This is the penultimate draft of an article has been published in
(2012) *European Journal for Philosophy of Science* 2(2), 219-232.

Pluralistic Physicalism and the Causal Exclusion Argument[1]

Markus I. Eronen
Ruhr-Universität Bochum
Institut für Philosophie II; GA03/150
Universitätsstraße 150
D-44780 Bochum
Germany

Abstract

There is a growing consensus among philosophers of science that scientific endeavors of understanding the human mind or the brain exhibit explanatory pluralism. Relatedly, several philosophers have in recent years defended an interventionist approach to causation that leads to a kind of causal pluralism. In this talk, I explore the consequences of these recent developments in philosophy of science for some of the central debates in philosophy of mind. First, I argue that if we adopt explanatory pluralism and the interventionist approach to causation, our understanding of physicalism has to change, and this leads to what I call pluralistic physicalism. Secondly, I show that this pluralistic physicalism is not endangered by the causal exclusion argument.

[1] This article is based on chapters 9-11 of my PhD thesis *Reduction in Philosophy of Mind: A Pluralistic Account* (Ontos, 2011).

1. Introduction

There is a growing consensus among philosophers of science that scientific endeavors of understanding the human mind or the brain exhibit *explanatory pluralism*. According to explanatory pluralism, science essentially involves explanations at different analytical levels and of different kinds, and will continue to do so in the future. Furthermore, interlevel and cross-discipline connections have a fundamental role in the advancement of science. Explanatory pluralism is thus an alternative both to strong reductionism and to the kind of antireductionism that claims that the special sciences are totally independent from the physical sciences.

In a related development, several philosophers have in recent years presented accounts of causation in terms of interventions and manipulability. This "interventionist" account of causation has received wide acceptance from both the philosophical and the scientific community. Its core idea is that causal relationships are relationships that are potentially exploitable for purposes of manipulation and control.

My aim in this paper is to explore the consequences of these recent developments in philosophy of science for some of the central debates in philosophy of mind. First, I argue that if we adopt explanatory pluralism and the interventionist approach to causation, our understanding of physicalism has to change, and this leads to what I call *pluralistic physicalism*. Secondly, I show that this pluralistic physicalism is not endangered by the causal exclusion argument. Although my focus is on the philosophy of the cognitive sciences, the considerations may also apply more broadly.

In the next section, I will discuss explanatory pluralism in more detail and give a rough definition. In section 3, I will briefly present the interventionist approach to causation. In section 4, I will argue that if we accept explanatory pluralism and the interventionist account, this naturally leads to what I call pluralistic physicalism. Finally, in section 5, I will argue that the causal exclusion argument is not a problem for this kind of pluralism.

2. Explanatory pluralism

Recently several philosophers of neuroscience, biology, and psychology have defended *explanatory pluralism* as an approach to the relations between sciences and different analytical levels (e.g., Bechtel (2008), Brigandt (2010), Craver (2007), Looren de Jong (2002), McCauley & Bechtel (2001), Mitchell (2003), Wimsatt (1976, 2007)). I take the core of explanatory pluralism to consist of the following four theses:

(1) For full understanding of human behavior (or the mind), explanations of different *kinds* are necessary
(2) For full understanding of human behavior (or the mind), explanations at different *levels* are necessary
(3) Successful explanations remain explanatory even when corresponding lower-level explanations are complete
(4) Interlevel connections and explanatory integration across disciplines are essential in explanatory enterprises

Thesis (1) is an acknowledgement of the fact there is no single pattern or structure to which all scientific explanations conform. Historically speaking, the most influential model of scientific explanation has been the deductive-nomological model (Hempel & Oppenheim 1948). For a long time it was hoped that this model, or at least something very similar, would capture the general pattern of scientific explanations. Unfortunately, these hopes were dashed, as it turned out that most scientific explanations do not fit the model. In fact, it is fairly clear that scientific explanations are too heterogeneous to fit any single model. Also when explaining the human mind or brain, we shouldn't expect the explanations to conform to a single pattern: we need, to name a few, mechanistic, causal, computational, and evolutionary explanations.

The second thesis reflects the fact that focusing on just one level of analysis is in most cases insufficient for full understanding of the phenomenon of interest. Levels are

here best understood as the "levels of mechanisms" (Bechel 2008; Craver 2007) in the system or phenomenon under consideration. For example, in order to understand the memory consolidation mechanism, we need to consider several compositional levels, and none of these levels is fundamental or sufficient for full understanding of the phenomenon. For instance, the molecular level is not sufficient, because we also need to understand the functional role of the mechanism and where it is situated in the overall system. The higher levels are not sufficient, because often the details of the composition are necessary for making the right predictions or explanations.

The third thesis is related to the second one, but is stronger, since it states that higher-level explanations are necessary not only now, but also in the foreseeable future. The importance of higher-level explanations is not due to some temporary incompleteness of lower-level theories. For example, even when we know the full story of memory consolidation all the way down to the molecular level, we will still need higher-level regularities characterizing the functioning of memory, since going down to the molecular level to seek explanations is in most cases both pointless and intractable due to the enormous complexity of the system (see, e.g., Dennett 1991 or Wimsatt 2007 for more).

The point of the fourth thesis is to emphasize the importance of explanatory integration and interlevel connections: the explanations of different fields and levels are not independent or isolated from each other. This is a crucial point that sets explanatory pluralism apart from more radical forms of pluralism and claims of disunity of science (e.g., Cartwright 1999).

Is explanatory pluralism compatible with reductionism? Of course, this depends on what is meant by reductionism. If we understand reductive explanation as downward-looking mechanistic explanation (Bechtel 2008), and reductionism as the view that all mental phenomena can be reductively explained, then explanatory pluralism and reductionism are indeed compatible. The claim that all mental phenomena can be reductively explained in the mechanistic sense does not contradict any of the four theses of explanatory pluralism. In fact, the wide acceptance of explanatory pluralism is closely related to the recent emergence of mechanistic explanation as the paradigm for the

philosophy of the life sciences (Bechtel 2008; Behtel & Richardson 1993; Craver 2007; Machamer et al. 2000).

If, on the other hand, reductionism is understood as "New Wave Reductionism" (Bickle 1998), "ruthless" reductionism (Bickle 2003), or "functional reductionism" (Kim 1998, 2005), then reductionism is not compatible with explanatory pluralism. Reductionists of these kinds would deny one or all of the first three theses of explanatory pluralism (see Walter & Eronen 2011 for more).

Explanatory pluralism is certainly not without its contenders, and the relevant criticisms need to be addressed. However, in this paper I will treat explanatory pluralism as a premise, along with the interventionist account of causation (next section). My aim here is to explore the implications of these views, not to defend them.

3. The interventionist account of causation

The approach to causation that most naturally fits explanatory pluralism is the "interventionist" account (Pearl 2000, Woodward 2003, 2008; Woodward & Hitchcock 2003, also Spirtes, Glymour, and Scheines 1993). Indeed, many explanatory pluralists have explicitly endorsed it as the right understanding of causation (e.g., Craver 2007). I will focus here on Woodward's (2003) version of interventionism, which is exceptional in its scope and clarity.

The guiding insight of the account is that causal relationships are relationships that are potentially exploitable for purposes of manipulation and control. To put it very roughly, in this model a necessary and sufficient condition for *X* to cause *Y* or to figure in a causal explanation of *Y* is that the value of *Y* would change under some intervention on *X* (in some background circumstances).

An intervention can be thought of as an (ideal or hypothetical) experimental manipulation carried out on some variable *X* (the independent variable) for the purpose of ascertaining whether changes in *X* are causally related to changes in some other variable *Y* (the dependent variable). Of course, several restrictions on interventions must be added – see Woodward (2003) for details. Interventions are not only human activities, there are

also ”natural” interventions, and the definition of an intervention makes no essential reference to human agency. This sets the interventionist account clearly apart from previous manipulability theories of causation (e.g., Menzies and Price 1993).

This framework captures the nature of causation as *difference-making*: if variable *X* is causally relevant for variable *Y*, changes in the value of variable *X* make a difference in the value of variable *Y* (in a range of circumstances). One consequence of this model is that relata of causation must be represented as variables, but states or properties can easily be represented as binary variables, such that, for example, 1 marks the presence of the property and 0 the absence of the property.

According to Woodward, causal relationships are relationships that are *invariant* under interventions – that is, they continue to hold under a range of interventions. Physical laws are (highly) invariant generalizations, but so are many biological, psychological, economical, etc., generalizations. This leads to a kind of causal pluralism: we have causal generalizations at several levels and in several different domains, and they need not be reduced to some physical processes. This directly supports explanatory pluralism.

The interventionist account also allows for several noncompeting representations of one and the same system. What variables we choose to include in the representation depends on the question at hand. However, this does not make causal judgments subjective, since the counterfactual patterns of dependence that make the causal claims true or false are mind-independent. Once the variables and representations are fixed, causal claims are true or false in a mind-independent way.

One problem in applying the interventionist account to the issues in philosophy of mind is that it seems to provide a rather weak, promiscuous, and most importantly nonreductive notion of causation that many philosophers of mind will find unsatisfactory (e.g., Kim 2005). These philosophers are after a productive or generative notion of causation that is more metaphysically robust and somehow grounded in fundamental physics. They argue that when we discuss issues like mental causation we should be interested in causation in such a stronger sense.

However, the problem with grounding causation in physics is that notions like cause and effect do not really play a role in our best physical theories (as famously

argued by Bertrand Russell (1912-13), and more recently by Ladyman and Ross (2007), Loewer (2007), Norton (2007), and many others). The fundamental laws of physics relate the totality of a physical state at one time to the totality of the physical state at later instants, but do not single out causes and effects among these states. If we want to find causes that physically "bring about" or "produce" their effects, or causes that are "sufficient" for their effects, we have to consider something like the entire state of the universe as the cause for even a small effect.

Of course, we can put labels onto relata that appear in physical equations and call some of them causes and others effects, but this is entirely superfluous to the physics itself. Causal notions do not in any way guide or restrict physical theory formation. Furthermore, there are cases even in Newtonian physics that go straight against our ideas of causation (Norton 2007), not to even speak of phenomena like quantum entanglement.

The interventionist account seems to capture the nature of causation both in special sciences and everyday life very well, and in fundamental physics, causal notions are unnecessary and superfluous. It then seems that the interventionist account, insofar as it is successful, gives us all we want from an account of causation. In the rest of the paper, I will assume that this is indeed the case.

4. From Explanatory Pluralism to Pluralistic Physicalism

What is the relation between explanatory pluralism and physicalism? What are the ontological implications of the interventionist account of causation? These questions have been largely neglected in the literature on explanatory pluralism and interventionism, mainly due to the tendency of philosophers working on these topics to eschew traditional metaphysical issues. However, instead of eschewing the metaphysics, one can also try to find out a scientifically relevant metaphysical position that fits explanatory pluralism and interventionism. This is my main goal in this section. Among other reasons, this is important for attempts to connect the new philosophy of science with the more classic metaphysical debates in philosophy, particularly philosophy of mind.

Traditionally, causal considerations have played a key role in the arguments for physicalism. For example, Kim (2005) argues along the following lines: Causal considerations rule out substance dualism, since it is inconceivable how the nonmaterial mental substance could causally interact with the physical substance that has only physical properties. Kim then continues by arguing that causal considerations also rule out property dualism: the famous causal exclusion argument purportedly shows that nonphysical properties cannot have causal powers of their own, which means that property dualism leads to the highly implausible conclusion that nonphysical properties are epiphenomenal.

However, if we adopt the interventionist account, this reasoning breaks down. In the interventionist framework, causation is a notion that is important in the special sciences but not in fundamental physics. Causes at different levels can happily coexist, and higher-level causes are not excluded by lower-level causes (more on this in the next section). This is in stark contrast with the view that causation is a physical matter or that all causes reduce to physical causes. It seems that causal considerations now lead toward some kind of pluralism instead of traditional physicalism.

Let us then take a closer look at the kind of pluralism I have in mind. I propose we should start by taking *robustness* as the criterion for what is real. The idea of robustness is drawn from the practice of scientific modeling, and has been most extensively discussed by William Wimsatt (2007). He roughly defines it as follows (2007, 196): "*Things are robust if they are accessible (detectable, measureable, derivable, defineable, producible, or the like) in a variety of independent ways*." For instance, the moon is a very robust thing, since it can be measured and detected and accessed in numerous ways that are independent from each other. Properties like temperature or mass are robust, since they are also measurable, detectable, etc., in a variety of independent ways. It is important that the different ways of access are *independent* from each other, since then the likelihood that they all are mistaken is a product of each one's independent likelihood to go wrong, and this product will be a very small number if there are many independent ways.

According to Wimsatt (1981; 2007), robustness is by no means a new idea, and has in fact been looming at the background throughout the history of philosophy,

particularly in the works of Aristotle, Galileo, Peirce, and Whewell. In the last century, the idea was discussed by Levins (1966) in connection to modeling in population biology, and Levins was apparently the first to use the term "robust" in approximately the present sense (see also Hacking (1983), who does not use the term but presents similar ideas in passing). However, in spite of its importance, robustness has never received broader attention of the philosophical community – only very recently there has been renewed interest in the idea (Calcott 2010, Weisberg 2006).

Wimsatt extends robustness to cover also theories, laws, explanations, and so on, but this makes the notion unnecessarily complicated. For the present purposes, we can define a version of robustness that concerns only properties: a property is robust if it is detectable, measurable or producible in a variety of independent ways. Based on this, we can formulate the core idea of robustness-realism as follows: *We are justified in believing that property P is real if and only if property P is robust, that is, it is detectable, measurable or producible in a variety of independent ways.*

This formulation may be in need of further refinement, but the basic idea is clear and plausible. It is also clear that if we take robustness as a guideline for building our ontology, plenty of higher-level or special science properties turn out real. For example, the properties of short-term memory, such as its approximate capacity, can be measured and studied with varying experimental setups that are independent of each other. Change blindness is a fairly recently discovered robust property of the visual system that is detectable and producable in a variety of independent ways. The same goes for psychological and special science properties in general, insofar as they are good scientific properties – as Wimsatt (2007, Ch. 4) points out, scientists generally use robustness analyses to determine whether a phenomenon is real or just an artifact. Using robustness as a guideline for what to consider real leads to a kind of ontological pluralism and a "tropical rainforest ontology" (Wimsatt 2007).

One should not understand ontological pluralism based on robustness as some kind of "spooky" pluralism that asserts that there are fundamentally different substances in the world. It merely expresses the fact that there are many different kinds of properties in the world, and that requiring that everything real is reducible to something physical or has physical causal powers does not make much sense.

Another important caveat is that I am not advocating a form of constructivism. The pluralism I am defending is rather a form of scientific realism. Our ideas about what is robust may change as science proceeds, but this does not mean that reality itself changes. The fact that property *P* is robust in our current analyses gives us justification for believing that *P* is real, but it does not in any sense "make" *P* real.

Let us now turn to the question whether robustness pluralism is an alternative to physicalism or a kind of physicalism. In addition to causal arguments that were discussed above, another motivation for physicalism has come from considerations based on the history of science. All hypotheses concerning non-physical forces that affect physical processes in a way that conflict with the laws of physics have consistently failed. Relatedly, as science has progressed, more and more phenomena have been successfully explained in broadly speaking physical terms – also phenomena that were previously thought to resist physical explanations. Perhaps the biggest triumph in this respect was the explanation of the fundamental processes of life in terms of DNA molecules. However, these inductive arguments do not directly support physicalism. They support a weaker thesis, which Ladyman and Ross (2007, 43) have dubbed the *Primacy of Physics Constraint* (PPC): "Special science hypotheses that conflict with fundamental physics, or such consensus as there is in fundamental physics, should be rejected for that reason alone. Fundamental physical hypotheses are not symmetrically hostage to the conclusions of the special sciences." That is, physics sets *constraints* for the theories of special sciences.

A robustness pluralist can happily accept the Primacy of Physics Constraint. The claim that there are irreducible higher-level properties in no way conflicts with the claim that fundamental physics constrains the theories or hypotheses of special sciences. This takes us to the point that instead of seeing robustness pluralism as an alternative to physicalism, it is perhaps more appropriate to see it as a *kind of physicalism.* Consider the following definition of physicalism (often called "supervenience physicalism"): Physicalism is true at a possible world *w* if and only if any world which is a (minimal) physical duplicate of *w* is a duplicate of *w* simpliciter (Jackson 1998, 12). Nothing what has been said above is in conflict with this. A robustness pluralist could also accept that the fundamental physical level in some sense determines all the higher-level properties. A

robustness pluralist could accept token physicalism. If criteria of this kind are sufficient for physicalism, and I believe they are, then the position I have defended could be called *pluralistic physicalism*. It provides a scientifically credible and philosophically interesting middle ground between reductive physicalism and more radical forms of pluralism.

What is then wrong with classical forms of physicalism and why is robustness pluralism preferable to them? I take the main problem with reductive physicalism (type physicalism) to be the familiar one. Putnam (1967) was the first to argue that it is extremely ambitious to assume that a psychological property could be identified with a single brain state, since psychological properties can have multiple different realizers. Recent analyses (e.g., Bechtel & Mundale 1999, Polger 2004, Shapiro 2000) have cast doubt on this idea of multiple realizability, both conceptually and empirically, and I agree with these critics in that philosophers of mind have overestimated the significance of multiple realizability. However, I also believe that proponents of multiple realizability are right in one sense: there are no one-to-one mappings from all higher-level properties to physical properties. The type physicalist solution to the reality of higher-level properties would require the following: for every single higher-level property that we want to retain in our ontology we will find a physical property that is identical to that higher-level property. I find this extremely implausible. Furthermore, if we look at scientific practice, special science properties are not considered real only insofar as they are identical to some physical properties – they are considered real insofar as they are robust.

What is the relation between pluralistic physicalism and traditional non-reductive physicalism? If nonreductive physicalism is understood as consisting of a moderate kind of physicalism (such as supervenience physicalism) and the view that special science properties are distinct from physical properties, then pluralistic physicalism is a form of nonreductive physicalism. However, traditional nonreductive physicalism carries more baggage than this. Most importantly, it also includes the following thesis about the ontological status of higher-level properties: higher-level properties are not identical to physical properties, but are physically *realized*. The problem with this "realization physicalism" is that its success hinges on the notion of realization, but it has turned out to be extremely difficult to spell out a notion of realization that would yield a plausible form

of nonreductive physicalism and make scientific sense (Polger 2004, 2007, Shapiro 2004, see also Eronen 2010-2011). Without such an account, realization physicalism collapses into either type physicalism or property dualism. In contrast, pluralistic physicalism abandons the idea of realization. It states that higher-level properties are real insofar as they are robust; they need not be "realized" (i.e., made real) by physical properties.

Generally speaking, higher-level (or mental) properties are very heterogeneous, and their relations to physical properties are complex and have to be analyzed case by case. These relations can spelled out in terms of constitution, mechanisms, determination, satisfaction of function, and so on, but there is no reason to expect a single notion, such as "realization", to apply in every case, and it is questionable whether "realization" even captures anything important that could not be accounted for with the other notions.

5. Pluralistic Physicalism and Causal Exclusion Worries

Perhaps the most formidable challenge to nonreductive ontological positions, including pluralistic physicalism, is the causal exclusion argument. Several different versions of the argument exist; the formulation here reflects the account of Jaegwon Kim (Kim, 2002, 2005), who has been the most ardent proponent of the exclusion argument. The argument is based on certain principles that together create a problem for mental causation (Kim, 2002, 278):

> *The Problem of Mental Causation*: Causal efficacy of mental properties is inconsistent with the joint acceptance of the following four claims: (1) physical causal closure, (2) exclusion, (3) mind-body supervenience, and (4) mental/physical property dualism (i.e., irreducibility of mental properties).

The principle of physical causal closure states that every physical occurrence has a sufficient physical cause. The principle of exclusion states that no effect has more than one sufficient cause, except in cases of genuine overdetermination, such as two bullets hitting the heart of a victim at exactly the same time, both causing death.

It is easy to see how the four principles taken together lead to trouble. Let us start by assuming that (the instantiation of) a mental property *M* causes (the instantiation of) another mental property *M**. Due to mind-body supervenience, *M* supervenes on some physical property *P*, and *M** supervenes on some physical property *P**. Since *M** supervenes on *P**, *M** must be necessarily instantiated whenever *P** is instantiated, no matter what happened before: the instantiation of *P** alone necessitates the occurrence of *M**. Thus, according to Kim, the only way that *M* can cause *M** is by causing *P**.

This is where the principle of causal closure kicks in: *P** must also have a sufficient physical cause. This means that *P** has a sufficient physical cause *P* and a mental cause *M*, and the exclusion principle states that one of these must go – if we would accept cases like this as genuine overdetermination, we would get massive overdetermination of physical effects by mental causes, which is highly implausible. Obviously *M* is the one that has to go, since if *M* was the only cause of *P**, this would violate the principle of physical causal closure. Therefore, *M* cannot be the cause of *M** or of any other mental or physical property. This holds for all mental properties, and we have the striking conclusion that, under mind-body supervenience, mental properties are causally impotent.

According to Kim, physical causal closure and mind-body supervenience are among the inescapable commitments of all physicalists. The exclusion principle is taken to be a general metaphysical constraint that can hardly be challenged. This leaves only mental/physical property dualism (i.e., the irreducibility of mental properties) as the principle that has to go. Therefore, Kim's conclusion is what he calls "conditional reductionism": "If mentality is to have a causal influence in the physical domain – in fact, if it is to have any causal efficacy at all – it must be physically reducible" (Kim, 2005, 161).

The argument is targeting mental properties, and I will mainly discuss mental causes in this section, but it should be noted that the argument works just as well for any nonphysical properties. One reason why mental properties are seen as particularly problematic is that it is generally assumed that biological, neural, chemical, etc., properties either count as broadly speaking "physical" properties, or are ontologically reducible to physical properties. Therefore, premise (4) does not hold for these properties,

and they are not threatened by the argument. Yet, the pluralism I have defended above can be taken to imply that these kinds of properties are in a sense distinct from physical properties, and therefore face the exclusion argument. For this reason, it is particularly important to show that there are no serious worries of causal exclusion.

Prima facie, it seems that mental causation is unproblematic in the interventionist framework. There are invariant psychological generalizations such that we can make interventions to mental states in order to change other mental states or physical behavior. For example, as Woodward (2008) points out, when you persuade someone, you manipulate her beliefs by providing information or material things, in order to change her other beliefs. Also many psychological and social science experiments involve intervening on the beliefs of the subjects, usually through verbal instruction, in order to change some other beliefs and observable behavior.

In a closer philosophical analysis, it indeed seems that the interventionist account vindicates mental causation. Recently several authors (e.g., Menzies 2008, Raatikainen 2010, Woodward 2008) have argued that if the interventionist account is correct, mental states can be causes of physical behavior, and they are not excluded by their physical realizers. On the other hand, Michael Baumgartner (2010) has argued that there is an interventionist version of the exclusion argument, and thus adopting the interventionist account does not make the problem of exclusion go away.

Instead of going through the details of these arguments, I argue that there is a deeper underlying problem that kicks in already before the arguments of either side can take off. The problem is that typical causal representations of mental causation fail to satisfy the conditions required of interventionist causal models. One of these conditions is that variables that are not related as cause or effect or as effects of a common cause have to be uncorrelated. In other words: conditional on its direct causes, each variable has to be independent of every other variable except its effects (this is often called the Causal Markov Condition, see Hausman & Woodward 1999 for other formulations and an extensive discussion of the condition). Although the exact formulation of this condition has been a matter of some debate, it is widely agreed that the condition (or at least something very close to it) is integral to causal modeling. If this condition is not

satisfied, the model is not a well-formed causal model, and drawing causal inferences from it is not possible.

The typical representations of mental causation in philosophy of mind *fail* to satisfy this condition. Kim's formulation of the exclusion argument is a good example: in this representation, mental property M causes another mental property M^*, physical property P causes another physical property P^*, M supervenes on P, and M^* supervenes on P^*. Due to supervenience, the values of M and P (as well as M^* and P^*) are correlated, and M depends on P. Whenever M changes, P also changes, and when the value of P is fixed, the value of M is also fixed. However, M does not cause P, P does not cause M, and they are not both effects of a common cause. Mind-body supervenience implies a non-causal correlation and dependency between the variable describing the mental property and the variable describing the physical property. Therefore, from an interventionist point of view, the representation is incorrect and has to be modified.

The obvious reductive solution to this problem would be to get rid of the mental variables, either by eliminating them or identifying them with physical variables. Then we would have only physical variables in the representation, and no non-causal relationships. However, the problem with this approach becomes obvious when we consider the fact that we can apply just the same reasoning to biological, chemical, neural, and macrophysical properties. They all supervene on lower-level physical properties. Therefore, we can simply draw the same picture again, replacing mental variables by, say, neural variables. Then it seems that since we got rid of the mental variables in the first case, we also have to get rid of the neural variables in the second case. Causation seems to be draining away towards some fundamental physical level, which is particularly strange if we consider the fact that there seems to be nothing resembling our ideas of causation at the fundamental physical level (see section 3). (This is a version of the *generalization argument* that has often been raised against Kim's exclusion argument (e.g., Block 2003, van Gulick 1992).)

The reductive approach of replacing or reductively identifying the higher-level variables also runs counter to scientific practice: when scientists have to choose between causal representations of a system, it is not the case that they always choose the maximally precise or lowest-level representation. The interests of the scientist determine

the explanandum, and once this is fixed, various empirical and theoretical considerations determine the right level at which the causal explanation is sought (Woodward 2010). One does not get rid of a good causal model just because the properties represented in it supervene on some lower-level properties.

This leads to a more scientifically plausible way of dealing with supervenience in causal representations. This would allow higher-level causal representations, but not allow including the supervenient base variables in the same representation. For example, we would not include neural variables in the same representation as the supervenient mental variables. We would have a plurality of causal representations, but no representations that include both supervenient variables and their base variables. As Hausman and Woodward (1999, 531) put it in a different context: "One needs the right variables or the right level of analysis – variables that are sufficiently informative and that are not conceptually connected."

This approach is simple, coherent, and scientifically credible. However, defending it convincingly also requires showing what exactly goes wrong in the exclusion argument. The argument seems to be valid, so at least one of its premises has to turn out false.

Let us start with the most likely candidate, the exclusion principle. This principle states that no effect has more than one sufficient cause, except in cases of genuine overdetermination. A straightforward interventionist rendering of this principle would be something along these lines: If variable *M* is a difference-making cause for *B*, there is no other difference-making cause for *B*, unless this is a genuine case of overdetermination. It is easy to see that this principle does not hold: there can be many difference-making causes to a single variable. However, this formulation is too general and not very fair – it should at least include the requirement that the competing causes are acting at the same instance in time (Menzies 2008). Taking this into account, we could formulate the principle as follows: If this particular instantiation of *M* (the variable *M* taking, say, value 1 instead of 0) is a difference-making cause for this particular instantiation of *B* (the variable B taking value 1 instead of 0), then there is no other difference-making cause for this particular instantiation of *B* (unless this is a case of overdetermination).

In my view, this principle is also problematic. Due to supervenience, there seems to be another difference-making cause for the particular instantiation of *B* in addition to the instantiation of *M*, namely an instantiation of the supervenience base of *M*. As I have argued above, we should not include this in the same representation as *M*, and in most cases considering the supervenience base is likely to be intractable or pointless, but we cannot rule out the possibility that at least sometimes it is possible (and makes sense) to build a representation where the supervenience base of *M* (but not *M* itself) is included and where it is the cause of *B*.

One can see this either as a denial of the exclusion principle or as systematic overdetermination. If one does not count cases like above as genuine overdetermination, then the exclusion principle is false. If one does count them as overdetermination, then we have systematic overdetermination. Both options have been traditionally considered unacceptable, but if we understand causation as a matter of manipulation and control (and not as physical "bringing about"), this kind of violation of the exclusion principle or acceptance of overdetermination is unproblematic (see also Bennett (2003), who casts doubt on the exclusion principle, independently of the notion of causation applied). There simply can be several difference-making causes at different levels for a given effect, and which level we focus on depends on the context and the question at hand.

It is also important to note that the position I have defended does not lead to any scientifically or physically dubious conclusions, such as non-physical causes that violate physical laws. I am not denying the principle of physical causal closure. Although there are also nonphysical difference-making causes for physical occurrences, every physical occurrence does have a physical difference-making cause.

To summarize, if we understand causation in interventionist terms, it is true that causal claims become very problematic when conjoined with supervenience claims. However, this does not mean that higher-level causes are excluded by the lower-level causes they supervene on. Which variables are retained in the representation depends on the question at hand. The exclusion argument can be tackled either by denying the exclusion principle or by accepting systematic overdetermination. Therefore, in the interventionist framework, the exclusion argument does not rule out higher-level causes.

6. Conclusion

Explanatory pluralism and the interventionist account of causation together form a coherent and scientifically plausible framework for the philosophy of the cognitive sciences. The ontological position most naturally fitting this framework is what I have called pluralistic physicalism. In spite of first appearances, this kind of pluralism is not undermined by the causal exclusion argument.

In a broader perspective, I hope to have shown that recent developments in philosophy of science have extremely important implications for traditional issues in philosophy of mind. Philosophers of mind should pay closer attention to the contemporary debates in philosophy of science, and both sides would benefit if these two subdisciplines came to interact more closely in the future.

Acknowledgements

I thank Vera Hoffmann-Kolss, Dan Brooks and Laura Bringmann for very helpful discussions and feedback on earlier drafts of this paper. I also thank the three anonymous referees of this journal, whose comments helped significantly improve the article. Finally, I am grateful to the Finnish Cultural Foundation for supporting this work financially.

References

Baumgartner, M. (2010). Interventionism and Epiphenomenalism. *Canadian Journal of Philosophy* 40, 359-383.

Bechtel, W. (2008). *Mental Mechanisms. Philosophical Perspectives on Cognitive Neuroscience*. London: Routledge.

Bechtel, W., and Mundale, J. (1999). Multiple realizability revisited. *Philosophy of Science* 66, 175–207.

Bechtel, W., and Richardson, R. C. (1993). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Princeton: Princeton University Press.

Bennett, K. (2003). Why the Exclusion Problem Seems Intractable, and How, Just Maybe, to Tract It. *Noûs* 37, 471-497.

Bickle, J. (1998). *Psychoneural Reduction: The New Wave*. Cambridge, MA: MIT Press.

Bickle, J. (2003). *Philosophy and Neuroscience: A Ruthlessly Reductive Account.* Dordrecht: Kluwer Academic Publishers.

Block, N. (2003). Do Causal Powers Drain Away? *Philosophy and Phenomenological Research* 67, 133-150.

Brigant, I. (2010). Beyond reduction and pluralism. *Erkenntnis* 73, 295–311.

Calcott, B. (2010). Wimsatt and the Robustness Family: Review of Wimsatt's Re-engineering Philosophy for Limited Beings. *Biology & Philosophy* 26, 281-293.

Cartwright, N. (1999). *The Dappled World: A Study of the Boundaries of Science*. Cambridge: Cambridge University Press.

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

Dennett, D. C. (1991). Real Patterns. *The Journal of Philosophy* 88, 27-51.

Eronen, M. I. (2010-2011). Replacing Functional Reduction with Mechanistic Explanation. *Philosophia Naturalis* 47-48, 125-153.

Hacking, I. (1983). *Representing and intervening. Introductory topics in the philosophy of natural science*. New York: Cambridge University Press.

Hausman, D. M. and Woodward, J. (1999). Independence, Invariance and the Causal Markov Condition. *British Journal for the Philosophy of Science* 50, 521-583.

Hempel, C. G., and Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science* 15, 135-175.

Jackson, F. (1998). *From Metaphysics to Ethics: A Defense of Conceptual Analysis.* Oxford: Oxford University Press.

Kim, J. (1998). *Mind in a Physical World*. Cambridge, MA: MIT Press.

Kim, J. (2002). Mental causation and consciousness: the two mind-body problems for the physicalist. In C. Gillett & B. Loewer (eds.) *Physicalism and Its Discontents*. Cambridge: Cambridge University Press, 271-283.

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

Ladyman, J. and Ross, D. (2007). *Every Thing Must Go: Metaphysics Naturalized*. Oxford: Oxford University Press.

Levins, R. (1966). The strategy of model building in population biology. *American Scientist* 54, 421-431.

Loewer, B. (2007). Mental Causation, or Something Near Enough. In B. P. McLaughlin and J. Cohen (eds.) *Contemporary Debates in Philosophy of Mind*. Malden, MA: Blackwell Publishing, 243-264.

Looren de Jong, H. (2002). Levels of explanation in biological psychology. *Philosophical Psychology* 15, 441-462.

Machamer, Peter K., Lindley Darden, and Carl Craver (2000). Thinking about mechanisms. *Philosophy of Science* 67, 1-25.

McCauley, R. N. and Bechtel, W. (2001). Explanatory Pluralism and Heuristic Identity Theory. *Theory & Psychology* 11, 736-760.

Menzies, P. (2008). The exclusion problem, the determination relation, and contrastive causation. In Hohwy, J. & Kallestrup, J. (Eds.) *Being Reduced*. Oxford: Oxford University Press, 196–217.

Menzies, Peter, and Huw Price (1993). Causation as a secondary quality. *The British Journal for the Philosophy of Science* 44, 187-203.

Mitchell, S. D. (2003). *Biological Complexity and Integrative Pluralism*. Cambridge: Cambridge University Press.

Norton, J. D. (2007). Causation as Folk Science. In H. Price & R. Corry (Eds.) *Causation, Physics, and the Constitution of Reality. Russell's Republic Revisited*. Oxford: Oxford University Press, 11-44.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge, UK: Cambridge university press.

Polger, T. W. (2004). *Natural Minds.* Cambridge: MIT Press.

Polger, T. W. (2007). Realization and the Metaphysics of Mind. *Australasian Journal of Philosophy* 85, 233-259.

Putnam, H. (1967). Psychological predicates. In W.H. Capitan & D.D. Merrill (eds.) *Art, Mind, and Religion*. Pittsburg: Pittsburg University Press, 37–48.

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

Russell, B. (1912-1913). On the Notion of Cause. *Proceedings of the Aristotelian Society* 13, 1-26.

Shapiro, Lawrence A. (2000). Multiple realizations. *The Journal of Philosophy* 97, 635-654.

Shapiro, Lawrence A. (2004). *The Mind Incarnate*. Cambridge, MA: MIT Press.

Spirtes, P., Glymour, C., and Scheines R. (1993). *Causation, Prediction, and Search*. New York: Springer-Verlag.

van Gulick, R. (1992). Three bad arguments for intentional property epiphenomenalism. *Erkenntnis* 36, 311-332.

Walter, S. and Eronen, M. I. (2011). Reductionism, Multiple Realizability, and Levels of Reality. In S. French & J. Saatsi (eds.) *Continuum Companion to the Philosophy of Science*. London: Continuum, 138-156.

Weisberg, Michael (2006). Robustness analysis. *Philosophy of Science* 73, 730-742.

Wimsatt, W. C. (1976). Reductionism, Levels of Organization, and the Mind-Body Problem. In Globus et al. (eds.) *Consciousness and the Brain. A Scientific and Philosophical Inquiry*. New York: Plenum Press, 205-267.

Wimsatt, W. C. (1981). Robustness, Reliability, and Overdetermination. In M. Brewer and B. Collins (eds.) *Scientific Inquiry and the Social Sciences*. San Fransisco: Jossey-Bass, 124-163. Revised reprint in Wimsatt (2007), 43-74.

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

Woodward, J. (2003). *Making Things Happen*. Oxford: Oxford University Press.

Woodward, J. (2008). Mental causation and neural mechanisms. In Hohwy, J. & Kallestrup, J. (Eds.) *Being Reduced*. Oxford: Oxford University Press, 218–262.

Woodward, J. (2010). Causation in biology: stability, specificity, and the choice of levels of explanation. *Biology & Philosophy* 25, 287-318.

Woodward, J. and Hitchcock, C. (2003). Explanatory Generalizations, Part I: A Counterfactual Account. *Noûs* 37, 1-24.