The Inherent Empirical Underdetermination of Mental Causation

Michael Baumgartner

Abstract

It has become a popular view among non-reductive physicalists that it is possible to devise empirical tests generating evidence for the causal efficacy of the mental, whereby the exclusion worries that have haunted the position of non-reductive physicalism for decades can be dissolved once and for all. This paper aims to show that these *evidentialist* hopes are vain. I argue that, if the mental is taken to non-reductively supervene on the physical, there cannot exist empirical evidence for its causal efficacy. While causal structures without non-reductive supervenience relations can be conclusively identified in ideal discovery circumstances, it is impossible, in principle, to generate evidence that would favor models with mental causation over models without. Ascribing causal efficacy to the mental, for the non-reductive physicalist, is a modeling choice that must be made on the basis of metaphysical background theories or pragmatic maxims guiding the selection among empirically indistinguishable models.

1 Introduction

When it comes to conceptualizing the relationship between the mental and the physical, the position of non-reductive physicalism has much appeal to many, both in philosophy and in science. It is committed to the following tenets: ($\mathcal{N}1$) the domain of the physical is causally complete, ($\mathcal{N}2$) mental properties supervene on physical properties without being reducible (identical) to the latter, and ($\mathcal{N}3$) some mental properties are causally efficacious. However, the notorious causal exclusion argument, as most famously advanced by Kim (e.g. in 2005), exposes a dangerous tension between the first two tenets and the third. It is unclear how mental properties can influence other mental properties, given that the latter are fully determined by their physical supervenience bases, and how they can impact on the causally complete domain of the physical, given that the mental is distinctly non-physical. If it is moreover assumed that the putative effects of the mental are not systematically overdetermined, it follows that mental causation in the sense of non-reductive physicalism is downright impossible.

While the debate on this problem has been conducted in the arena of metaphysical theorizing for decades, it has seen an *evidentialist turn* in recent years. Numerous non-reductive physicalists have argued that the proper arena to settle the question as to the causal efficacy of the mental is not metaphysics but empirical science. According to Shapiro and Sober (2007), Shapiro (2010, 2012), Raatikainen (2010, 2013), Menzies (2008), Campbell (2007, 2010), Woodward (2008a), or Andersen (2009) the existence of non-reductive mental causation can be, and indeed has been, established empirically. The core idea behind all of these proposals is that the methodological protocol entailed

by Woodward's (2003) popular interventionist theory of causation—*interventionism*, for short—allows for designing a test that generates evidence for non-reductive mental causation.

This paper takes issue with this evidentialist movement. My aim is to show that if the mental is taken to non-reductively supervene on the physical, there does not exist a test that could generate evidence favoring models with mental causation over models without. More specifically, I will argue that according to both the original version of interventionism presented in Woodward (2003) as well as the theory's latest amendments in Woodward (2015) the test design envisaged by evidentialists is unrealizable. Still, interventionism allows for laying out the blueprint of the best conceivable test for mental causation that is *de facto* realizable. It turns out, however, that all the evidence that can be generated by this best test underdetermines the inference to non-reductive mental causation. While there exist ideal discovery circumstances generating unambiguous evidence for ordinary causal structures, i.e. structures without supervenience relations, all the evidence on non-reductive mental causation, even if generated under optimal discovery circumstances, can equally be accounted for by models featuring mental causation and models without. For principled reasons, there is no evidence-based fact of the matter whether there exists non-reductive mental causation.

I conclude that ascribing causal efficacy to the mental, for the non-reductive physicalist, is a modeling choice that must be made based on metaphysical background theories or pragmatic maxims guiding the selection among empirically indistinguishable models. While this result, on the one hand, entails that the question as to the causal efficacy of the mental falls into the scope of philosophy after all, despite recent attempts at delegating the issue to empirical science, it, on the other hand, should also be taken as a reason to defuse the intensity with which debates on the causal efficacy of the mental are conducted in philosophy. For whether mental properties are modeled by variables with exiting causal arrows, ultimately, is only of representational relevance.

The paper is organized as follows. Section 2 reviews the test design that evidentialists believe to generate evidence for mental causation. In section 3, I show that tests of this design are unrealizable and introduce the best realizable test, which, however, turns out to inherently underdetermine the inference to non-reductive mental causation in section 4. Finally, section 5 reviews different approaches to resolve the underdetermination non-empirically.

2 Evidentialist Reasoning

Before reviewing the evidentialist reasoning, a conceptual clarification is required. Mental causation comes in two variants: mental-to-mental and mental-to-physical causation. In the exclusion debate, the latter variant is commonly assumed to be primary in the sense that the possibility of mental-to-mental causation presupposes the possibility of mental-to-physical causation, but not vice versa (e.g. Kim 2005, 20). Hence, if mental-to-physical causation is impossible, the same holds for mental-to-mental causation. As the bulk of the evidentialist literature, therefore, focuses on mental-to-physical causation,

I will also develop my main argument against that background. Section 4 will generalize my findings for mental-to-mental causation.

The basic idea of evidentialists is that the causal efficacy of mental properties can be empirically revealed by intervening on them and recording ensuing changes in other (mental or physical) properties. Some evidentialists, for instance Campbell (2010, 16-18), Raatikainen (2010), and Andersen (2009, 210-227), believe that tests uncovering mental causation can be designed in the same vein as tests for ordinary causation. According to this view, the standard test design—reflecting Mill's method of difference—to determine whether a mental variable M_1 (i.e. a variable representing a mental property) is a cause of a physical variable P_2 requires surgically intervening on M_1 by ensuring that the intervention I is not connected to P_2 on a causal path around M_1 and homogenizing (e.g. holding fixed) all causes of P_2 that are not located on a path through I and M_1, that is, all *off-path* causes of P_2. In an experimental setting, this design is realized by actual manipulations of relevant variables (Raatikainen 2010; Campbell 2010), whereas in observational settings background influences are homogenized away, for example, by randomization as in RCTs (Campbell 2010) or by appropriate conditionalization as in Bayes nets methods (Spirtes et al. 2000).

Figure 1 provides a graphical representation of this standard test design applied to the Kim-style structure, around which, traditionally, debates on mental causation revolve. M_1 and M_2 are two mental variables with their corresponding physical supervenience bases P_1 and P_2, the former of which is assumed to be causally sufficient for the latter subject to the causal completeness of the physical. Relative to that setting, the question of mental-to-physical causation is whether M_1 is a cause of P_2, despite the latter's complete causal determination by P_1. Since non-reductive physicalists endorse the non-reducibility of mental properties, P_1 does not represent the same property as M_1. The two variables, hence, are non-identical. As (directed) causal paths are ordered n-tuples of variables (Spirtes et al. 2000, 8-9), the pair $\langle P_1, P_2 \rangle$ corresponds to a different path than the scrutinized path $\langle M_1, P_2 \rangle$. This, in turn, entails that P_1 is an off-path cause of P_2 that, according to the standard test design, must be fixed (or otherwise homogenized), which is represented by square brackets in figure 1. The standard test then amounts to manipulating M_1 by means of some intervention variable I. A correlation of M_1 and P_2 under such a manipulation is evidence for M_1 being a cause of P_2.

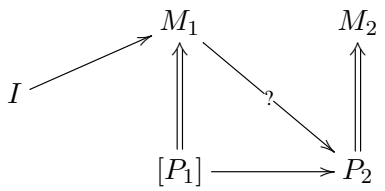

Figure 1: Standard test design for ordinary causation applied to the case of mental causation. "$\rightarrow$" stands for direct causation, "$\Uparrow$" for non-reductive supervenience, "$[P_i]$" for homogenizing P_i.

Before discussing the applicability of this test to non-reductive mental causation, a caveat is needed. Figure 1 models real-life structures involving mental causation in a very condensed manner. In particular, mental properties typically have complex supervenience bases featuring many physical properties and intricate mechanisms among them. Representing all of this by single variables P_1 and P_2, although standard in the literature on causal exclusion, amounts to a far-reaching simplification. A natural way of interpreting Kim-style diagrams is to view P_1 as a coarse-grained variable that takes a value p_i iff the mechanism represented by P_1 is in an overall state that realizes a corresponding value m_i of the mental variable M_1, such that $P_1 = p_i \rightarrow M_1 = m_i$, for all i—and analogously for P_2 and M_2.

Against that backdrop, it is easily seen that the test in figure 1 cannot be performed—neither in its experimental nor its observational variant. Due to the supervenience of M_1 on P_1, it is not possible to intervene on M_1 while holding P_1 fixed; nor is it possible to randomize or homogenize P_1 across variations in M_1; nor to conditionalize on specific values of P_1 while investigating whether changing values of M_1 affect the probability distribution of P_2. Whenever P_1 is fixed, either by intervention or conditionalization, so is M_1; and whenever P_1 is randomized, so is M_1.

Some evidentialists are well aware of the unrealizability of the test in figure 1. For instance, Shapiro and Sober (2007) and Shapiro (2010, 2012) criticize Kim's exclusion argument as relying on an erroneous premise according to which that test is realizable after all. According to Shapiro and Sober (2007, 241), the test in figure 1 is *irrelevant* to determining the causal efficacy of mental properties because it examines whether M_1 has an influence on P_2 *in addition* to P_1, which, for Shapiro and Sober, is the wrong question to ask. They contend that even though M_1 does not have an influence on P_2 beyond P_1, it can still cause P_2. The test they deem appropriate for uncovering this kind of non-supplemental influence and, hence, tailor-made for evaluating the causal efficacy of non-reductively supervening properties is sketched in the following passage from Shapiro (2010).

> The right test to perform will hold fixed, *not* the supervenience base of M_1, but the common cause that M_1 and P_2 share. The common cause of M_1 and P_2 will precede M_1 [...]. P_0, not P_1, is the common cause of M_1 and P_2. This means that it is P_0 that should be held fixed when testing whether M_1 is a cause of P_2. But what happens when P_0 is held fixed and M_1 is wiggled? Because changing M_1 is impossible without simultaneously changing M_1's supervenience base P_1, and because P_1 is a cause of P_2, a change in M_1 *does* result in a change in P_2. *This is evidence that M_1 is a cause of P_2.* (Shapiro 2010, 600-601 [with adapted notation])

Even though this quote only mentions the fixing of a single common cause of M_1 and P_2, *viz.* P_0, it is clear that the general test design laid out here requires that *all* common causes of M_1 and P_2 be fixed (cf. also Shapiro and Sober 2007, 238-239). That is, in order to test whether M_1 is a cause of P_2, not M_1's supervenience base must be fixed but all common causes of M_1 and P_2. Prima facie, there could exist two types of common causes of M_1 and P_2: (i) common causes that influence P_2 via P_1 and (ii) common causes that influence P_2 directly. In figure 2, which depicts the test design advanced by Shapiro

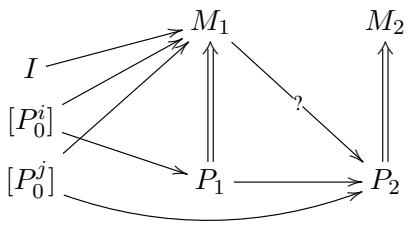

Figure 2: Evidentialist test for mental causation. "$[P_0^i]$" and "$[P_0^j]$" represent the homogenizing of all common causes of M_1 and P_2.

and Sober, P_0^i is a placeholder for all common causes of the first type and P_0^j stands for all common causes of the second type. The test for non-reductive mental causation, thus, amounts to this:

Evidentialist Test: To test whether a mental variable X is a cause of an effect Y of its own supervenience base, fix all common causes of X and Y and intervene on X; a correlation of X and Y under such an intervention is evidence for X being a cause of Y.

Before the next section assesses whether the Evidentialist Test serves its purpose, let us briefly review the theoretical framework in which Shapiro and Sober (2007, 237, 256) embed their proposal: Woodward's (2003) *interventionism*, which currently is one of the most popular theories of causation. Interventionism spells out causation in terms of difference-making under suitable interventions. More specifically, the theory—as presented in Woodward (2003)—provides the following (type-level) *definition of causation (M)*: X causes Y with respect to (w.r.t.) a set of analyzed variables $\mathbf{V}$ iff there exists a possible intervention on X w.r.t. Y that is associated with a change in Y when all off-path causes of Y in $\mathbf{V}$ are held fixed (Woodward 2003, 59). (M) is complemented by a *definition of intervention (IV)*: a variable I taking one of its values, $I = i_n$, is an intervention on X w.r.t. Y iff $I = i_n$ surgically fixes the value of X without having an impact on Y that is not mediated via X and without being correlated with any off-path causes of Y (Woodward 2003, 98).

3 Evidentialist Reasoning Revisited

To assess the prospects of the Evidentialist Test, note first that it not only requires to fix certain variables, *viz.* the common causes of the target variables, but also to intervene on a specific variable, *viz.* the mental variable. An intervention must be surgical, which means in particular that it must be independent of all off-path causes of the ultimate outcome. Applied to the structure in figure 2, this means that a cause of M_1 only counts as an intervention on M_1 w.r.t. P_2 if it is uncorrelated with all off-path causes of P_2. However, we have already seen in the previous section that there exists at least one off-path cause of P_2 with which all causes that induce changes in M_1 are necessarily correlated, *viz.* P_1; and, for the non-reductive physicalist, P_1 is not identical with M_1, meaning it is located

on another causal path to P_2. In consequence, there cannot exist a surgical intervention on M_1 w.r.t. P_2, which, in turn, entails that the Evidentialist Test cannot be performed after all, notwithstanding the fact that it does not demand the fixing of P_1.

This is an instance of a more general problem that has been pointed out by Baumgartner (2010): macro properties that non-reductively supervene on micro properties cannot be intervened on relative to effects of their supervenience bases because they are necessarily correlated with changes in their supervenience bases, which are non-identical to the former and thus off-path with respect to downstream effects. Furthermore, as the interventionist definition of causation (M) imposes intervenability as a necessary condition for causation, the impossibility to intervene on non-reductively supervening macro properties entails their causal inertness. While this problem does not affect macro properties that are taken to be reducible to underlying micro properties (cf. Baumgartner 2013), mental properties constitute a particularly recalcitrant instance of the problem because their non-reducibility seems exceedingly plausible. In consequence, not only is it impossible to perform the Test but, according to interventionism, which is the very theoretical context in which the Test is embedded, its non-performability entails that M_1 does not cause P_2. In other words, interventionism gives rise to its very own causal exclusion argument, which, to make matters worse, rests on considerably weaker premises than the standard Kim-style argument, as it neither assumes an exclusion principle nor the causal completeness of the physical (cf. Baumgartner 2010). That means the project of devising empirical tests for mental causation, which defines the evidentialist movement, is completely superfluous because interventionism entails that the mental is causally inert *prior to all evidence*.

Some authors believe that this interventionist exclusion argument is an artifact of an ill-chosen variable set (e.g. Eronen 2012; Yang 2013; Raatikainen 2013). The reason offered is that, in ordinary contexts of causal discovery, it is common to demand that analyzed variable sets satisfy what Woodward (2015, 316) calls the principle of *Independent Fixability*, which requires that variables can, at least in principle, be set to all their values independently of each other. Independent Fixability is clearly violated by the set $\mathbf{V} = \{M_1, P_1, M_2, P_2\}$, for the mental variables and their supervenience bases, in principle, cannot be set to all value configurations. Accordingly, Eronen (2012), for instance, claims that if the supervenience relations were removed by selecting proper subsets $\mathbf{V}' = \{M_1, P_2\}$ and $\mathbf{V}'' = \{M_1, M_2\}$, the interventionist exclusion argument would be blocked and interventionism would straightforwardly establish M_1 as a cause of P_2 and M_2, respectively.

However, that suggestion is due to a misapprehension of the definition of an intervention (IV). Contrary to the definition of causation (M), (IV) is *not relativized* to a variable set. "[T]he intervention must be uncorrelated with *all* potential confounders, not just with all confounders that happen to be in some variable set" (Woodward 2008b, 202). In other words, whether some manipulation I counts as an intervention on X w.r.t. Y depends on the existence of a causal influence of I on Y that is not mediated via X *in the world* and not in some suitably chosen model. By way of example, taking the pill does not count as an intervention on pregnancy w.r.t. thrombosis, irrespective of whether the analyzed variable set contains a variable for oestrogen ratio in the blood. The reason is that, in the

world, the pill causally influences thrombosis by increasing the oestrogen ratio, that is, on a causal route around pregnancy.[1] Analogously, as it follows from non-reductive physicalism that P_1 has an influence on M_2 on a causal route not containing M_1, all putative interventions on M_1 w.r.t. P_2 need to be independent of changes in P_1, even if the system is analyzed using variable sets, as $\mathbf{V}'$ or $\mathbf{V}''$, that satisfy Independent Fixability. The main corollary that drives the interventionist exclusion argument, *viz.* that (IV)-defined interventions on M_1 w.r.t. P_2 are impossible, holds regardless of a chosen variable set. To avoid interventionist exclusion, it does not suffice to tweak the analyzed variable set, rather the core definitions of interventionism must be modified.

By now, there are numerous suggestions available as to how best to modify these definitions (Eronen and Brooks 2014; Woodward 2015; Weslake forthcoming). The details of these proposal diverge considerably, but their consequences for our current purposes amount to the same: an appropriate version of interventionism must neither entail that P_1 has to be held fixed nor that interventions on M_1 w.r.t. P_2 must be independent of changes in P_1. In Woodward's (2015) latest version of interventionism, to which I shall subsequently refer as *interventionism**, this goal is achieved by introducing exemption clauses for supervenience relations, more specifically, by modifying (IV) and (M) as follows:

(IV*) A variable I taking one of its values, $I = i_n$, is an intervention on X w.r.t. Y iff $I = i_n$ fixes the value of X without having an impact on Y that is not mediated via X *or via a variable, which is related in terms of supervenience to* X *and* Y*,* and without being correlated with any off-path causes of Y, *except for those related in terms of supervenience to* X *and* Y.

(M*) X causes Y w.r.t. $\mathbf{V}$ iff there is a possible (IV*)-intervention on X that changes Y when all off-path variables in $\mathbf{V}$ are held fixed, *except for those related in terms of supervenience to* X *and* Y.

Interventionism* does not demand that interventions on M_1 w.r.t. P_2 be independent of *all* off-path causes of P_2, but only of those that are not related in terms of supervenience to M_1 and P_2. The fact that all putative interventions on M_1 are necessarily correlated with changes in P_1, hence, no longer renders interventions on M_1 w.r.t. P_2 impossible. It follows that interventionism* does not exclude the causal efficacy of the mental on *a priori* grounds, which, in turn, renders the evidentialist project of devising an empirical test for mental causation meaningful in the first place.

The next question then becomes what the prospects of the Evidentialist Test are when it is interpreted against the background of interventionism*. To answer this, assume that a particular variable I taking one of its values, $I = i_n$, is an (IV*)-intervention on M_1 w.r.t. P_2 such that changing M_1 via $I = i_n$ is associated with a change in P_2 when all off-path causes of P_2 except for P_1 are held fixed. As M_1 supervenes on P_1, the change induced on M_1 via $I = i_n$ is necessarily associated with a change in P_1. If it moreover holds—as it should in homogeneous contexts of causal discovery—that all off-path causes

[1] Correspondingly, interventionism would not be an adequate theory of causation if the notion of an intervention were relativized to a variable set (Woodward 2008b, 201-203; Baumgartner 2013, 11-13).

of P_1 are fixed, (M*) entails that I is not only a cause of M_1 but also of P_1. This can be realized in one of two ways: either I causes M_1 and P_1 along *one* causal path, say, $I \longrightarrow M_1 \longrightarrow P_1$, or along *two* paths, $M_1 \longleftarrow I \longrightarrow P_1$. The former option is excluded since the instances of M_1 are realized or constituted by the instances of P_1, meaning they spatiotemporally overlap and, thus, stand in a *non-causal* form of dependence.[2] In light of the non-identity of M_1 and P_1 and the standard definition of (directed) causal paths in terms of ordered n-tuples of variables, it follows that I causes M_1 and P_1 along two different paths, *viz.* $\langle I, M_1 \rangle$ and $\langle I, P_1 \rangle$ with $M_1 \neq P_1$. In other words, I is a common cause of M_1 and P_1.

This argument can be generalized. The non-reductive supervenience of the mental on the physical yields that every cause inducing a change on the mental level is necessarily associated with a change on the physical level and, hence, is a cause of both the mental and physical changes. Moreover, due to the non-causal nature of the relationship between the mental and its physical supervenience base, all causes of the mental are common causes of the mental and their supervenience bases. That is, *all* (IV*)-interventions on M_1 necessarily are common causes of M_1 and P_1. M_1 can only be manipulated with a *fat hand* (cf. Baumgartner and Gebharter (2015)).[3]

A side corollary of this result is that *direct* common causes of M_1 and P_2 in the vein of P_0^j in figure 2 do not exist; rather, all common causes of M_1 and P_2 have the structural features of P_0^i. More importantly, the systematic fat-handedness of interventions on mental variables entails that the Evidentialist Test cannot be performed, even if the surgicality constraint is lifted in the vein of interventionism*. It is impossible to hold all common causes of M_1 and P_2 fixed and still intervene on M_1 w.r.t. P_2, for fixing all common causes of M_1 and P_2 is tantamount to fixing all candidate intervention variables. When all common causes are homogenized, there is nothing left to wiggle M_1 with.

In order to intervene on a mental variable, at least one of the common causes it shares with its supervenience base must be allowed to change. If the common causes of a mental phenomenon and its underlying physical mechanism cannot all be homogenized when testing the former's causal efficacy, methodological prudence still demands that they be homogenized *as much as possible*. Optimally, all common causes of M_1 and P_2 are fixed *except for one*; and this unrestrained common cause is then used as (IV*)-intervention to test whether M_1 is a cause of P_2. That is, the best conceivable test for non-reductive mental causation that can actually be performed—in optimal laboratory circumstances—is the following:

Evidentialist Test*: To test whether a mental variable X is a cause of an effect Y of its own supervenience base, fix all common causes of X and Y except for one (IV*)-intervention variable I and intervene on X via I; a correlation of X and Y under such an (IV*)-intervention is evidence for X being a cause of Y.

[2] For an extended argument as to why constitution is a non-causal form of dependence see Craver and Bechtel (2007).

[3] A fat-handed intervention is an intervention that influences its effects along two (or more) different causal paths (Scheines 2005, 931-32).

If there is any way to generate empirical evidence for non-reductive mental causation, it must be on the basis of a test design along the lines of Test*. The reason is that all tests with stricter homogeneity requirements cannot be performed on systems featuring mental properties that are assumed to non-reductively supervene on their physical realizers. Hence, the crucial follow-up question—to be answered in the next section—is whether performing Test* fulfills its purpose.

4 Underdetermination

To determine whether Test* can produce empirical evidence for mental causation, consider a concrete application of it as depicted in figure 3(a). All common causes P_0^i of M_1 and P_2, except for one, are held fixed (represented by square brackets), I being the one non-fixed variable. I satisfies the constraints of an (IV*)-intervention variable on M_1 w.r.t. P_2: it is a cause of M_1 whose only influence on P_2 that goes around M_1 is mediated via M_1's supervenience base P_1, and it is uncorrelated with all off-path causes of P_2, except for P_1. Let us assume that wiggling M_1 via I is associated with changes in P_2, that is, our exemplary Test* generates a (perfect) correlation of M_1 and P_2. Does such a correlation amount to evidence in favor of M_1 causing P_2?

Even though this instance of Test* is performed against a *maximally homogenized* causal background, the fact remains that I is a common cause of M_1 and P_2—and fat-handed interventions generate confounded data, which are uninformative as regards the relationship between the targeted variables. More concretely, the correlation between M_1 and P_2 resulting from our exemplary Test* can be fully accounted for by the mere fact that the two variables are wiggled with a fat hand. Hence, there is no need at all to stipulate the existence of an additional causal dependence between M_1 and P_2. Model (a) in figure 4, which features M_1 as cause of P_2, and the epiphenomenalist model 4(b), which does not, imply the very same correlations under manipulations of M_1 via I in contexts where all other common causes P_0^i are held fixed.

As all interventions on non-reductively supervening mental properties necessarily are fat-handed, the above argument can be repeated for any (IV*)-intervention on M_1 w.r.t. P_2 that results in a correlation of M_1 and P_2. Relaxing the constraints imposed on interventions along the lines of interventionism* entails that correlations of mental variables

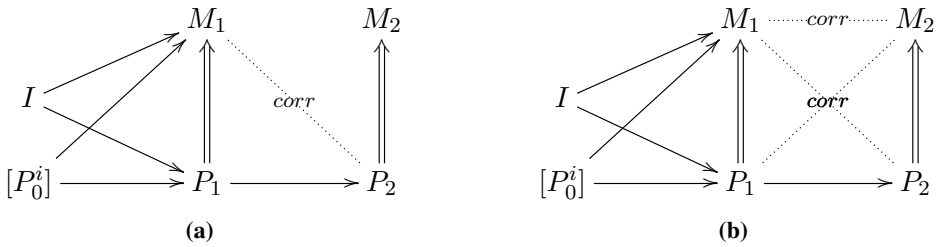

Figure 3: Two applications of Test* where P_0^i is a placeholder for all common causes of M_1 and P_2 except for one, and I is that one remaining common cause. Dotted lines represent correlation.

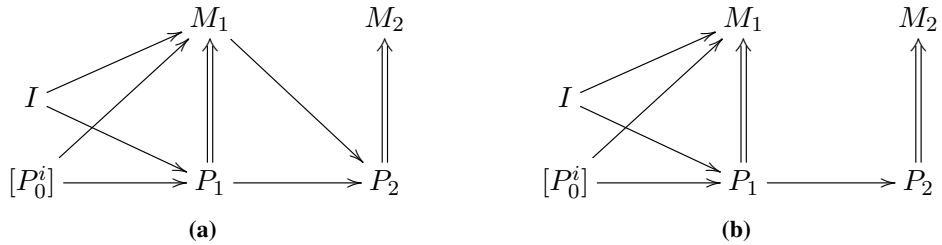

Figure 4: Two models that equivalently account for the Test*-result in figure 3(a).

and downstream effects of their supervenience bases can always be accounted for by the mere fat-handed nature of corresponding interventions—even under optimal laboratory conditions. Data produced by applications of Test* do not provide a rationale for an inference to mental-to-physical causation. For every model featuring mental-to-physical causation there exists a pure common-cause model (i.e. a model with causally inert mental properties) that entails the very same correlations under interventions in maximally homogenized backgrounds and, hence, cannot be distinguished from the former model empirically.

The inference to mental-to-mental causation is empirically underdetermined in an analogous manner. To see this, assume that another instance of Test* generates a correlation not only of M_1 and P_2 but also of M_1 and M_2, as depicted in figure 3(b). In that case, the change induced on P_1 via I is associated with changes in both P_2 and M_2 (when all off-path causes are fixed). It follows from interventionism* that P_1 is not only a cause of P_2 but also of M_2, meaning there exists a causal path from I via P_1 to M_2 that goes around M_1. Yet, although the change in I is a common cause of the changes in M_1 and M_2, it nonetheless passes as (IV*)-intervention on M_1 w.r.t. M_2, for P_1 represents the supervenience base of M_1. As in case of mental-to-physical causation, however, such a fat-handed intervention reveals nothing determinate on the relationship between M_1 and M_2. One viable model that accounts for a correlation of M_1 and M_2 indeed features mental-to-mental causation, as depicted in figure 5(a); but there likewise exists an epiphenomenalist model without mental causation that accounts for such a correlation equally well, *viz.* the one in figure 5(b). The correlation between M_1 and M_2 is either due to a causal dependence or to the fat-handed nature of I, but the Test*-result in figure 3(b) provides no rationale whatsoever for preferring one option over the other.

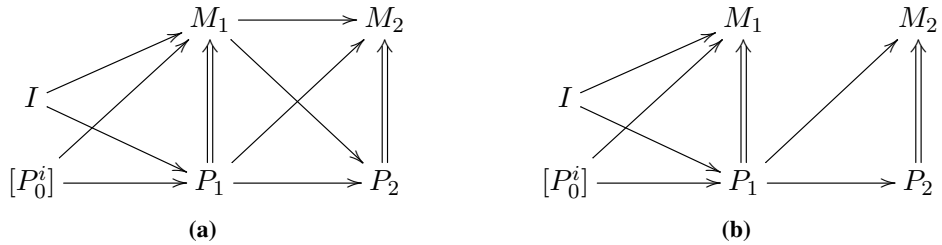

Figure 5: Two models that equivalently account for the Test*-result in figure 3(b).

Analogously to the interventionist exclusion argument, the empirical underdetermination of non-reductive mental causation is not an artifact of an ill-chosen variable set. Any intervention with a causal influence on the ultimate outcome that is not mediated via its direct target variable generates confounded data, if the influence around the target is not suppressed. This fact is completely independent of whether the preferred model contains a variable on the causal path around the target. More concretely, a study investigating the connection between pregnancy and thrombosis will produce confounded data, if pregnancy is manipulated via pill intake and the oestrogen ratio is not controlled for—whether or not the implemented model features a variable for oestrogen ratio. The same holds in the case of non-reductive mental causation. Subject to the assumed non-reducibility of the mental to the physical, all (IV*)-interventions on M_1 w.r.t. P_2 have a causal influence on P_2 that is not mediated via M_1 in the world, and not merely in some poorly designed model. As this influence on P_2 around M_1 cannot be suppressed, all (IV*)-interventions inevitably produce confounded data. That is, the empirical underdetermination of mental causation is due to the way in which the non-reductive physicalist presumes the analyzed properties to interact and interdepend *in the world*.

What is more, the inference to non-reductive mental causation is empirically underdetermined in a way that differs from underdetermination known from ordinary causal inference. It is a frequent phenomenon in causal discovery that empirical data can be accounted for by different causal models that score equally well on all parameters of model fit (cf. e.g. Spirtes et al. 2000, 59-72). Such underdetermination ultimately stems from our limited capacities for controlling background influences in ordinary discovery contexts. The resulting noise in real-life data yields that the latter, ever so often, do not unambiguously reflect underlying causal structures. But this common form of empirical underdetermination can be resolved in ideal discovery contexts. In the absence of non-reductive supervenience relations, causal structures satisfy Independent Fixability—as causal interactions can be broken, that is, effects can be suppressed via suitable interventions even after their causes have occurred. In such contexts, it is possible to surgically intervene on scrutinized causes. It follows that cause-effect pairs can be isolated from confounding background influences, meaning that ordinary contexts of causal discovery can be idealized to such a degree that crucial experiments become possible that produce unconfounded data providing conclusive evidence for causal dependencies. The paradigm example of such an ideal discovery context is the experimental setting envisaged in Mill's method of difference. If surgical interventions on a variable X are associated with changes in a variable Y, when *all* other causes of Y—and not just all other causes not subject to some exemption clause—are fixed, it conclusively follows that X is a cause of Y.

Such ideal discovery circumstances do not exist for mental causation as conceived by the non-reductive physicalist. Manipulating mental properties is tantamount to manipulating their non-identical physical supervenience bases, which are part of a causally complete causal network. Contrary to causal relationships, however, the dependence between mental properties and their physical realizers cannot be broken and, hence, renders the principle of Independent Fixability unsatisfiable. It is impossible to surgically intervene on mental properties and to isolate pairs of mental properties and downstream effects.

As a consequence, even in ideal discovery contexts in which the causal background is maximally homogenized, it is impossible to produce unconfounded data furnishing evidence for mental causation. When investigating the potential effects of non-reductively supervening mental properties, data confounding is introduced by the very manipulations intended to uncover mental causation. Data produced by the best realizable test for mental causation—Test*—can always be accounted for by a model with mental causation and one without. In contrast to the case of ordinary causation, there cannot exist an *experimentum crucis* for the causal efficacy of the mental. Being empirically underdetermined is an *inherent* feature of non-reductive mental causation. Contrary to the hopes of evidentialists, the debate between the friends of mental causation and their epiphenomenalist opponents cannot be settled empirically.

5 Resolving the Underdetermination

That this debate cannot be decided empirically does not entail that it cannot be decided in other ways. Accordingly, this section discusses non-empirical approaches to resolve the inherent underdetermination of the inference to non-reductive mental causation.

A first type of approach is pursued by Woodward's (2015) interventionism*. According to (M*) and (IV*), Test*-results as depicted in figure 3 entail mental causation. If (IV*)-interventions on M_1 are associated with changes in P_2 (and M_2), when all off-path causes are fixed, it follows from (M*) that M_1 is a cause of P_2 (and M_2). That is, interventionism* prefers models 4(a) and 5(a) over their empirically indistinguishable epiphenomenalist rivals. More generally, interventionism* entails *universal mental causation* in the following sense: Whenever an (IV*)-intervention $I = i_n$ on a mental variable M is associated with changes in an effect Y of M's supervenience base, M is a cause of Y, even if I is connected to Y on a path that does not go through M but through its supervenience base. Or differently, according to interventionism*, a correlation of M and Y is never due to the fat-handedness of pertinent (IV*)-interventions but is always due to causation.

This general preference for models with mental causation yields that, if there exists even a single epiphenomenalist structure of type 4(b) or 5(b) in the world, interventionism* is an inadequate theory of causation, for it would in that case erroneously ascribe causal relevance to the mental. In other words, endorsing the adequacy of interventionism* *qua* theory of causation is tantamount to assuming the nonexistence of epiphenomenalist structures featuring causally inert mental properties that non-reductively supervene on their physical realizers. That is, interventionism* resolves the empirical underdetermination of non-reductive mental causation by excluding the epiphenomenalist models on *a priori* grounds. Woodward (2015, 340-342) justifies this with a metaphysical principle that is deeply embedded in the whole interventionist framework: "no causal difference without a difference in manipulability relations". By virtue of this principle, models as the ones in figures 4 or 5, which behave alike under all possible manipulations, do not represent two but only one causal structure. The only remaining question, for Woodward (2015, 342-344), then is whether diagrams of types

4(a) and 5(a) or of types 4(b) and 5(b) are more suitable to represent that underlying structure. He opts for the diagrams featuring mental variables at tails of causal arrows because they render difference-making and manipulability relations more transparent.

This is not the place for a detailed assessment of this proposal. A few observations suffice for our purposes. First, the general preference for models featuring mental causation squares nicely with the basic ideas behind interventionism. Thus, although the empirical underdetermination is resolved by means of a metaphysical principle, that principle is not *ad hoc* but theoretically motivated. However, second, this proposal does not resolve the debate between friends of mental causation and epiphenomenalists by acknowledging both positions as viable options and then adjudicating between them. Instead, it simply denies the viability of the position defended by the epiphenomenalist camp, which, accordingly, will hardly be convinced by this line of reasoning. Third, and most importantly, Woodward's (2015) proposal does not serve the purposes of the evidentialists, as it does not pave the way for establishing the existence of non-reductive mental causation on evidence-based grounds. Rather, its underlying metaphysical (or definitional) commitments simply stipulate that difference-making relations in which mental variables are involved never have epiphenomenalist roots.[4]

It is plain that the same type of approach to resolving the empirical underdetermination is also available to epiphenomenalists. Instead of relying on a theory of causation whose metaphysical commitments prefer models with mental causation, they can invoke a theory imposing principles in favor of epiphenomenalist models. A principle to that effect would, for instance, be this: "causal structures do not contain redundant elements". There exist well-known theories of causation implementing that principle in some form or other. For example, Suppes' (1970) probabilistic theory of causation or the theory underlying Bayes-net procedures of causal inference (Spirtes et al. 2000) define causes to be non-redundant probability-changers of their effects, *viz.* probability-changers for which no off-screeners exist.[5] Applied to our concrete example, this theory entails that, even if M_1 turns out to be a probability-changer of P_2, it does not pass as cause of P_2, because P_1, which is a sufficient cause of P_2 subject to the causal completeness of the physical, screens off P_2 from all other variables including M_1 and, hence, precludes the existence of any other causes (Gebharter 2015). Any theory of causation implementing such a non-redundancy principle gives preference to epiphenomenalist models as 4(b) and 5(b) because they reproduce the Test*-data in figure 3 in a redundancy-free manner, whereas models 4(a) and 5(a) feature redundant causal connections.

All the previous remarks on the interventionist resolution of the underdetermination apply to this epiphenomenalist proposal in reverse. The non-redundancy principle stipu-

[4]To be fair, this is my reconstruction of how interventionism* resolves the empirical underdetermination of mental causation. Woodward (2015) himself does not view interventionism* as establishing mental causation by means of a definitional fiat. As he contends that mental causation models and their epiphenomenalist rivals do not represent different causal structures, he denies that there is empirical underdetermination to be begin with. A consequence of this view, however, is that the ongoing debate between non-reductive physicalists and epiphenomenalists is a meaningless pseudo-debate over the truth of causal models that do not represent different structures. Plainly, this perspective likewise undermines the evidentialist project, for there is no reason to design tests that could adjudicate between positions that do not differ in the first place.

[5]A variable X *screens off* two variables A and B iff $\mathrm{P}(B|A) \neq \mathrm{P}(B)$ and $\mathrm{P}(B|X \wedge A) = \mathrm{P}(B|X)$.

lates the nonexistence of causal structures with redundant elements on *a priori* grounds; and as epiphenomenalist structures generally feature less causal dependencies than their empirically indistinguishable rivals with mental causation, it follows that structures of the latter type are excluded for containing redundancies. Likewise, the non-redundancy principle is not *ad hoc* but has a long tradition in philosophical theorizing about causation. At the same time, the principle does not resolve the debate between friends of mental causation and epiphenomenalists in an adjudicative manner but simply puts a stop to it by denying the viability of the former position. And clearly, it also does not establish the nonexistence of mental causation on evidence-based grounds. Rather, it stipulates that difference-making relations in which mental variables are involved are always due to the fat-handedness of relevant interventions.

An alternative approach, which does not resolve the underdetermination in an *a priori* manner, goes back to Yablo (1992) and has recently been reactivated by List and Menzies (2009). The idea is to demand that causes be *proportional* to their effects, meaning that changes in the causes are always associated with changes in the effects and that these changes are related by a constant multiplier. List and Menzies advance proportionality as a requirement causes *must* satisfy. That is, if it turns out that the values of M_1 but not the ones of P_1 are proportional to the values of P_2, List and Menzies conclude that M_1 is a cause of P_2, whereas P_1 is not. Imposing proportionality as necessary condition on causation can thus yield downward exclusion, which, in turn, invalidates the causal completeness of the physical. This strong proportionality approach has been criticized by many (e.g. Bontly 2005; Shapiro and Sober 2012; Woodward forthcoming; Hoffmann-Kolss 2014). The main objection is that only a small subset of real-life causal structures actually satisfy proportionality. More often than not, the values of causes and effects are related by some non-linear function that violates proportionality. Accordingly, subject to all standard (type-level) theories of causation, a variable X counts as a cause of another variable Y if there exists at least one configuration of background conditions α in which X makes a (correlational/counterfactual/probabilistic etc.) difference to Y—even if changes in X are not associated with changes in Y in all configurations different from α.

Although proportionality must, hence, not be imposed as a necessary condition on causation, it can still be advanced as a *pragmatic* maxim that helps to resolve cases of empirical underdetermination. More concretely, if the Test* in figure 3 reveals that M_1 is proportional to P_2 or M_2, the models featuring mental causation can be argued to be preferable over their epiphenomenalist counterparts, because they render the tight interdependence between M_1 and P_2/M_2 more transparent. They are more informative regarding manipulation and control.

As recourse to pragmatic maxims is a common strategy to choose among empirically indistinguishable models, other maxims might be brought to bear as well. For instance, the pragmatic virtue of *simplicity* gives preference to epiphenomenalist models. They account for the data generated by Test* by introducing fewer dependencies than the models with mental causation. Another pragmatic virtue to be considered is *coherence* with, say, standard theoretical commitments in a scientific community. In light of the widespread acceptance of mental causation in the disciplines investigating the mental, coherence

could be argued to resolve the underdetermination in favor of models featuring mental variables at tails of causal arrows. Other pragmatic maxims as *predictive strength* and *explanatory power* could likewise be employed, but their ramifications for the selection among models with mental causation and models without are less straightforward and, thus, cannot be assessed here. What matters for our purposes is that there exist alternatives to resolving the empirical underdetermination by means of metaphysical principles. Some pragmatic maxims of model selection prefer models with mental causation, while others favor their epiphenomenalist rivals.

All in all, when it comes to resolving the underdetermination of the inference to mental causation non-empirically, both camps in the debate on the causal efficacy of non-reductively supervening mental properties have well established metaphysical principles and corresponding theories of causation as well as common pragmatic maxims of model selection at their disposal that are favorable to their respective positions.

6 Conclusion

In the past 10 years, numerous philosophers with sympathies for non-reductive physicalism have argued that it is possible to devise empirical tests that produce evidence in favor of the causal efficacy of the mental. This paper has shown, first, that the test designs envisaged by these evidentialists are unrealizable and, second, that the best realizable test—Test*—is incapable of generating evidence that would in any way favor models with mental causation over models without. Moreover, we have seen that the inference to non-reductive mental causation is empirically underdetermined for principled reasons: even in ideal discovery circumstances no *experimentum crucis* for mental causation is possible.

Still, some theories of causation have a built-in preference for models with mental causation and other theories for models without. Likewise, some pragmatic maxims of model selection are favorable to mental causation while others support epiphenomenalism. All of this shows that in order to make headway in the debate between friends of non-reductive mental causation and their epiphenomenalist opponents it is inevitable to carefully weigh up the pros and cons of the different theories of causation and pragmatic maxims. This has not been the place to take on this task. For the argument of this paper it suffices to stress that it is not a task that can be fulfilled by conducting experiments and collecting data. It is not a task for empirical science but for a meta-level discipline as philosophy.

Plainly, this result will not come as a surprise to metaphysically minded philosophers. By contrast, it should force empirically minded philosophers to recognize that ordinary causal claims not involving supervenience relations have a fundamentally different epistemological status than causal claims involving non-reductively supervening mental properties. While the truth of the former can be established empirically, there simply is no evidence-based fact of the matter whether there exists non-reductive mental causation. For the empirically minded philosopher, modeling mental properties by variables with exiting causal arrows should only be of instrumental and not of factual

relevance—it should be a mere modeling choice. As such, it must not be accompanied by heated discussions.

Acknowledgements

I am grateful to the audiences at the workshop on *Emergence, Exclusion and Causation*, University of Glasgow, April 2016, and at the conference on *Causality in the Sciences of the Mind and Brain*, University of Aarhus, June 2016. Moreover, I thank three anonymous reviewers for their helpful comments and suggestions.

Funding

Research for this article was supported by the Swiss National Science Foundation, grant number PP00P1_144736.

References

Andersen, H. (2009). *The Causal Structure of Conscious Agency*. Ph. D. thesis, University of Pittsburgh. http://d-scholarship.pitt.edu/9254/.

Baumgartner, M. (2010). Interventionism and epiphenomenalism. *Canadian Journal of Philosophy 40*, 359–384.

Baumgartner, M. (2013). Rendering interventionism and non-reductive physicalism compatible. *Dialectica 67*, 1–27.

Baumgartner, M. and A. Gebharter (2015). Constitutive relevance, mutual manipulability, and fat-handedness. *British Journal for the Philosophy of Science*. doi: 10.1093/bjps/axv003.

Bontly, T. D. (2005). Proportionality, causation, and exclusion. *Philosophia 32*(1), 331–348.

Campbell, J. (2007). An interventionist approach to causation in psychology. In A. Gopnik and L. Schulz (Eds.), *Causal Learning. Psychology, Philosophy, and Computation*, pp. 58–66. New York: Oxford University Press.

Campbell, J. (2010). Control variables and mental causation. *Proceedings of the Aristotelian Society (Hardback) 110*(1pt1), 15–30.

Craver, C. and W. Bechtel (2007). Top-down Causation Without Top-down Causes. *Biology & Philosophy 22*(4), 547–563.

Eronen, M. (2012). Pluralistic physicalism and the causal exclusion argument. *European Journal for the Philosophy of Science 2*, 219–232.

Eronen, M. I. and D. S. Brooks (2014). Interventionism and supervenience: A new problem and provisional solution. *International Studies in the Philosophy of Science 28*(2), 185–202.

Gebharter, A. (2015). Causal exclusion and causal Bayes nets. *Philosophy and Phenomenological Research*. doi: 10.1111/phpr.12247.

Hoffmann-Kolss, V. (2014). Interventionism and higher-level causation. *International Studies in the Philosophy of Science 28*(1), 49–64.

Kim, J. (2005). *Physicalism or Something Near Enough*. Princeton: Princeton University Press.

List, C. and P. Menzies (2009). Non-reductive physicalism and the limits of the exclusion principle. *The Journal of Philosophy 106*, 475–502.

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*, pp. 196–217. New York: Oxford University Press.

Raatikainen, P. (2010). Causation, exclusion, and the special sciences. *Erkenntnis 73*, 349–363.

Raatikainen, P. (2013). Can the mental be causally efficacious? In K. Talmont-Kaminski and M. Milkowski (Eds.), *Regarding the mind naturally: naturalist approaches to the sciences of the mental*, pp. 138–166. Newcastle Upon Tyne: Cambridge Scholars Publisher.

Scheines, R. (2005). The similarity of causal inference in experimental and non-experimental studies. *Philosophy of Science 72*(5), 927–940.

Shapiro, L. (2010). Lessons from causal exclusion. *Philosophy and Phenomenological Research LXXXI*, 594–604.

Shapiro, L. and E. Sober (2007). Epiphenomenalism. The dos and don'ts. In G. Wolters and P. Machamer (Eds.), *Thinking about Causes: From Greek Philosophy to Modern Physics*, pp. 235–264. Pittsburgh: University of Pittsburgh Press.

Shapiro, L. and E. Sober (2012). Against proportionality. *Analysis 72*(1), 89–93.

Shapiro, L. A. (2012). Mental manipulations and the problem of causal exclusion. *Australasian Journal of Philosophy 90*, 507–524.

Spirtes, P., C. Glymour, and R. Scheines (2000). *Causation, Prediction, and Search* (2nd ed.). Cambridge: MIT Press.

Suppes, P. (1970). *A Probabilistic Theory of Causality*. Amsterdam: North-Holland.

Weslake, B. (forthcoming). Exclusion excluded. *International Studies in the Philosophy of Science*.

Woodward, J. (2003). *Making Things Happen. A Theory of Causal Explanation*. New York: 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*, pp. 218–262. New York: Oxford University Press.

Woodward, J. (2008b). Response to Strevens. *Philosophy and Phenomenological Research LXXVII*, 193–212.

Woodward, J. (2015). Interventionism and causal exclusion. *Philosophy and Phenomenological Research 91*(2), 303–347. doi: 10.1111/phpr.12095.

Woodward, J. (forthcoming). Intervening in the exclusion argument. In H. Beebee, C. Hitchcock, and H. Price (Eds.), *Making a Difference*. Oxford: Oxford University Press.

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

Yang, E. (2013). Eliminativism, interventionism and the overdetermination argument. *Philosophical Studies*, 321–340.