CHAPTER X

REALIZATION AND ROBUSTNESS: NATURALIZING NONREDUCTIVE PHYSICALISM

MARKUS I. ERONEN

Introduction

In philosophy of mind, one of the standard positions regarding the nature of mental properties is *nonreductive physicalism*: mental properties are physical or physically realized, but nevertheless irreducible. The ontological framework for traditional nonreductive physicalism crucially relies on the notion of *realization.* The idea is that mental properties are ontologically distinct from physical properties, but at the same time causally efficacious and naturalistically acceptable due to the fact that they are physically realized.

In this article, I argue that this ”realization physicalism” fails to deliver what it promises: it does not secure the ontological autonomy of mental properties, and does not provide an answer to the causal exclusion problem. I also present an alternative account that shows how mental properties can be ontologically distinct from physical properties while being also causally relevant. I apply the notion of *robustness* to develop a novel understanding of the ontological status of mental properties, and draw from the *interventionist* account of causation to dissolve worries of causal exclusion. This yields a more naturalistic and scientifically credible alternative to realization physicalism.

Realization and non-reductive physicalism

Realization physicalism[1] is closely connected to functionalism. One of the core ideas of functionalism[2] is that a given mental property is not identical to any physical property, since each functionally defined mental property can be ”realized” by several different physical properties that perform this function (Putnam 1967). This implies that mental states are ontologically irreducible – they are not just physical states under different descriptions.

However, the view that mental properties are distinct from physical properties immediately evokes the dreaded causal exclusion argument, which has been most extensively defended and developed by Jaegwon Kim (1993, 1998, 2005). The exclusion argument states that mental properties cannot have causal powers of their own: since all physical occurrences have sufficient physical causes, any additional mental causes would either overdetermine physical effects or violate physical laws, and both are unacceptable results (I will discuss the argument in more detail in the section “Facing the causal exclusion argument”).

Realization physicalism is supposed to yield a form of nonreductive physicalism that avoids this problem. The solution of realization physicalism is that mental properties are not entirely distinct from physical properties, since they are *physically realized*. In virtue of being physically realized, they can also be causally efficacious without violating physical laws.

It is obvious that all the metaphysical work in this account is done by the notion of realization. This notion, if it is to make realization physicalism a viable solution, has to fulfill two requirements: it must allow physically realized mental properties to be in a substantial sense ontologically distinct from physical properties, and it must allow them to be causally efficacious.

I will mainly focus here on the most influential and common strategy of making sense of realization, the so-called *subset strategy*. This strategy is explicitly supported in somewhat different forms at least by Clapp (2001), Shoemaker (2001) and Pereboom (2002). In addition to these explicit proponents, the model is often implicitly assumed and applied in many of the central debates in current philosophy, including those concerning reduction, emergence, and physicalism (see Gillett (2010) for more).

---

1 I have adopted the term ”realization physicalism” from Polger (2007).

2 Functionalism is to be understood here as role functionalism, not as the reductive filler-functionalism defended by Lewis (1972).

A central background assumption of the subset model is the causal theory of properties (which goes back to Shoemaker 1980). According to this theory, a property is individuated by the causal powers it contributes to an individual. For example, the property of being metallic contributes to an individual powers such as the power to conduct electricity and heat, and the total set of causal powers that the property of being metallic contributes is what makes it the property it is.

The main idea of the subset model of realization is expressed very succinctly by Clapp (2001, 129): "*P realizes Q* if and only if (def.), where *p* and *q* are the sets of powers constituting *P* and *Q*, $q \subset p$." In other words, property *P* realizes property *Q* if and only if all the causal powers of *Q* are also causal powers of *P*. That is, the causal powers of *Q* form a *subset* of the causal powers of *P*.[3]

Let us suppose that mental properties are physically realized in accordance to this model. This means that for each mental property *M* we can find a physical property *P* such that the causal powers of *M* are a subset of the causal powers of *P*. How does this guarantee the autonomy of mental properties and solve the causal exclusion problem? The idea is that mental properties and physical properties are distinct, but not *entirely* distinct. The physical realizer properties in a sense contain the mental properties as their parts. Therefore, as Clapp (2001) and Shoemaker (2001) argue, there is no more causal competition between mental and physical properties than there is between the whole and its parts. In some situations it is more natural to consider the part rather than the whole as causally efficacious.

Consider the famous example from Yablo (1992): A pigeon is trained to peck at all objects that are red, and is presented with an object that is scarlet. The property of being scarlet is a subset realizer of the property of being red. Is it the property of being red or the property of being scarlet that causes the pigeon to peck? It seems more natural to consider the property of being red as the cause, since it is what makes the causal difference – the pigeon would have still pecked if the object had not been scarlet (but some other shade of red). Appealing to this analogy, the

3 Strictly speaking, properties do not *have* causal powers and causal powers do not *constitute* properties. Properties contribute causal powers to individuals. However, it is common in the debate to simply talk of properties having causal powers. For the sake of simplicity, I will also continue to use expressions like this. The causal powers of a property can be taken to mean the causal powers the property contributes to an individual.

proponents of the subset model argue that there are situations where we naturally consider the mental property rather than its physical realizer as the causally efficacious property. Thus, the subset model supposedly shows how mental properties can be distinct from physical properties, yet causally efficacious.

Problems of realization

Realization physicalism is an elaborate position, but leads to several fundamental problems. These have been clearly stated by Gillett (2010), Polger (2004) and Walter (2010) (see also Shapiro 2004 for an extensive critical discussion of the notion of realization in general), and I will only briefly summarize them here.

First of all, it is not at all clear that the subset-realized mental properties are causally efficacious. According to the subset view of realization, the causal powers of a given mental property form a subset of the causal powers of the realizing physical property. This is supposed to guarantee the causal efficacy of mental properties without violating physical laws. However, what this means is that the mental property has no causal powers that go beyond those of the physical property. If realization physicalism is correct, then the physical properties alone are sufficient for bringing about all the effects; there is nothing the mental properties add or even could add to the causal nexus of the world. Mental properties are not causally efficacious *qua* mental properties, they are causally efficacious *qua* parts of physical properties.

Relatedly, it is questionable whether mental properties are on this view truly distinct from physical properties. Mental properties are parts of physical properties and have no causal powers that go beyond those of the physical properties. Mental properties make no further contribution to the causal relations in the world; physical properties do all the causal work. It seems that for the sake of ontological parsimony we could simply retain the realizer properties in our ontology and do without the realized mental properties.

In general, it seems that the subset view of realization leads to a reductive view of mental properties and their causal powers, which is in stark contrast to the nonreductive aspirations of its proponents. On this view, mental properties are not truly distinct from physical properties, and have no additional causal powers. Physical properties fully account for the causal relations in the world. This seems very much like ontological reduction! A realization physicalist might still claim that we need mental

properties (or concepts) for practical reasons and that they can figure in explanations, but even if this is true, realization physicalism fails to solve the metaphysical problem of the autonomy of mental properties and mental causation.

Furthermore, it can also be questioned whether mental properties are in fact realized in such a way as the subset view suggests. Supporters of realization physicalism have given few if any examples of the physical realization of actual psychological properties studied by the psychological sciences. As Gillett (2010) points out, the subset view does not seem to capture any scientifically relevant relation between properties. This suggests that even if the subset account delivered what it promises, it could be just a philosophical fiction with no connection or relevance to science.

Based on these considerations, it is fairly clear that the subset model fails to provide a solid basis for nonreductive physicalism.[4] However, there are also other prominent accounts of realization – perhaps they could be of help to the realization physicalist?

Gillett (2003, 2010) has defended an alternative model of realization that differs from the subset model in two crucial respects. First of all, the subset model is ”flat” while Gillett's model is ”dimensioned”. This means that in the subset model the realized and realizer properties need to be properties of the same individual, while in Gillett's model the realizers may also be properties of the individual's constituents. Secondly, in Gillett's model the causal powers of the realized property need not form a subset of the causal powers of the realizer property. Instead, the requirement is that the causal powers of the realized property are contributed to the individual *in virtue of* the causal powers contributed by the realizer properties.

As interesting as this account is, it is not clear whether it is an account of ”realization” in the same sense as the subset model. As Polger (2007) has pointed out, the dimensioned account of realization fails to accommodate some of the central examples of realization in the literature, such as the realization of abstract algorithms or computations. It seems to be more like an account of scientific composition, and provides no novel solution to the problem of causal exclusion and the autonomy of mental states (see also Polger & Shapiro 2008). Accordingly, Gillett (2010) has

4 I believe this brief discussion is enough to show how problematic the subset model is, but the troubles of the model do not end here. For example, one could also question the assumption that all mental properties can be exhaustively defined in terms of their causal role.

recently defended the dimensioned model as an account of the scientific ”making-up” relation and argued that it forms a basis for ontological reductionism.

The notion of realization also figures prominently in the work of Jaegwon Kim (1992, 1998, 2005), but like Gillett, Kim ends up defending ontological reductionism. Kim supports the following ”causal inheritance principle”: if a functional property *E* is instantiated on a given occasion in virtue of one of its realizers, *Q*, being instantiated, then the causal powers of this instance of *E* are identical with the causal powers of this instance of *Q*. Based on this principle and considerations of causal exclusion, Kim reaches the conclusion that mental properties can be causally efficacious only if they are identical to physical properties. The leaves no room for any robust notion of ”realization”, since mental properties are either identical to physical properties or eliminated from the ontology (see Eronen 2010-2011 for more).

To sum up, the subset model fails to provide a basis for nonreductive physicalism, and Gillett's and Kim's views of realization lead to ontological reductionism. It seems that the notion of realization is of little help in defending nonreductive physicalism.

Robust mental properties

As we have seen above, the main role for the notion of realization in philosophy of mind has been to form a solid ontological basis for nonreductive physicalism. In this section, I show that we do not need realization for this: we can better understand the autonomy of mental properties and their relations to physical properties *without* the notion of realization. I will introduce a scientifically credible form of nonreductive physicalism that does not appeal to physical realization at all.

I propose we should understand the reality of mental properties in terms of *robustness*. 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, measurable, derivable, definable, 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 present 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 very 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 (e.g., Calcott 2010; Kuorikoski et al. 2010; Weisberg 2006; Woodward 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-reality 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* (see Eronen 2011 and Eronen forthcoming for more).

If we accept robustness as a guideline for building our ontology, it is clear that plenty of mental 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 producible in a variety of independent ways. The same goes for mental or psychological properties in general, insofar as they are good scientific properties.

Robustness realism thus provides an answer to the ontological status of mental properties and shows how they can be real while also being distinct from physical properties. Robust mental properties are real in their own right – they need not be physically "realized", i.e. "made real", by physical properties.

However, another key issue to which realization physicalism also supposedly provides an answer is the relation between mental and physical properties. Realization physicalism states that physical properties realize mental properties. Robustness alone does not say anything about this. If we adopt robustness-realism instead of realization physicalism, how

should we then understand the relations between mental properties and physical properties?

In general, realization physicalists have been very ambitious in assuming that there is a single notion that relates all mental properties to physical properties, and that this notion is sufficient for giving a satisfying answer to the ontological status of mental properties.[5] The relations between mental and physical properties are complex and have to be understood in terms of many different notions. Mental properties are very heterogeneous: even if we restrict our focus to the properties studied by cognitive psychology, they include properties varying from change blindness to the capacity of short-term memory and to cognitive dissonance.[6] There is no reason to expect that we could use a single notion, such as realization, to account for the way in which these properties are related to lower-level properties. It is not even clear whether we need a notion of realization – does it add something substantial to the set of more well-defined relations, such as identity, composition, determination, supervenience, etc.? It appears that at least the currently available notions of realization fail to do this (see Polger 2010 for more).

In contrast to realization, one recently much-discussed approach that has been helpful in analyzing the relations between mental properties and physical properties is the *mechanistic explanation* paradigm (Bechtel 2008; Behtel & Richardson 1993; Craver 2007; Machamer et al. 2000). Many (if not all) mental properties can be seen as resulting from the functioning of lower-level neural mechanisms, or to put it in another way, many mental properties can be seen as higher-level properties of multilevel mechanisms. They are higher-level properties in the sense that they are properties of the mechanism as a whole, not any of its components.

To give some very rough and simplified examples, the functioning of the neural mechanism centered round the cellular process of Long Term Potentiation (LTP) in the hippocampus results in spatial memory formation. Properties such as the capacity of short term spatial memory or the rate of deterioration of the memory trace are properties of the memory system as a whole. The molecular processes in photoreceptor cells result in

5 See also Polger 2004, ch. 4, 2010 and Polger & Shapiro 2008, who emphasize the heterogeneity of putative cases of realization.

6 In fact, these are better characterized as functions and capacities and effects rather than properties (see Cummins 2000); I only talk of mental properties because this is the common practice in philosophy of mind. Nothing crucial turns on this.

the visual system adapting to the ambient light, i.e., in the higher-level property of light adaptation. The ability to distinguish contrasts in extremely varying illumination conditions is a property of the system as a whole. These robust properties are not identical to any properties of the components of the system.

One might also call cases where a neural mechanism is performing a higher-level function ”realization”. As Wilson & Craver (2007) point out, this comes close to how the term ”realization” is used in the cognitive sciences: when scientists state that they are looking for, say, the neural realization of memory consolidation, what they typically mean is that they are looking for the neural mechanism of memory consolidation. However, it is obvious that this kind of weaker notion of realization differs fundamentally from the philosophical notions of realizations, such as the one applied in the subset model. One interesting question that is not yet resolved is what is the ontological relation between the overall (higher-level) function and the mechanism that performs it, but the existing philosophical accounts of realization are of little help in answering this (see also Polger 2010 for more).

Another open question is whether there is multiple realization in the sense that there are (or could be) several distinct mechanisms for a given function. Perhaps any relevant differences in the mechanism also necessarily result in relevant differences in the ”realized” higher-level function (cf. Shapiro 2000). However, if we accept the approach in this paper, the issue of multiple realizability vs. type physicalism becomes rather peripheral or irrelevant. I have argued that mental properties are real in virtue of being robust, not in virtue of being identical to physical properties or physically realized. In the next section, I will argue that they need not be identical to physical properties in order to be genuine causes. This suggests that the question of type physicalism becomes far less pressing than has been traditionally thought. In other words, I am not arguing that type physicalism is false, but I do argue that its truth is not necessary for guaranteeing mental causation or the reality of mental properties.

Furthermore, it is important to emphasize that the position I have defended is not a radical departure from physicalism, or a radical form of nonreductionism (such as emergent property dualism). I accept that (many, if not all) mental properties are potentially reductively explainable in the sense of being mechanistically explainable. I find it plausible that (many) mental properties result from the organized functioning of underlying neural mechanisms, and that they are in some sense determined by neural properties. I also grant that neuroscience sets constraints for psychological

theories and that neuroscience and psychology ”co-evolve” instead of being independent from each other. For these reasons, I prefer to call the position I have defended *pluralistic physicalism* to distinguish it from traditional forms of nonreductive physicalism (Eronen forthcoming). The key elements of pluralistic physicalism are thoroughgoing explanatory and causal pluralism, and robustness as the criterion for the reality of properties. This provides a naturalistic and scientifically credible position that gives an account of the ontological status of mental properties without relying on the notion of realization. In the next section, I show how a pluralist of this kind can deal with the causal exclusion argument.

Facing the causal exclusion argument

All metaphysical positions that are nonreductive enough to deny the identity of mental and physical properties face the causal exclusion argument. However, I believe that underlying the worries of causal exclusion is an outdated understanding of the nature of causation and its role in science. Philosophers of mind (e.g., Kim 1998, 2005) commonly think of causation as a relation where the cause generates, produces, or brings about the effect. This also naturally binds causation to physics, since the generation or production of an effect is clearly a physical matter. The assumption is that causation is in the first place physical business, and that the burden is on the proponent of higher-level causation to show how higher-level causation is compatible with physical causation. However, what exactly is the nature of this physical causation, or of causation in general, is an issue that few philosophers of mind have tackled.

If we take a more naturalistic stance and look at the role of causal notions in actual science, the picture changes. Causal notions are constantly employed in the special sciences and play a crucial role there, but in fundamental physics the situation is quite different. According to a venerable tradition in philosophy of science that goes back to Russell (1912-13) and has gained broad support in the last years (e.g., Ladyman & Ross 2007; Loewer 2007; Norton 2007), causal notions are not important in fundamental physics, and we find there nothing that resembles our common sense ideas of causation. 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. 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. Furthermore, there are cases even

in Newtonian physics which go straight against our ideas of causation – for instance, effects that take place with no observable causes (Norton 2007) – not to even speak of phenomena like quantum entanglement. There are also differing views regarding the role of causation in fundamental physics (see, e.g., Frisch 2009), but in this paper, I will assume that the noncausal view is correct, and explore its consequences for the causal exclusion argument.

In contrast to fundamental physics, it is uncontroversial that special sciences (including psychology) are busy with uncovering causal relations in the world, and an important aspect of research in the special sciences is singling out causal relations from mere correlations. As several philosophers have recently argued, causal relations in the special sciences are best understood as relations that are potentially or ideally exploitable for manipulation or control. Causes are such that intervening on them makes a difference to the effect. This is the ”interventionist” account of causation that has recently come to prominence (Pearl 2000; Woodward 2003, 2008; Woodward & Hitchcock 2003, also Spirtes, Glymour, and Scheines 1993).

I will focus here on James Woodward's (2003) account of interventionism, since it is exceptionally clear and elaborate. 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). Interventions are not only human activities, there are also ”natural” interventions, and the definition of an intervention makes no essential reference to human agency. Of course, several restrictions must be imposed on acceptable interventions – the necessary details are in Woodward (2003).

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 feature 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. In what follows, I often talk of variables and properties interchangeably, since nothing crucial turns on this.

The interventionist account seems to capture the nature of causation both in special sciences and everyday life very well. In fundamental physics, causal notions are apparently 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. Let us assume this is the case. How does our understanding of the causal exclusion argument then change?

The argument can be presented in a simple and clear form as the following five principles that cannot all be true (Bennett 2008):

Distinctness: Mental properties are distinct from physical properties.
Completeness: Every physical occurrence has a sufficient physical cause.
Efficacy: Mental events sometimes cause physical ones, and sometimes do so in virtue of mental properties.
Non-Overdetermination: The effects of mental causes are not systematically overdetermined; they are not on a par with the deaths of firing squad victims.
Exclusion: No effect has more than one sufficient cause unless it is overdetermined.

Several authors (e.g., List & Menzies 2009; Raatikainen 2010; Shapiro & Sober 2007; Woodward 2008, unpublished manuscript) have recently argued that interventionism provides a nonreductive solution to the exclusion argument. Building on the work of these authors, I will tackle the above general form of the argument and show why it does not carry through in the interventionist framework. One of the five principles has to turn out false, and since I do not appeal to type physicalism, it cannot be Distinctness. I will focus here on the most likely candidates, Exclusion and Non-Overdetermination.

Let us start with Exclusion. A straightforward interventionist rendering of this principle would be something along these lines: If variable *M* is a difference-making cause for variable *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 at different times and contexts. 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. Let us suppose for the sake of example that neuroscientists have discovered a neural state such that this state results in Markus uttering ”Hello” whenever certain background conditions *B* obtain. We can represent this state with variable *N*, and assume that it can take values a, b, and c, where only c results in Markus uttering ”Hello”. The event of Markus uttering ”Hello” can be represented with binary variable *P* (such that 1 represents Markus uttering ”Hello”). Let us further suppose that there is a mental state, Markus' desire or intention or decision to say hello, that supervenes on *N* and that also causes Markus to utter ”Hello” (whenever certain background conditions *B* obtain). We can represent this with variable *M*, which can take values p and q, where only q results in Markus uttering ”Hello” (i.e., variable *P* taking value 1 instead of 0). Since *M* supervenes on *N,* it must be the case that whenever there is a change in the value of *M*, there is a change in the value of *N*, but *N* can change without there being a change in *M*. Let us therefore assume that values a and b of variable *N* correspond to value p of variable *M*, while value c of *N* corresponds to value q of *M*.

In this case, it appears that we can intervene on *N* to change *P* and we can also intervene on *M* to change *P*. Therefore, both *N* and *M* seem to be difference-making causes of *P*. It also seems that both the particular instantiation of *N* (the variable *N* taking value c) and the particular instantiation of *M* (the variable *M* taking value q) can be difference-making causes for the particular instantiation of *P* (the variable P taking value 1).

This example is contrived, but I find it extremely plausible that there are at least some real-life or scientific cases that follow this pattern. It seems that in situations like this we can form several noncompeting representations of the same situation: in one, the particular instantiation of the neural state *N* is a difference-making cause for the particular instantiation of *B*, and in another the particular instantiation of *M* is the difference-making cause.

Baumgartner (2010) and Hoffmann-Kolss (unpublished manuscript) have argued that in situations like this the mental variable cannot be a genuine cause, because it does not fulfill the requirements for interventionist causation. Due to supervenience, interventions on variable *M* always result in changes in variable *N*. This is a problem, because Woodward's definition of an intervention explicitly requires that the intervention to assess whether *M* is a cause of *P* should not change any

variable which is a cause of *P* and is not on the causal path that goes through *M* (Woodward 2003, 98). Variable *N* is not on the causal path from *M* to *P* and it is a cause of *P*. This seems to imply that interventions of the right kind to determine whether *M* is a cause of *P* are not possible, and *M* cannot be a cause of *P*.

However, one has to be careful when assessing situations like this.[7] Due to supervenience, the variables *M* and *N* are non-causally correlated. One of the fundamental rules of causal modeling is that you should not have variables that are non-causally correlated in the same representation. This constraint is commonly formulated as the Causal Markov Condition: conditional on its direct causes, every variable is independent of every other variable, except its effects (see, e.g., Hausman & Woodward 1999, 2004 for more). Representations where there are mental variables that supervene on physical variables clearly violate this condition. Supervenience creates a problematic kind of non-causal correlation.

There is an obvious reductive solution to this: in each situation, we can get rid of the higher-level causes, and finally we will have only physical causes and no non-causally dependent variables, and consequently no violation of the Causal Markov condition. However, this approach leads to an even more fundamental problem: as I pointed out above, there are good reasons to believe that causation is a notion for the special sciences and not for fundamental physics. If this true, claiming that higher-level causes should be eliminated in favor of physical causes is absurd. Causation would drain to some fundamental physical level, where there is no causation. This solution also runs counter to scientific practice: As Woodward (2010) points out, 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. It is not the case that scientists always choose the maximally precise or lowest-level representation.

In my view, the better option is to accept that it is possible to build several noncompeting representations of the same situation, each having (instantiations of) different properties as the difference-making cause of one and the same (instantiation of a) property. In the above case, one can decide to include either the mental variable *M* or the neural variable *N* in

7 The following approach to dealing with the exclusion problem in the interventionist framework was suggested to me by Dan Brooks, for which I am very grateful.

the representation, depending on the context and the interests of the inquirer.[8]

This can be seen as a denial of the principle Exclusion, or alternatively as a denial of Non-Overdetermination. This depends on whether cases such as the above, where there are two difference-making causes for one effect, count as overdetermination. I am inclined to think that they do not – they surely do not resemble classic cases of overdetermination, such as two bullets hitting the heart of the victim at exactly the same moment. If this is the case, this means that the exclusion principle is false. The revised exclusion principle formulated above states: ”If this particular instantiation of *M* is a difference-making cause for this particular instantiation of *B*, then there is no other difference-making cause for this particular instantiation of *B* (unless this is a case of overdetermination).” We have now seen that there can also be other difference-making causes of the particular instantiation of *B*, and in cases that do not count as overdetermination.

Denying Exclusion has been traditionally considered unacceptable, but if we understand causation as a matter of difference-making and manipulation and control (and not as physical “bringing about”), violation of the exclusion principle does not pose any fundamental problems (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 contextual matters.

To conclude: if we understand causation in difference-making terms, the exclusion argument fails to provide convincing grounds for denying the causal efficacy of mental properties. The exclusion principle turns out false in the interventionist framework. This might be counterintuitive, but

---

8 One possible problem for this position arises from the fact that interventions (in Woodward 2003) are not defined relative to a representation or variable set: whether we can intervene on M with respect to B is independent of the representation or variable set, and there are reasons to suspect that we cannot intervene on M with respect to B, because M supervenes on N, which is also a cause of B (Baumgartner 2010). However, this problem can be avoided by adopting Woodward's (unpublished manuscript) revised definition of an intervention (IV*), where changes in the supervenience base of M are not taken into account when assessing whether there is an intervention on M with respect to B. I thank Michael Baumgartner for bringing this problem to my attention.

the other option (accepting only fundamentally physical causes) is even more counterintuitive and problematic.

Conclusions

Realization physicalism can be seen as an attempt of naturalizing the mental while keeping it ontologically autonomous. I have argued that realization physicalism does not succeed in this. It is scientifically implausible and fails to give a solution the problem of causal exclusion and the autonomy of the mental properties.

The more naturalistic approach that I have defended recognizes mental properties as real insofar as they are robust, and appreciates the heterogeneity of mental properties and their relations to physical properties. The causal exclusion problem can be resolved, since when we adopt the difference-making approach to causation, the exclusion principle does not hold. Mental properties need not be identical to physical properties or physically realized in order to be real and causally efficacious.

References

Baumgartner, Michael. 2010. "Interventionism and Epiphenomenalism." *Canadian Journal of Philosophy* 40: 359-383.

Bechtel, William. 2008. *Mental Mechanisms. Philosophical Perspectives on Cognitive Neuroscience*. London: Routledge.

Bechtel, William, and Robert C. Richardson. 1993. *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Princeton: Princeton University Press.

Bennett, Karen. 2003. "Why the Exclusion Problem Seems Intractable, and How, Just Maybe, to Tract It." *Noûs* 37: 471-497.

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

Clapp, Lenny. 2001. "Disjunctive Properties: Multiple Realizations." *The Journal of Philosophy* 98: 111-136.

Craver, Carl F. 2007. *Explaining the Brain*. Oxford: Oxford University Press.

Cummins, Robert. 2000. "'How does it work?' versus 'what are the laws?' Two conceptions of psychological explanation." In *Explanation and Cognition*, ed. F. Keil and R. Wilson, 117-144. Cambridge: MIT Press.

Eronen, Markus I. 2010-2011. "Replacing Functional Reduction with Mechanistic Explanation." *Philosophia Naturalis* 47-48: 125-153.

Eronen, Markus I. 2011. *Reduction in Philosophy of Mind: A Pluralistic Account*. Frankfurt (Main): Ontos.

Eronen, Markus I. Forthcoming. "Pluralistic Physicalism and the Causal Exclusion Argument." *European Journal for Philosophy of Science.*

Frisch, Mathias. 2009. "'The Most Sacred Tenet'? Causal Reasoning in Physics." *British Journal for the Philosophy of Science* 60: 459-474.

Gillett, Carl. 2003. "The Metaphysics of Realization, Multiple Realizability, and the Special Sciences." *The Journal of Philosophy* 100: 591-603.

Gillett, Carl. 2010. "Moving beyond the subset model or realization: The problem of qualitative distinctness in the metaphysics of science." *Synthese* 177: 165-192.

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

Hausman, Daniel M., and James Woodward. 1999. "Independence, Invariance and the Causal Markov Condition." *British Journal for the Philosophy of Science* 50: 521-583.

Hausman, Daniel M., and James Woodward. 2004. "Manipulation and the Causal Markov Condition." *Philosophy of Science* 71: 846-856.

Hoffmann-Kolss, Vera. Unpublished manuscript. "The Supervenience Argument Is Alive and Kicking."

Kim, Jaegwon. 1992. "Multiple realization and the metaphysics of reduction." *Philosophy and Phenomenological Research* 52: 1–26.

Kim, Jaegwon. 1993. "The non-reductivist's troubles with mental causation." In *Mental Causation*, ed. J. Heil and A. Mele, 189-210. Oxford: Clarendon Press.

Kim, Jaegwon. 1998. *Mind in a Physical World*. Cambridge, MA: MIT Press.

Kim, Jaegwon. 2005. *Physicalism, or Something Near Enough*. Princeton: Princeton University Press.

Kuorikoski, Jaakko, Aki Lehtinen, and Caterina Marchionni. 2010. "Economic Modelling as Robustness Analysis." *British Journal for the Philosophy of Science* 61: 541-567.

Ladyman, James, and Don Ross. 2007. *Every Thing Must Go: Metaphysics Naturalized*. Oxford: Oxford University Press.

Levins, Richard. 1966. "The strategy of model building in population biology." *American Scientist* 54: 421-431.

Lewis, David. 1972. "Psychophysical and theoretical identifications." *Australasian Journal of Philosophy* 50: 249-258.

List, Christian, and Peter Menzies. 2009. "Non-Reductive Physicalism and the Limits of the Exclusion Principle." *Journal of Philosophy* 106: 475-502.

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

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

Menzies, Peter. 2008. "The exclusion problem, the determination relation, and contrastive causation." In *Being Reduced*, ed. J. Hohwy and J. Kallestrup, 196-217. Oxford: Oxford University Press.

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

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge, UK: Cambridge University Press.

Pereboom, Derk. 2002. "Robust Nonreductive Materialism." *The Journal of Philosophy* 99: 499-531.

Polger, Thomas W. 2004. *Natural Minds*. Cambridge: MIT Press.

Polger, Thomas W. 2007. "Realization and the Metaphysics of Mind." *Australasian Journal of Philosophy* 85: 233-259.

Polger, Thomas W., and Lawrence Shapiro. 2008. "Understanding the dimensions of realization." *The Journal of Philosophy* 105: 213-222.

Polger, Thomas W. 2010. "Mechanisms and explanatory realization relations." *Synthese* 177: 193-212.

Putnam, Hilary. 1967. "Psychological predicates." In *Art, Mind, and Religion*, ed. W.H. Capitan and D.D. Merrill, 37-48. Pittsburg: Pittsburg University Press.

Raatikainen, Panu. 2010. "Causation, Exclusion, and the Special Sciences." *Erkenntnis* 73: 349-363.

Russell, Bertrand. 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.

Shoemaker, Sydney. 1980. "Causality and Properties." In *Time and Cause*, ed. P. van Inwagen, 109-136. Dordrecht: D. Reidel.

Shoemaker, Sydney. 2001. "Realization and Mental Causation." In *Physicalism and Its Discontents*, ed. C. Gillett and B. Loewer, 74-98. Cambridge: Cambridge University Press.

Spirtes, Peter, Clark Glymour, and Richard Scheines. 1993. *Causation, Prediction, and Search*. New York: Springer-Verlag.

Walter, Sven. 2010. "Taking realization seriously: no cure for epiphobia." *Philosophical Studies* 151: 207-226.

Weisberg, Michael. 2006. "Robustness analysis." *Philosophy of Science* 73: 730-742.

Wilson, Robert A., and Carl F. Craver. 2007. "Realization: Metaphysical and scientific perspectives." In *Handbook of the Philosophy of Psychology and Cognitive Science*, ed. P. Thagard, 81-104. Amsterdam: Elsevier.

Wimsatt, William C. 1976. "Reductionism, Levels of Organization, and the Mind-Body Problem." In *Consciousness and the Brain. A Scientific and Philosophical Inquiry*, ed. Gordon G. Globus, Grover Maxwell, and Irwin Savodnik, 205-267. New York: Plenum Press.

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

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

Woodward, James. 2003. *Making Things Happen*. Oxford: Oxford University Press.

Woodward, James. 2006. Some varieties of robustness. *Journal of Economic Methodology* 13: 219-240.

Woodward, James. 2008. "Mental causation and neural mechanisms." In *Being Reduced*, ed. J. Hohwy and J. Kallestrup, 218-262. Oxford: Oxford University Press.

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

Woodward, James. Unpublished manuscript. "Interventionism and causal exclusion." (http://philsci-archive.pitt.edu/id/eprint/8651)

Woodward, James, and Christopher Hitchcock. 2003. “Explanatory Generalizations, Part I: A Counterfactual Account.” *Noûs* 37: 1-24.

Yablo, Stephen. 1992. “Mental Causation.” *The Philosophical Review* 101: 245-280.