Causation, Exclusion, and the Special Sciences **Abridged draft 9.11.2007**

Panu Raatikainen

1. Introduction

Thinkers sympathetic to the autonomy of special sciences typically think, not only that special sciences are not reducible to the fundamental physical level, but also that the properties studied by special sciences have causal powers of their own. More physicalistically inclined philosophers forcefully attack the latter idea. 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 Exclusion Problem

The great majority of philosophers are now convinced that the identity theory of mind cannot be correct. The main reason is "the multiple realizability argument": It seems plausible that a particular mental state can be realized by many states (see e.g. Putnam 1967, Fodor 1968). But how, then, can mental states, if they are not physical, have physical effects, such as behaviour, as a consequence? The problem becomes particularly acute in the form of "the exclusion problem." That is, consider the following five *prima facie* plausible theses:[1]

(1) *Distinctness:* Mental properties are distinct from physical properties.

(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 more than one sufficient cause unless it is overdetermined.

The problem now is that these claims seem to be incompatible. I think it is fair to say that there is no widely accepted and truly convincing solution. In this brief note, I attempt to shed some new light on the problem of mental causation (and more generally, of the downward causation) 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 debate on mental causation

[1] See, e.g., Malcolm (1968), Peacocke (1979), Schiffer (1987) and, in particular, Kim (1989), (1992), (1998), (1999) and (2006). See also Bennett (2007).

and has considerable intrinsic plausibility. Hence, it should be interesting in the present context to see if it can provide any clarification.

3. The Interventionist Theory of Causation

Recently, an 'interventionist' theory of causation has emerged in the philosophy of science, developed especially by James Woodward (1997, 2000, 2003, 2004). 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 (such as the problem of pre-emption).

The interventionist theory of causation has been developed into a sophisticated theory, though 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. Causal claims relate, in this approach, variables, say *X* and *Y*, that can take at least two values. The idea now is that were there an intervention on the value of *X*, this would also be an intervention on 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.

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.[2] 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, so-called "active counterfactuals." These are such that their antecedent is made true by an intervention. They have the form:

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

Now, this is not the right place to try to defend this theory.[3] Suffice it to say that it is in various ways a promising and intuitively attractive theory, and seems to be gaining ground in the philosophy of science. What I want to do in this paper is only to consider the problem of mental causation from the perspective of such a 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 effect peoples' behaviour by manipulating their beliefs and/or desires. Two characteristics of the interventionist approach deserve special attention here: First, it is nowhere required that a cause is in any substantial sense

[2] For an exact definition, see Woodward (2000), (2003).

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

physical. All that is required is that it would make sense in principle to manipulate it. Second, no strict laws are required to subsume the cause and the effect in order for there to be causation (pace Davidson etc.). Nevertheless, these observations do not, as such, answer the exclusion worry. However, I aim to show that the interventionist theory of causation can in fact be helpful in countering the problem.

4. The Argument

Now there is an argument, discovered independently by the present author and Peter Menzies (see Raatikainen 2008, Menzies 2008)[4], which shows that from the interventionist perspective, a mental state can truly be causally relevant, and moreover, that - at least in some ways of conceptualizing the situation - the underlying physical state may fail to be such. I shall outline the argument, and then elaborate some more detailed issues.

I prefer to present the argument with the help of a concrete example: Assume that John desperately wants beer. This is part of our constant background. 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 behaviour? Or is it rather John's brain state (or whatever underlying physical state), call it *B*?

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 beer from the refrigerator. John then gives up the belief that there is beer in the refrigerator. Consequently, 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 below).

In order to evaluate whether we should consider John's belief or his brain state as the cause of his behaviour (going to the refrigerator), 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).

[4] Also Carl Craver (2007, pp. 223-4) briefly sketches a similar argument, giving credit to Eric Marcus. Thus such an argument seems to be very much in the air.

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.[5]

Now 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.

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.

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.

5. Elaboration of the Argument

The above argument certainly deserves, and requires, further elaboration. To begin with, in the interventionist approach, one can distinguish 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 a non-direct cause. The notion of a total cause allows a rather simple interventionist definition,[6] 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.[7] 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 consider explicitly as variables, and which ones are left out

[5] Woodward has in fact certain reservations about the standard Lewis-Stalnaker analysis of counterfactuals. But in the present example, its possible problems appear to be irrelevant. In the interventionist literature, counterfactuals are evaluated instead by systems of equations. In the present case, this approach gives apparently the same results. I have leaned here on the possible world approach because it is more familiar.

[6] (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).

[7] As, for example, in Hesslow'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.

as fixed background conditions. 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.

More importantly, it is common to distinguish *type-level causation* from *token-level causation*. Now the interventionist approach is most directly applicable to type-level causes, and indeed, much of the interventionist literature has focused solely on type-level causation. The cases in which many philosophers (and especially philosophers of mind) have been most interested in are, on the other hand, token-level causes (as our above example about John's belief). Does this undermine our argument? Though this is a reasonable question, the answer is arguably negative. Namely, Woodward (2003, pp. 74-86) shows, in some detail, that the interventionist approach can well be extended to handle also token-level causation. This requires, however, some modifications. Let us take a bit closer look at this issue.

Roughly, instead of possible values of variables, one considers actual values, e.g., ($X = x_1$) and ($Y = y_1$), and takes ($X = x_1$) to be *an actual cause* of ($Y = y_1$) if the following two conditions are satisfied:

(AC1) The actual value of $X = x_1$ and the actual value of $Y = y_1$.

(AC2) There is at least one route R from X to Y for which an intervention on X will change the value of Y, given that all other direct causes Z_i of Y that are not in this route have been fixed at their actual values. (It is assumed that all direct causes of Y that are not on any route from X to Y remain at their actual values under the intervention on X.)

According to such characterization, John's belief, in the above example, can indeed be viewed as an *actual cause* of his behaviour. But how about the underlying physical state, e.g., the brain state B? An intervention on X will not necessarily change the value of Y, because – again – an intervention on X may be such that it changes John's brain state from B to B', and there is no change in Y. So (AC2) is not satisfied. Therefore, having B does not count, in this setting, as an actual cause of John's behaviour.

Above, we assessed active counterfactuals by the familiar possible-world analysis of counterfactuals, because of its familiarity. This is, however, not how they are actually evaluated in the interventionist literature. Rather, the relevant counterfactuals are evaluated by systems of equations. However, this does not change the conclusion. The actual value of $X = x_1$ (the brain state B occurs), and the actual value of $Y = y_1$ (John walks to the kitchen). If then, counterfactually, an intervention were to change the value of X from x_1 to x_2 (the brain state B does not occur), the value of Y might remain the same, if John would go to another brain state B' which nevertheless realizes the belief that there is some beer in the refrigerator.

Finally, of course, one must make sure that the alleged intervention I is indeed a genuine intervention. To begin with, could I, in our simple example (i.e., Peter's hypothetical interference), cause Y directly without going through X? It does not seem so: if Peter's utterance failed to change X, John's belief, John would have still gone to the 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 behaviour changes too (other things being equal). And quite clearly, in the counterfactual 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.

6. The Question of Overdetermination Revisited

It is a very 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 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 which realizes it is much more intimate than between individual shooters of the squad (cf. Loewer 2001, Funkhauser 2001).

Now from the interventionist perspective, 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 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). 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, the key premise of the exclusion argument – (4) No overdetermination – and accordingly, the whole argument, seem to fail to make sense.

7. Conclusion

Mental states or events – and more generally, any properties etc. studied by special sciences which are multiply realizable – can thus be as causally relevant as anything can, and be causes of physical events. Does this vindicate the emergentist claim that a higher-level property may have causal powers of its own? This depends a lot on how one understands "causal power" and what exactly does one mean by "having causal powers of its own". And I, for one, find it rather unclear 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 affirmative. If, on the other hand, something more substantial 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 already interesting and reassuring enough.

References

Bennett, Karen 2007: "Mental Causation", *Philosophy Compass* 2 (2), 316–337.
Block, Ned and Jerry Fodor 1972: 'What Psychological States Are Not'. *Philosophical Review* 81, 159-181.
Collingwood, R.G. 1940: *An Essay on Metaphysics*. Oxford: Clarendon Press.
Crane, Tim 2001: *Elements of Mind*. Oxford: Oxford University Press.
Fodor, Jerry 1968: *Psychological Explanation.* New York: Random House.
Funkhauser, Eric 2002: "Three Varieties of Causal Overdetermination", *Pacific Philosophical Quarterly* 83 (4), 335–351.
Gasking, Douglas 1955: 'Causation and Recipes'. *Mind*, 64, 479-487.
Craver , Carl 2007: *Explaining the Brain*, Oxford: Clarendon Press.
Hesslow, Germund 1976: "Discussion: two notes on the probabilistic approach to causality", *Philosophy of Science* 43, 290-92.
Kim, Jaegwon 1989: 'The Myth of Nonreductive Physicalism', reprinted (1993) in his *Supervenience and Mind*. Cambridge: Cambridge University Press, pp. 265-284.
———. 1992. "'Downward Causation'in Emergentism and Nonreductive Physicalism",in: Beckermann, A., Flohr, H., & Kim, J., (eds.) *Emergence or Reduction?* Berlin: Walter de Gruyter., 119–138.
———. 1998: *Mind in a Physical World*. Cambridge, MA: Bradford.
———. 1999: "Making Sense of Emergence", *Philosophical Studies* 95, 3–36.
———. 2006: "Emergence: Core ideas and issues", *Synthese* 151, 347–354.
Loewer, Barry 2001: "Review of Jaegwon Kim, Mind in a physical world. An essay on the mind-body problem and mental causation", *Journal of Philosophy, 98*, 315–324.
Malcolm, Norman 1968: 'The Compatibility of Mechanism and Purpose'. *The Philosophical Review,* 78, pp. 468-482.
Menzies, Peter 2008: "Exclusion problem, the dermination relation, and contrastive causation", in: Jakob Hohwy & Jesper Kallestrup (eds.), *Being Reduced – New Essays on Reductive Explanation and Special Science Causation*, OUP. Forthcoming.
Menzies, Peter and Price, Huw 1993: 'Causation as a Secondary Quality'. *British Journal for the Philosophy of Science*, 44, pp. 187-203.
Peacocke, Christopher 1979: *Holistic Explanation.* Oxford: Clarendon Press.
Pearl, Judea 2000: *Causality*. New York: Cambridge University Press.
Putnam, Hilary 1967: 'Psychological Predicates', in W.H. Capitan and D.D. Merrill (eds.), *Art, Mind, and Religion*. Pittsburgh: University of Pittsburgh Press, pp. 37-48.
—— 1992: *Renewing Philosophy*. Cambridge, MA: Harvard University Press.
Raatikainen, Panu 2008: "Mental causation, interventions, and constrasts". Forthcoming.
Schiffer, Stephen 1987: *Remnants of Meaning.* Cambridge, MA: Bradford.
Spirtes, Peter, Clark Glymour and Richard Scheines 2000: *Causation, Prediction, and Search*, 2nd ed. New York: MIT Press.
von Wright, Georg Henrik 1971: *Explanation and Understanding.* Ithica: Cornell University Press.
Woodward, James 1997: 'Explanation, Invariance, and Intervention', in *PSA* 1996, volume 2, pp. 26-41.
—— 2000: 'Explanation and Invariance in the Special Sciences'. *British Journal for the Philosophy of Science*, 51, pp. 197-254.
—— 2001: 'Causation and Manipulability', in *The Stanford Encyclopedia of Philosophy (Fall 2001 Edition)*, Edward N. Zalta (ed.), URL =
<http://plato.stanford.edu/archives/fall2001/entries/causation-mani/>.
—— 2003: *Making Things Happen: A Theory of Causal Explanation.* New York: Oxford University Press.
—— 2004: "Counterfactuals and Causal Explanation", *International Studies in the Philosophy of Science* Vol. 18, 41-72.