Causation, Exclusion, and the Special Sciences

Forthcoming in *Erkenntnis*

Panu Raatikainen

1. Introduction

Thinkers sympathetic to the autonomy of the special sciences – and emergentists in particular – typically think, not only that the special sciences are not reducible to the fundamental physical level, but also that the properties studied by the special sciences have causal powers of their own. More physicalistically inclined philosophers forcefully attack the latter idea. Their master argument is the so-called exclusion problem. It is based essentially on the idea that the physical world is causally closed, i.e., that everything physical which happens must be a result of purely physical causes.

The discussion has been most vigorous in the domain of the philosophy of mind, but the structure of the argument is entirely general, and is applicable to any special science with some common properties. Nevertheless, for concreteness, and because the issue is most familiar from that context, I shall discuss the problem in the context of the philosophy of mind. All the same, my arguments and conclusions, if sound, are applicable across the board.

2. The Causal Exclusion Problem

If only the identity theory of mind had worked, there would not be any problem with mental causation. For if mental events were (type) identical with physical events, it would not be surprising at all that mental events could cause physical events. However, the great majority of philosophers are now convinced that such a view cannot be correct. The main reason for this is the so-called multiple realizability argument: It seems plausible that a particular mental kind (property, state, or event) can be realized by many distinct physical kinds (see e.g. Putnam 1967, Fodor 1968, Block and Fodor 1972). In what follows, I shall simply assume that the possibility of multiple realizability is a fact – regardless of the right conclusions to be drawn from this.

However, the denial of the identity theory re-opens the door to the problem of mental causation: How can mental events, if they are not physical, have physical effects, such as behavior, as a consequence? The problem becomes particularly acute in the form of what is often called "the exclusion problem."[1] That is, consider the following five *prima facie* plausible theses:[2]

[1] Variants of the exclusion problem have been presented, e.g., by Malcolm (1968), Peacocke (1979) and Schiffer (1987). More recently, Kim (1989, 1992) and Papineau (1993, 2001) in particular have pressed the exclusion argument in defence of physicalism; and as Papineau (2001) argues, something like the exclusion argument (and the assumption of completeness of physics; see below) also drives the traditional type-identity theory of Smart (1959), Lewis (1966, 1972) and Armstrong (1968).

[2] I have borrowed this elegant way of summarizing the exclusion problem from Bennett (2007).

(1) *Distinctness:* Mental properties (and perhaps events) are distinct from physical properties (or events) (i.e., the type-identity theory is false).

(2) *Completeness:* Every physical occurrence has a sufficient physical cause.

(3) *Efficacy:* Mental events sometimes cause physical events, and sometimes do so in virtue of their mental properties.

(4) *No overdetermination:* The effects of mental causes are not systematically over-determined.

(5) *Exclusion:* No effect has (at a particular time *t*) more than one sufficient cause unless it is overdetermined.

The problem is that these claims seem to be incompatible.[3] Philosophers, being what they are, have certainly tried to wriggle out of this deadlock in all possible ways. But most often, (4) and (5) are taken as background assumptions which are not questioned.

Physicalists typically commit themselves to (2), so, for them, the problem must be with either (1) or (3). But the denial of (3) amounts to the conclusion that mental causation is an illusion, that is, epiphenomenalism. This unattractive result has led many physicalists to deny (1) and to attempt to vindicate, after all, reductionism, or the identity theory, in some form. Emergentists, unwilling to give up (1) or (3), end up denying (2) and accepting 'downward' causation. Yet, (2), the causal completeness of the physical world, is certainly intuitively appealing. Indeed, even many philosophers less inclined towards physicalism (e.g. Chalmers 1996) see it as compelling. So perhaps it should not be given up too quickly without an argument or analysis.

What makes the exclusion problem so difficult is this: It is not that something about the mental makes it unsuitable to be a cause (as some other arguments questioning mental causation suggest). Rather, the problem is in the physical. Given that every physical event already has a sufficient physical cause, there is no room for the mental to cause anything, even if the mental were in principle able to work as a cause of something.

One can, of course, dispute the details of the exclusion argument, and its fine points certainly deserve more attention. The literature, however, is very wide ranging, and it is simply not possible to review the numerous attempted solutions in the space of a paper like this. It is perhaps fair to say, though, that there does not exist any widely accepted solution. Although many want to resist the argument, philosophers' opinions vary greatly on what exactly goes wrong in it (cf. Bennett 2007).

In any case, much of this literature seems to me to be beside the point. Some assume the outdated idea that causation requires laws; others the arguably misguided conception of causation as a sort of

---

[3] Note that the exclusion argument, if sound, generalizes: it threatens to make all properties studied by special sciences (e.g. properties of biology) which are purportedly distinct from (i.e., not type-identical with) underlying physical properties causally inefficacious. Hence, there is much more at stake here than just the mental.

production, involving contact and transmitting something; and still others take too much for granted the classical Lewisian counterfactual theory of causation which also faces serious problems. Some differentiate causation and explanation, and propose that though there is no genuine causation at the level of the mental (or, in general, at the level of the special sciences), explanations in terms of mental states or events (or whatever) may nevertheless be useful. I, for one, find such "solutions" unattractive. Would it not be preferable to have an account in which (causal) explanations explain by citing genuine causes? Still others rely on rather specific assumptions about the metaphysics of properties or events. Though I think such approaches are not without interest, it would certainly be better to have a way out which is independent of such detailed metaphysical views.

Consequently, I strongly feel that one needs a new start here. Accordingly, instead of doing more armchair metaphysics, I will attempt to shed some new light on the problem of mental causation (and more generally, on causation in the special sciences) by taking into account certain advances in recent theorizing on causation in the philosophy of science. The theory of causation I rely on has been developed independently of the whole debate on mental causation and the exclusion problem. Moreover, it has considerable intrinsic plausibility. Hence, it should be interesting in the present context to see if it can provide any clarification.

3. Causation and Fundamental Physics

Before going to the theory of causation in question, however, I want to make a brief digression to causation at the fundamental physical level. For, the exclusion argument apparently generalizes (see fn 2) and entails, from the physicalistic perspective, that there is genuine causation only at the fundamental micro-physical level. It would be important to note, though, that this is emphatically not how numerous distinguished philosophers of physics and experts in the theory of causation view the issue. Namely, quite independently of any considerations of mental causation, the exclusion problem, or of the idea of downward causation, the whole idea of causation becomes problematic at the level of fundamental physics.

To begin with, it is a historical fact that the notion of cause has disappeared from physics as the subject has developed (see Kuhn 1971; cf. Loewer 2001). More importantly, many philosophers who apparently know their physics have argued that the whole idea of causation is not even applicable to fundamental physics, or is incompatible with it. Very briefly, and roughly, one of the problems is that in some cases, one has to specify the entire state of the whole universe at one time in order to determine the state of even a small region at some later time. And in such a case, it is difficult indeed to consider anything as a cause (Latham 1987, Redhead 1990, Field 2003, cf. Loewer 2001, Hitchcock 2007, Elga 2007; this idea goes back, of course, to Russell 1912-13).

More specifically, from the point of view of the interventionist approach to causation (to which we shall later turn in this paper), one may doubt the meaningfulness of the notion of cause in the context of fundamental physics. Judea Pearl, a key developer of the interventionist theory, writes: "If you wish to include the whole universe in the model, causality disappears because interventions disappear – the manipulator and manipulated lose their distinction" (Pearl 2000, p. 350; see also Hitchcock 2007). James Woodward, another important figure in the interventionist approach, is (characteristically to him) more cautious, but still says that "causal ascription becomes less natural

and straightforward – increasingly strained – when candidate causes expand to include the state of the entire universe" (Woodward 2007, 93).

Some may be prepared to follow Russell and abandon the whole notion of causation as an outdated folk notion which has no place in advanced science. However, as Nancy Cartwright (1979) has argued, totally abandoning the concept of causation would cripple science (cf. Field 2003; Hitchcock 2007): she does not have in mind fundamental physics, but more ordinary science, such as the search for the causes of cancer. The notion of cause is intimately connected with the distinction between effective and ineffective strategies. For example, if smoking causes lung cancer, then stopping smoking is an effective strategy for avoiding cancer. Consequently, most philosophers who are skeptical about causation at the fundamental physical level have still concluded that causation has a firm place at least in the special sciences.

Now, in what follows, I do not want to commit myself to the view that there is no causation in fundamental physics – I do not even want to pretend that I am competent to judge the issue. However, it is important to keep in mind that many able philosophers have concluded this. But be that as it may, it just is not the case – contrary to what numerous physicalistic metaphysicians take for granted – that causation is *uncontroversially* present in fundamental physics. Therefore, such philosophers should perhaps think twice before declaring, from the armchair, that there is causation *only* at the fundamental physical level.

4. The Interventionist Theory of Causation

Recently, a 'manipulationist' or 'interventionist' theory of causation has emerged in the philosophy of science. It has been developed especially by James Woodward (1997, 2000, 2003), although related ideas have been put forward, e.g., by Pearl (2000) and Spirtes, Glymour and Scheines (2000).[4] This theory can be viewed as a variant of the counterfactual theories of causation, but it is particularly attractive in its avoidance of many well-known problems of the more traditional counterfactual theories. The theory can also be seen as a sophisticated version of the general idea of causes as difference-makers.[5] Furthermore, the interventionist theory also embodies the idea that causal claims are essentially contrastive (see below).

One way of motivating this approach is to ask the questions: What is the point of our having a notion of causation (in contrast to, say, a mere notion of correlation) at all? What role or function does this concept play in our lives? Why do we care to distinguish between causal and merely correlational relationships? (cf. Woodward 2003, p. 28) According the interventionist approach, the answer is that such knowledge of genuine causal relationships is, *sometimes*, practical and applicable: by manipulating the cause we can influence the effect. If there is a real causal relationship between *A* and *B*, manipulating *A* is a way to change *B*; Mere correlation between *C*

[4] Manipulationist theories of causation, in fact, have a longer history. Earlier variants include Collingwood 1940, Gasking 1955, von Wright 1971, and Menzies and Price 1993. These tend to be, however, problematically anthropocentric, subjectivistic and reductionistic, and are moreover threatened with circularities. The more recent interventionist variants apparently avoid such problems; cf. Woodward 2001.

[5] For the idea of causes as difference-makers, see Menzies 2007; cf. Menzies 2008; List and Menzies 2009.

and *D*, on the other hand, just disappears if one attempts to affect *D* by manipulating *C*. (Obviously, our knowledge of causal relationships and our interest in them need not be restricted only to applicable causal relations; it can certainly be purely theoretical and based on curiosity. Not all science is applied science.) Thus, we can try to find a cure for AIDS, suppress poverty or prevent eutrophication of the Baltic Sea on the basis of knowledge about the causal relationships associated with them.

Real causal relationships can, in favorable circumstances, be distinguished from accidental correlations experimentally, by manipulating the initial conditions (the putative causes) and investigating whether this has consequences on the effects (surely, this is often in practice impossible). The interventionist theory of causation thus emphasizes the close connection between causal thinking and experimental research (and manipulation and control).

The interventionist theory of causation has been developed into a sophisticated theory, but its basic idea can be explained quite simply. It connects causal claims with counterfactual claims concerning what would happen to an effect under interventions on its putative cause. Roughly, *C* causes *E* if and only if an intervention on *C* would bring about a change in *E*. Slightly more exactly, causal claims relate, in this approach, variables, say *X* and *Y*, that can take at least two values. These may often be some magnitudes (such as temperature, electric charge or pressure), but in simple cases, they may also be just discrete alternative events or states of affairs. The idea now is that were there an intervention on the value of *X*, this would also result a change in the value of *Y*.

Heuristically, one may think of interventions as manipulations that might be carried out by a human agent in an idealized experiment. Nevertheless, the approach is in no way anthropocentric, and intervention can be defined in purely causal terms (that a causal vocabulary is presupposed means that the theory does not aim to give a reductive analysis of causation. This does not make the approach viciously circular: "*X* causes *Y*" is explicated with the help of *other* causal relations and correlational information.)

In order to distinguish genuine causation from other ways in which an intervention *I* that changes *X* might be associated with changes in *Y*, some further conditions must be added. Roughly, it is required that *I* does not cause *Y* directly via a route that does not go through *X*, that *I* not be correlated with other causes of *Y* besides those causes that lie on the causal route (if any) from *I* to *X* to *Y*, and so on.[6]

As was already noted, this approach is a version of the counterfactual theories of causation. According to the interventionist account, whether a relation is causal can be evaluated with the help of counterfactuals which have to do with the outcomes of hypothetical interventions. Such counterfactuals are called "active counterfactuals." These are such that their antecedents are made true by an intervention. Active counterfactuals have the form:

> If *X* were to be changed by an intervention to such and such a value,
> the value of *Y* would change.

[6] For an exact definition, see Woodward 2000, 2003.

It has become increasingly popular to think that causal claims do not in fact describe a simple binary relation between two events, but rather involve (even if often only implicitly) a contrastive class for both cause and effect, that is, they contrast alternatives to the putative cause and effect (see e.g. Hitchcock 1996; Menzies 2008), The interventionist approach to causality, which relates variables, also incorporates this idea: variables can take different values; different choices of possible alternative values lead to different contrast classes. In its context, some contrasts need to be fixed; otherwise, causal claims are not even unambiguous. If, for example, *X* could take as its value either x_1 or x_2, and *Y* either y_1 or y_2, the relevant causal claim, with the contrasts made explicit, could be:

X's being x_1 (rather than x_2) causes Y's being y_1 (rather than y_2).

Note that different choices of contrasts, say, x_3 and y_3, for the same x_1 and y_1, for example, lead to different causal claims, some of which may be false, some true. The most natural, "default" contrast, though, is – unless there is some specific reason to choose differently – that the presence rather than absence of the property (or whatever) at issue is caused by the presence of the another appropriate property (or whatever) rather than absence of it (see Woodward 2003, p. 67-8, 145-6; cf. Woodward 2008, p. 235-236; see also Menzies 2008).[7] In this standard case, the relevant active counterfactual would have the form:

(1) If *X*'s being x_1 were to be changed by an intervention to *X*'s not being x_1, then *Y* would change from being y_1 to not being y_1.

But it should be kept in mind that there may be special grounds to choose the contrast differently.

Now, this is not the right place to try to defend the interventionist theory of causation.[8] Suffice it to say that it is in various ways a promising and intuitively attractive theory, and seems to be successfully gaining ground in the philosophy of science, and the theory of causation in general. What I want to do in this paper is only to consider mental causation and the exclusion problem from the perspective of this theory of causation.

The first thing to note is that mental states or events are perfectly legitimate candidates for the role of causes in the proposed account. It is indeed commonplace to affect peoples' behaviour by manipulating their beliefs and/or desires. For example, Nazi propaganda was able to bring about violence towards Jews in *die Kristallnacht* by making people believe that there was a Jewish conspiracy behind the murder of a certain German diplomat. Less dramatic examples can easily be found in marketing and advertising, and the psychological research supporting them.

[7] For more about default contrasts, see also Hitchcock 2007 and Menzies 2009.

[8] Woodward 2003 is a book-length defense of this approach; see also Woodward 2004.

Two characteristics of the interventionist approach deserve special attention in the present context. First, it is nowhere required that a cause must be individuated with physicalistic concepts. All that is required is that it would make sense to manipulate it (although, it is obviously not required that it is in all cases humanly possible to manipulate it in practice). Second, no laws are required to subsume the cause and the effect in order for there to be causation. Among other things, this undermines a key premise of Davidson's anomalous monism. Nevertheless, these observations do not, as such, answer the worry about the exclusion problem. However, I aim to show that the interventionist theory of causation can in fact be helpful in our attempts to answer it.

5. The Argument

Now there is an argument, discovered independently (at least) by the present author and Peter Menzies (see Raatikainen 2006, 2007; Menzies 2008),[9] which shows that from the interventionist perspective, a mental state can truly be a cause of, e.g., behavior; and more drastically, that – at least in some ways of conceptualizing the situation[10] – the underlying physical state may fail to be the cause.

I prefer to present the argument with the help of a concrete (and perhaps even a bit oversimplifying) example: Thus, assume that, at the moment, John desperately wants a beer. This is part of our constant background, which does not vary. Suppose, then, that he forms a firm belief (say, he suddenly remembers that he has earlier bought a six pack of beer and put it in the refrigerator) that there is some beer in the refrigerator. Consequently, he walks to the refrigerator to get a beer. Suppose that this is what actually happens (i.e., this is stipulated to be our actual world below). Can John's belief now be taken as the cause of his behavior? Or is it rather John's brain state (or brain event; or whatever underlying physical state), call it *B*, at the moment?

Let us imagine, counterfactually, the following intervention *I*: Peter, John's roommate, walks into the room and informs John that he has drunk all John's beers in the refrigerator (even if Peter's actions were not fair, John has no reason to doubt that Peter is telling the truth). John then gives up the belief that there are beers in the refrigerator. Accordingly, John, instead of going to the refrigerator, leaves for the closest grocery to buy more beer.

John either has the belief that there is some beer in the refrigerator ($X = x_1$), or he does not have it ($X = x_2$). In the former case, he goes to the refrigerator ($Y = y_1$), in the latter case he goes to the grocery ($Y = y_2$). Let us suppose, for simplicity, that these cases exhaust all possible cases. It looks as if Peter's hypothetical interference satisfies all the conditions of a proper intervention (see also below).

---

[9] Also Carl Craver (2007, pp. 223-4) briefly sketches what seems to amount to the same argument, giving credit to Eric Marcus (unpublished). Thus, such an argument seems to be very much in the air. [While this paper was under review, Woodward 2008 also came out; Woodward himself is developing there many ideas that are quite similar to those expressed in this paper. I have added a couple of references to Woodward's paper in order to help comparison.]

[10] That is, with certain natural ways of choosing the contrasts.

There are in fact two significantly different kinds of causal claims that can be considered from the interventionist perspective, claims about the causal relevance[11] of a variable *X* to another variable *Y*, and claims about a variable's particular value's (e.g. $X = x_1$) being a cause of a particular value of another variable (e.g. $Y = y_1$), given the contrasts.[12] Let us first reflect on the former.

For a variable *X* to be causally relevant for another variable *Y*, it is sufficient, according to the interventionist account, that *some* changes, produced by some intervention, in *X* lead to a change in *Y*. It should be noted just how weak a requirement this really is (though, not trivial: mere correlations fail to satisfy it). Now it can be seen quite easily (see also below) that the above variable *X* (about John either having the belief or not) is causally relevant for *Y*: an intervention, e.g. Peter's hypothetical interference, which changes the value of *X*, brings about a change in the value of *Y*.

But how about the brain state *B*? This depends vitally on how we set the contrasts and choose the relevant variables. We may well let the variable *Z* (for the alleged cause) to range over a number of different possible, mutually exclusive brain states of John, including *B* above (i.e. the brain state which actually realizes John's belief that there is some beer in the refrigerator); let $Z = z_1$ just in the case when John is in the brain state *B*. In that case, the variable *Z* is also causally relevant for *Y*: at least *some* changes in *Z* lead to a change in *Y* too.[13] However, this is still a rather modest conclusion, and should by no means be thought as suggesting that *X* and *Z* are somehow in competition here, in the spirit of the exclusion argument. The situations where several variables are causally relevant in this sense to an effect variable are very common (in any case, they could not possibly be competing in this sense; see Section 7). This is just a consequence of the fact that very little is required for such a causal relevance between variables, and the conclusion is not particularly exciting. As Woodward himself puts it, the bare claim that *X* is causally relevant for *Y* is "not very informative"; "what one would really want to know", he continues, "is not just whether there is some manipulation of (intervention on) *X* that will change *Y*. One would also like to have more detailed information about just which interventions on *X* will change *Y*" (Woodward 2003, p. 66).

If, on the other hand, we prefer instead to follow Woodward and focus on the natural or default contrast, the alternative values of *Z* would be just "John has the brain state *B*" ($Z = z_1$) and "John does not have the brain state *B*" ($Z = z_2$). In this case, things become much more interesting, as we will soon see. In order to evaluate whether we should consider John's belief or his brain state (or

[11] In the interventionist literature, if the variable *X* is causally relevant for the variable *Y*, it is often said that *X* causes *Y*. This manner of speaking admittedly deviates from the normal usage. In what follows, I only talk about "causal relevance" in such cases, just in order to keep these two kinds of causal claims clearly distinguished. But this is purely verbal choice from my part – nothing really hinges on this choice.

[12] It is sometimes suggested that these correspond one-one to the so-called type-causation and token-causation, familiar from the philosophical literature on causation. However, I do not think that the issue is this simple. In particular, I think that both these sorts of claims can be meaningfully made at least in the type level (cf. Menzies 2008).

[13] Consequently, if we were to follow the somewhat deviant interventionist manner of speaking (cf. fn 11), we could say both that *X* causes *Y* and that *Z* causes *Y*.

both?) as the cause of his behavior (going to the refrigerator) *in this sense*, let us consider the following two *active counterfactuals*:

> (1) If John's belief that there is beer in the refrigerator were to be changed by an intervention to not having the belief, he would have gone to the grocery (and not to the refrigerator).
>
> (2) If John's brain state *B* were to be changed by an intervention to not having that state, he would have gone to the grocery (and not to the refrigerator).

Now according to the standard possible-world analysis of counterfactual conditionals, '$P \rightarrow Q$' is true if and only if either there is no *P*-world, or some *P* & *Q*-world is more similar to the actual world than any *P* & not-*Q*-world. The analysis makes '$P \rightarrow Q$' trivially true when *P* is impossible, which is when there is no *P*-world.[14]

Obviously it would have been possible that John had neither the belief nor the brain state *B*; hence, we must focus on the second case. It is quite clear that (1) emerges as true; only by postulating some further differences from the actual world can we make the antecedent true but the consequent false. Hence, John's belief is indeed causally relevant for his behavior.

But what about (2)? Given that we have granted the possibility of multiple realizability, it should be possible for there to be another brain state *B´*, one that is different from *B*, which can also realize the belief that there is some beer in the refrigerator. Hence, there is a possible world *w* in which an intervention changes John's brain state from *B* to *B´*, and John nevertheless goes to the refrigerator and not to the grocery. So this is a *P* & not-*Q*-world. Moreover, *w* seems to be, by all standards, much more similar to the actual world than the one where John does not believe that there is some beer in the refrigerator and consequently goes, instead of the refrigerator, to the grocery. Hence, (2) apparently comes out as false.[15] And consequently, if we hang onto the default contrast, the variable *Z* (whether John's has *B* or not) is not even causally relevant for the variable *Y* (whether John goes to the refrigerator, or to the grocery).

Let us next look at causal claims about particular values of variables, and first, with respect to John's belief. Now the causal claim "John's having the belief (that there is beer in the refrigerator) caused him to go to refrigerator", in symbols, "$X = x_1$ causes $Y = y_1$", is true if and only if, first, it is actually the case that $X = x_1$ and $Y = y_1$, and second, if an intervention were to change the value of *X*

---

[14] Woodward has certain reservations about the standard Lewis-Stalnaker analysis of counterfactuals. But in the present example, its possible problems appear to be irrelevant. The relative similarity between worlds seems to be sufficiently clear in these cases, and no violation of the laws of nature, or "miracles", are involved. Neither is Lewis's ultra-realism about possible worlds assumed.

In the interventionist literature, counterfactuals are evaluated instead by systems of equations. In the present simple case, this approach gives apparently the same results. I have leaned here on the possible world approach because it is more familiar.

[15] In my own case, this argument was inspired by Tim Crane's quite similar argument with respect to a more traditional counterfactual approach to causation (Crane 2001, 64-65), though others seem to have been able to arrive at the idea independently of it.

from x_1 to x_2, the value of Y would change from y_1 to y_2 (which amounts to our active counterfactual (1) above). It follows immediately from the above considerations that this causal claim is true.

The case of brain states (or whatever underlying physical properties) is also straightforward here. We have stipulated that the actual values of Z and Y are z_1 and y_1, respectively. However, if we again focus on the default contrast, the relevant second condition is simply our above counterfactual (2), and comes out as false. It would be therefore wrong to say that $Z = z_1$ causes $Y = y_1$. In other words, the causal claim, with contrasts made explicit,

> John's having the brain state *B* (rather than not having it) caused his going to the refrigerator (rather than to the grocery),

is false. Thus, according to this analysis, the brain state *B* is not, contrary to all appearances, the cause of John's behavior (his going to the refrigerator), but John's belief is. Consequently, mental states (or events) can be genuine causes, i.e., there is, in a sense, downward causation. Of course, the occurrence of *B* is surely sufficient for the effect, John's behavior, but that does not make it the *cause* of the latter. Being sufficient condition for the occurrence of something, and being its difference-making cause, must thus be clearly distinguished.

6. Some Elaboration of the Argument

The above argument certainly deserves, and requires, further elaboration. To begin with, in the interventionist approach, one can distinguish between various different notions of cause. First, one can contrast the notion of *a contributing cause* with the notion of *a total cause.* And, second, there is the notion of *a direct cause,* in contrast to the notion of a non-direct cause. The notion of a total cause allows a rather simple interventionist definition,[16] but the notions of a contributing cause and of a direct cause involve certain difficulties, and require more sophisticated definitions. Nonetheless, the notion of a contributing cause is needed primarily in the cases of cancellation.[17] And whatever are the complications with mental causation, there does not generally seem to occur such cancellations. Moreover, whether some cause is direct or not depends heavily on our way of conceptualizing the situation – on which factors we decide to consider explicitly as variables. Consequently, one need not perhaps worry too much about such fine distinctions here, and one may focus on the general idea of the interventionist approach.

Of course, one must also make sure that the alleged intervention *I* is indeed a genuine *intervention*. To begin with, could an intervention *I*, in our simple example (for example, Peter's hypothetical interference), cause *Y* directly without going through *X?* It does not seem so: if an intervention (such as Peter's utterance) failed to change *X*, John's belief, John would have still gone to the

[16] (TC) *X* is a *total cause* of *Y* if and only if there is a possible intervention on *X* which will change *Y* (or the probability distribution of *Y*); see Woodward 2003, p. 45, 51.

[17] As, for example, in Heslow's (1976) classical example, in which birth control pills both directly cause an increased probability of thrombosis, but also lower the probability of pregnancy, which is itself a positive probabilistic cause of thrombosis.

refrigerator (i.e., no change in *Y*). Or could *I* be correlated with other causes of *Y* besides those causes that lie on the causal route (if any) from *I* to *X* to *Y*? Again, if *I* did not change *X*, John's belief, there does not seem to be any other route through which it could influence *Y*.

Could there be a common cause for *X* and *Y* such that *X* and *Y* are not causally related? In that case, it should be possible to vary the value of *X* by an intervention without a change in *Y* (while everything else remains unchanged). Once more, this is apparently impossible. If John's belief is changed, his behavior changes too (other things being equal). And quite clearly, in the counter-factual scenario, Peter's report is a cause of the change of John's belief. In sum, Peter's interference can indeed be taken as a true intervention.

7. The Question of Overdetermination

It is a plausible and widely accepted thesis that everything that exists supervenes on the fundamental physical level, i.e., that the physical facts determine all possible higher-level facts, with metaphysical necessity. At least, it seems that any physicalist must assume so.

Now the philosopher's standard examples of apparently rare cases of overdetermination are such as a death caused by several members of a firing squad shooting simultaneously. As has been noted by some philosophers even independently of the interventionist approach, the relation between a mental state and its underlying physical state is much more intimate than between e.g. the individual shooters of the squad (see e.g. Loewer 2001, Funkhauser 2001).

From the interventionist perspective, however, somewhat surprising consequences follow for the whole overdetermination issue. That is, even raising the question whether a mental state and the physical state realizing it overdetermine the effect or not, requires that we consider a causal system which includes a variable for both. However, this is turn commands that one can, at least in principle, vary their values independently of each other (like one could, by a hypothetical intervention, prevent one shooter firing his gun without affecting the others, in the above firing squad case). But in as much as it is necessary that the facts of the physical level determine the mental level (supervenience), this is simply impossible, and consequently, the question of overdetermination does not even make sense in this context. And if this is so, a key premise of the exclusion argument, (4), seems to fail to make sense.[18] This gives us, from the interventionist point of view, another independed reason for doubting the whole exclusion argument.

8. Completeness and Exclusion Revisited

Consider now again the two other premises of the exclusion argument, namely, Completeness and Exclusion:

(2) *Completeness:* Every physical occurrence has a *sufficient* physical *cause*.

[18] Woodward now seems to have ended up with a similar conclusion (see Woodward 2008, Section 6).

(5) *Exclusion:* No effect has more than one *sufficient cause* unless it is overdetermined.

Note now that from the point of view of our preferred view of causation here, both these assumptions involve confusing causes with sufficient conditions. There are causes, which are difference-makers; and there are sufficient conditions, which are wholly different issues and not causes of any sort; there are no such things as *sufficient causes*. Hence, I do not think that these two assumptions are so much false (or true) as mongrels based on a conceptual confusion which fail to make clear sense. After all, the whole point of the exclusion argument and the debate surrounding it is to ask whether the mental is capable of being a *cause* of something physical. But then, surely the argument and its premises should talk about causes and not be formulated in terms of sufficient conditions.

For somewhat similar reasons, List and Menzies (2009) propose that we would revise Exclusion (as they have formulated it)[19] and see what happens:

Revised Exclusion: For all distinct properties *M* and *B* such that *M* supervenes on *B*, *M* and *B* are not both *difference-making* causes of a property *A*.

List and Menzies then demonstrate – interestingly and quite surprisingly – that this revised exclusion principle is not in general true.

But what if we also revise (2) and write it in terms of difference-making causes:

(2´) *Revised Completeness:* Every physical occurrence has a physical difference-making cause.

Once again, the right conclusion depends on the contrasts chosen. However, if we keep on concentrating on the default contrasts – and it is unclear what other contrasts could even be meant in a statement as general as this – our key argument above provides a counter-example for this Revised Completeness. That is, however intuitively appealing, Completeness (or "the causal closure of the physical"), when cleaned from confusions, turns out after all to be false.[20]

9. Conclusion

Mental states or events – and more generally, any properties etc. studied by the special sciences which are multiply realizable, and thus can not be identified with properties of the lower physical level – can be as causally relevant as anything can, and be genuine causes of physical events. Does this vindicate the emergentist claim that a higher-level property may have causal powers of its own?

[19] Their formulation of Exclusion as well as of the whole exclusion argument is a bit different from mine. I have not attempted here to formulate their result in my setting, or to make our terminologies commensurable. I merely want to mention their work as it is clearly related to my considerations about Revised Completeness, and indeed inspired my way of presenting the issue here.

[20] As it happens, Menzies (2008) draws the same conclusion.

This depends a lot on how one understands "causal power" and what exactly one means by "having causal powers of its own". And I, for one, find it rather unclear as to what, more precisely, such slogans mean. If it means merely that something is causally relevant, and can be concluded to be a cause, the answer is, in the light of the above arguments, affirmative. If, on the other hand, something stronger and more metaphysical is demanded, it is not at all clear that the claim can be supported. But in any case, perhaps even the former, more modest conclusion is reassuring enough.

Acknowledgements. Earlier versions of this paper have been presented in the "Reduction and the Special Sciences" –conference in Tilburg (April 2008) and in the "Emergence afternoon" at the University of Helsinki (May 2007). I would like to thank all those who participated in the discussions. I am most indebted to Tim Crane, Jaakko Kuorikoski, Peter Menzies and Petri Ylikoski for helpful discussions on the topics of this paper. I am also grateful to two anonymous referees for their useful comments.

References

Armstrong, D. (1968). *A Materialist Theory of the Mind*. London: Routledge.
Bennett, K. (2007). Mental causation. *Philosophy Compass*, 2 (2), 316–337.
Block, N. & Fodor, J. (1972). What psychological states are not. *Philosophical Review*, 81, 159–181.
Cartwright, N. (1979). Causal laws and effective strategies. *Nous*, 13, 419–38.
Chalmers, D. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.
Collingwood, R.G. (1940). *An Essay on Metaphysics*. Oxford: Clarendon Press.
Crane, T. (2001). *Elements of Mind*. Oxford: Oxford University Press.
Craver, C. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Clarendon Press.
Elga, A. (2007). Isolation and folk physics. In H. Price & R. Corry (Eds.), *Causation, Physics, and the Constitution of Reality*. Oxford: Oxford University Press, 106–119.
Field, H. (2003). Causation in a physical world. In M. Loux & D. Zimmerman (Eds.), *Oxford Handbook of Metaphysics*. Oxford: Oxford University Press 2003, 435–60.
Fodor, J. (1968). *Psychological Explanation*. New York: Random House.
Fodor, J. (1974). Special sciences: Or the disunity of science as a working hypothesis. *Synthese*, 28, 97–115.
Gasking, D. (1955). Causation and recipes. *Mind*, 64, 479–487.
Hitchcock, C. (1996). The role of contrast in causal and explanatory claims. *Synthese*, 107, 395–419.
Hitchcock, C. (2007). What Russell got right. In H. Price & R. Corry (Eds.), *Causation, Physics, and the Constitution of Reality*. Oxford: Oxford University Press, 45–65.
Kim, J. (1989). The myth of nonreductive physicalism. Reprinted (1993) in J.Kim, *Supervenience and Mind*. Cambridge: Cambridge University Press, pp. 265–284.
Kuhn, T. (1971). Concepts of cause in the development of physics. Reprinted in T. Kuhn, *Essential Tension*. Chicago: The University of Chicago Press, 21–30.
Latham, N. (1987). Singular causal statements and strict deterministic laws. *Pacific Philosophical Quarterly*, 68, 29–43.
Lewis, D. (1966). An argument for the identity theory. *Journal of Philosophy*, 66, 23–35.
Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy*, 50, 249–258.
List, C. & Menzies, P. (2009). Non-reductive physicalism and the limits of the exclusion principle. (Forthcoming).
Loewer, B. (2001). Review of J. Kim, Mind in a Physical World. *Journal of Philosophy*, 98, 315–324.
Malcolm, N. (1968). The compatibility of mechanism and purpose. *The Philosophical Review*, 78, 468–482.

Menzies, P. (2007). Causation in context. In H. Price & R. Corry (Eds.), *Causation, Physics, and the Constitution of Reality*. Oxford: Oxford University Press, 191–223.
Menzies, P. (2008). Exclusion problem, the determination relation, and contrastive causation. In J. Hohwy & J. Kallestrup (Eds.), *Being Reduced—New Essays on Reduction, Explanation and Causation.* Oxford: Oxford University Press, 196–217.
Menzies, P. (2009). Platitudes and counterexamples. (Forthcoming).
Menzies, P. & Price, H. (1993). Causation as a secondary quality. *British Journal for the Philosophy of Science*, 44, 187–203.
Norton, J. (2007). Causation as a folk science. In H. Price & R. Corry (Eds.), *Causation, Physics, and the Constitution of Reality*. Oxford: Oxford University Press, 11–44.
Papineau, D. (1993). *Philosophical Naturalism*, Blackwell, Oxford.
Papineau, D. (2001). The rise of physicalism. In B. Loewer & C. Gillett (Eds.), *Physicalism and Its Discontents*. Cambridge: Cambridge University Press, 3–36.
Peacocke, C. (1979). *Holistic Explanation.* Oxford: Clarendon Press.
Pearl, J. (2000). *Causality*. New York: Cambridge University Press.
Putnam, H. (1967). Psychological predicates. In W.H. Capitan & D.D. Merrill (Eds.), *Art, Mind, and Religion*. Pittsburgh: University of Pittsburgh Press, 37–48.
Raatikainen, P. (2006). Mental causation, interventions, and constrasts. Unpublished manuscript. Available via: *Online Papers on Consciousness* (complied by David Chalmers and David Bourget): <http://consc.net/online/7.7>.
Raatikainen, P. (2007) Reduktionismi, alaspäinen kausaatio ja emergenssi. *Tiede & Edistys* 4/2007, 284-296 ['Reductionism, downward causation, and emergence'; in Finnish].
Redhead, M. (1990). Explanation. In D. Knowles (Ed.), *Explanation and Its Limits*. Cambridge: Cambridge University Press, 135–154.
Russell, B. (1912-13). On the notion of cause. *Proceedings of the Aristotelian Society,* 13, 1–26.
Smart, J.J.C. (1959). Sensations and brain processes. *Philosophical Review,* 68, 141–156.
Schiffer, S. (1987). *Remnants of Meaning.* Cambridge, MA: Bradford.
Spirtes, P., Glymour, C. & Scheines, R. (2000). *Causation, Prediction, and Search*, 2nd ed. New York: MIT Press.
von Wright, G. H. (1971). *Explanation and Understanding*. Ithica: Cornell University Press.
Woodward, J. (1997). Explanation, invariance, and intervention. In *PSA* 1996, volume 2, 26–41.
Woodward, J. (2000). Explanation and invariance in the special sciences. *British Journal for the Philosophy of Science*, 51, 197–254.
Woodward, J. (2001). Causation and manipulability. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Fall 2001 Edition)*, URL = <http://plato.stanford.edu/archives/fall2001/entries/causation-mani/>.
Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.
Woodward, J. (2004). Counterfactuals and causal explanation. *International Studies in the Philosophy of Science*, 18, 41–72.
Woodward, J. (2007). Causation with a human face. In H. Price & R. Corry (Eds.), *Causation, Physics, and the Constitution of Reality*, Oxford: Oxford University Press, 66–105.
Woodward, J. (2008). Mental causation and neural mechanisms. In J. Hohwy & J. Kallestrup (Eds.), *Being Reduced—New Essays on Reduction, Explanation and Causation.* Oxford: Oxford University Press, 218–262.