true. Non-reductive physicalists with sympathies for interventionism additionally claim that the truth of "M causes P^*" can be assessed by means of interventionism. Subject to the strong reading of (M'), however, such claims are simply meaningless or ill-defined, because such downward causal dependencies violate (MAN) and (FIX). Subject to the weak reading, interventionism is not applicable to non-reductionist downward causation. In consequence, irrespective of whether one adopts the strong or the weak reading of (M'), Woodward's weakening of interventionism does not account for downward causation as would be required by non-reductive physicalists.

Furthermore, although replacing (M) by Woodward's strong reading of (M') in premise (1) would block the argument advanced in the previous section, such a replacement would immediately give rise to another interventionist argument targeting non-reductive physicalism, *viz.* to an argument showing the meaninglessness of the claim that there exist cases of macro-to-micro causation which, after all, is one of the core tenets of non-reductive physicalism. The result of replacing (M) by the strong reading of (M'), call it $(1')$, in combination with (2) and (3) entails that all statements of type "M causes P^*" are ill-defined or meaningless. All in all, thus, neither the interventionist theory presented in Woodward (2003) nor the one suggested in Woodward (2008) secures non-reductive physicalism against the challenge posed by the problem of causal exclusion.

5 Conclusion

In sum, the first part of this paper has shown that the currently most highly acclaimed version of interventionism, i.e. the one presented in Woodward (2003), gives rise to an argument that rules out the possibility of supervening macro properties causally influencing effects of their supervenience bases. This interventionist exclusion argument differs from classical exclusion arguments as the one advanced in Kim (2003) or Kim (2005) in two important respects: First, Woodward's interventionist notion of causation (M) does not rule out the overdetermination of micro effects in principle and, second, the interventionist exclusion argument only presupposes that there exists a micro cause for every micro effect, whereas classical exclusion arguments additionally assume that these micro causes are causally *sufficient* for their effects. That is, one of the presently most influential theories of causation gives rise to an exclusion argument that is based on premises that are considerably weaker than the ones of classical exclusion arguments.

The second part of the paper has revealed that, even though the original interventionist exclusion argument can be blocked by replacing (M) by the weakened version of interventionism suggested in Woodward (2008), (M') nonetheless does not provide the basis for an interventionist theory of non-reductionist downward causation. According to (M'), downward causal claims involving supervening variables are either meaningless or not analyzable within the interventionist framework. In either case, non-reductive physicalism is in no way supported by interventionism. Non-reductive physicalists wanting to account for macro-to-micro causation in interventionist terms, hence, have to reject both the theoretical framework presented in Woodward (2003) and the one indicated in Woodward (2008). This finding, of course, does not entail that there may not exist different weakenings of Woodward's original theory—weakenings that at the same time block the interventionist exclusion argument and yield a fruitful account of non-reductionist downward causation. Yet, instead of taking on the difficult task of modifying interventionism to suit their purposes, non-reductive physicalists tend to endorse Woodward's interventionist account in an unqualified manner. This paper has shown that this endorsement is thoroughly counterproductive. It radically undermines non-reductive physicalism, rather than securing it against exclusion arguments.

Acknowledgements

I am grateful to Delphine Chapuis, Mehmet Elgin, Derek Turner and Jim Woodward for very helpful comments and discussions. Moreover, I thank two anonymous referees of this journal for valuable comments on an earlier draft. Finally, I am indebted to the Center for Philosophy of Science of the University of Pittsburgh and to the Swiss National Science Foundation for generous support of this work (grant PBBE1 – 117031).

Notes

[1] For recent versions of the problem cf. Kim (2003) or Kim (2005). Moreover, while the problem was originally seen to only threaten the special case of mental-to-physical causation, it has meanwhile been generalized for any kind of macro-to-micro causation (cf. Bontly 2002, Kim 2005, 55).

Finally, note that a causal and an explanatory version of the exclusion problem can be distinguished (cf. Sabates 2001). This paper will only be concerned with the causal variant. For a persuasive overview over different forms and readings of exclusion arguments cf. Walter (2008).

[2] Cf. e.g. Yablo (1992), Jackson (1996), Horgan (2001), or very recently List and Menzies (unpublished).

[3] In modal systems containing the Barcan Formula (BF) $\Diamond\exists x Fx$ is even equivalent to $\exists x\Diamond Fx$. For details see e.g. Hughes and Cresswell (1996, 246). Irrespective of whether BF is presupposed, I take a *de dicto* reading of (M) that sidesteps metaphysical questions as to the manner of existence of possibilia to be more in line with the basic non-metaphysical approach followed in Woodward (2003). The subsequent discussion of the interventionist exclusion argument, however, in no way hinges on this preference of a *de dicto* reading of (M).

[4] Strictly speaking, of course, (a), (b), and (c) bring about differences in the logical form of the analysis of *direct causation* contained in (M). For brevity, I subsequently simply speak of *the logical form of (M)* in this context.

[5] While Woodward, in various passages (e.g. Woodward 2003, 112–113), explicitly characterizes the manipulability of X with respect to Y as a necessary condition of X causing Y, the fixability of the remaining variables is never explicitly identified as necessary condition for causation.

[6] For such "neuron diagrams", which have meanwhile become a classical source of exemplary causal structures, cf. Lewis (1986).

[7] As Woodward (2003, 82–85) shows, his account of *token-level* causation is compatible with systematic overdetermination as well.

[8] To be clear, Woodward (2008) does not himself claim to be weakening or even modifying his original theory. Yet, as shall be substantiated below, that is what he seems to be doing nonetheless—very implicitly.

[9] Woodward indicates that, for simplicity, (M^s) only accounts for deterministic causal dependencies. This restriction, however, is of no relevance to our current purposes.

[10] To be consistent with the notation used in this paper I change "(M)" to "(M^s)" in this quotation.

[11] To make this contrast between the original and the weakened version of Woodward's theory most transparent, compare the passage from Woodward (2008) quoted above with Woodward (2003, 112–113), where manipulability is characterized as necessary condition for causation.

References

Baumgartner, M. forthcoming. Interventionism and epiphenomenalism. *Canadian Journal of Philosophy*.

Bontly, T. D. 2002. The supervenience argument generalizes. *Philosophical Studies* 109: 75–96.

Horgan, T. 2001. Causal compatibilism and the exclusion problem. *Theoria* 15: 95–116.

Hughes, G. E., and M. J. Cresswell. 1996. *A new introduction to modal logic*. London: Routledge.

Jackson, F. 1996. Mental causation. *Mind* 105: 377–413.

Kim, J. 1989. Mechanism, purpose, and explanatory exclusion. *Philosophical Perspectives* 3: 77–108.

——— 2003. Blocking causal drainage and other maintenance chores with mental causation. *Philosophy and Phenomenological Research* 67: 151–176.

——— 2005. *Physicalism, or something near enough*. Princeton, NY: Princeton University Press.

Lewis, D. 1986. Postscript to 'causation'. In D. Lewis, *Philosophical Papers*, vol. 2, 172–213. Oxford: Oxford University Press.

List, C., and P. Menzies. unpublished. *Non-reductive physicalism and the limits of the exclusion principle*. ⟨ URL: http://eprints.lse.ac.uk/20118/ ⟩ Accessed on 17 February 2009.

Marras, A. 2007. Kim's supervenience argument and nonreductive physicalism. *Erkenntnis* 66: 305–327.

McLaughlin, B. P. 1995. Varieties of supervenience. In *Supervenience: New essays*, edited by E. Savellos and U. Yalcin, 16–59. Cambridge: Cambridge University Press.

Menzies, P., and H. Price. 1993. Causation as a secondary quality. *British Journal for the Philosophy of Science* 44: 187–203.

Raatikainen, P. unpublished. Causation, exclusion, and the special sciences. ⟨ URL: http://philsci-archive.pitt.edu/archive/00003879/ ⟩ Accessed on 17 February 2009.

Sabates, M. H. 2001. Varieties of exclusion. *Theoria* 16: 13–42.

Shapiro, L. forthcoming. Lessons from causal exclusion. *Philosophy and Phenomenological Research*.

Shapiro, L., and E. Sober. 2007. Epiphenomenalism. The dos and don'ts. In *Thinking about causes: From Greek philosophy to modern physics*, edited by G. Wolters and P. Machamer, 235–264. Pittsburgh, PA: University of Pittsburgh Press.

Walter, S. 2008. The supervenience argument, overdetermination, and causal drainage: Assessing Kim's master argument. *Philosophical Psychology* 21: 673–696.

Woodward, J. 2003. *Making things happen*. Oxford: Oxford University Press.

——— 2008. Mental causation and neural mechanisms. In *Being reduced: New essays on reductive explanation and special science causation*, edited by J. Hohwy and J. Kallestrup, 218–262. Oxford: Oxford University Press.

Yablo, S. 1992. Mental causation. *Philosophical Review* 101: 245–280.