Not Published October 22, 2006.

Mental Causation, Interventions, and Contrasts

Panu Raatikainen

The problem of mental causation is discussed by taking into account some recent developments in the philosophy of science. The problem is viewed from the perspective of the new interventionist theory of causation developed by Woodward. The import of the idea that causal claims involve contrastive classes in mental causation is also discussed. It is argued that mental causation is much less a problem than it has appeared to be.

1. Introduction

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).

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." That is, consider the following five *prima facie* plausible theses:[1]

(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 overdetermined.

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

The problem is that these claims are apparently incompatible. Philosophers, being what they are, have certainly tried to wriggle out of this dilemma in all possible ways, but I think it is fair to say that there is no generally accepted and truly convincing solution. The predicament is real. Claims (1), (4) and (5) are background assumptions which would be, to my mind, quite implausible to deny. In actuality, the conflict is between (2) and (3). If the argument is valid, one of the two must go. Physicalists typically commit themselves to (2), so, for them, the problem is with (3), and genuine mental causation is an illusion. But this amounts to epiphenomenalism. Others, unwilling to give up the intuitively appealing idea of mental causation, end up denying (2) and accepting 'downward' causation. Yet, it is hard to deny the intuitive appeal of (2), the causal completeness of the physical world.

[1] Variants of the exclusion problem have been presented, e.g., by Malcolm (1968), Peacocke (1979), Schiffer (1987), and Kim (1989). I have borrowed this elegant way of summarizing the exclusion problem from Karen Bennett (2006).

What makes the exclusion problem so difficult is this: It is not that something wrong with 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.

In this brief note, I attempt to shed some new light on the problem of mental causation by taking into account certain advancements in recent theorizing on causation in the philosophy of science. This approach, I hope, will turn out to be useful, given that is acquired certain precedence independently of the debate on mental causation and has some intrinsic plausibility. Hence, it should be interesting in the present context to see if it can provide any clarification.

2. The Interventionist Theory of Causation

Recently, a 'manipulationist' or 'interventionist' theory of causation has emerged in the philosophy of science, 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).[2] This theory is a variant of 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

[2] 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 and reductionistic, and are threatened with circularities. The more recent interventionist variants apparently avoid such problems. Cf. Woodward 2001.

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 be an intervention on *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 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 (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.[3]

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 antecedent is 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.

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

Now, this is not the right place to try to defend this theory. 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' behavior 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.

Two characteristics of the interventionist approach deserve special attention in the present context. First, it is nowhere required that a cause is in any substantial sense physical. 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 strict laws are required 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 mental causation and the exclusion problem. However, I aim to show that the interventionist theory of causation can in fact be helpful in our attempt to answer the exclusion problem.

3. A Case Study

Assume that, at the moment, John desperately wants beer. This is part of our 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), call it *B*, at the moment?

Imagine next, 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 (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 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.

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

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.[4]

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.

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 slightly 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.[5]

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 downward

[4] Woodward has certain reservations about the standard Lewis-Stalnaker analysis of counterfactuals. But in the present example, its undeniable 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.

[5] This argument was inspired by Tim Crane's related considerations with respect to a more traditional counterfactual approach to causation; see Crane 2001, 64-65.

causation. In other words, it seems that the assumption of the causal completeness of the physical world must be, after all, given up. [6]

4. Causes and Contrasts

Behind the idea of the "completeness" of the physical world is apparently the following picture of causation: The physical world has a "built-in" structure; it naturally divides into objective physical events. These stand in causal relation with each other, such that every event has, in normal cases (i.e., when there is no overdetermination), exactly one unique (immediate) sufficient cause. It is a widespread view (e.g., Davidson) that whereas explanations may be interest relative, causal relations certainly are "out there" independently of our interests and our ways of describing them. But is this view correct?

Certain recent developments in the philosophy of science suggest that it is not. In the theory of explanation, it has already become common (beginning from van Fraasen 1980; Garfinkel 1981) to emphasize the role of contrastive classes in explanation. However, it has been now argued, e.g., by Christopher Hitchcock (1996), that causal claims themselves 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. In fact, Woodward's interventionist approach to causality, which relates variables, also incorporates this idea. Similar suggestions have been put forward, with varying conclusions, e.g., by Dretske (1977), Achinstein (1983), Bennett (1988) and Putnam (1992).

For example, consider the following simple causal claim:

> Susan's theft of the bicycle caused her to be arrested.

[6] NOTE ADDED IN MAY 2007: In a talk given at the *Emergence*-conference in Dublin, in April 2007, Peter Menzies presented what amounts to essentially the same argument, though in a bit more abtract form. We have both been quite excited about the fact that we have independently arrived at the same idea.

One can now interpret its contrasts differently. For example:

> Susan's *theft* of the bicycle, rather than her purchase of it, caused her to be arrested.
>
> Susan's theft of the *bicycle*, rather than a car, caused her to be arrested.

It is quite clear that the former is true, whereas the latter is false. Hence, what contrast class is presupposed can be relevant to the truth value of a causal claim.

Now, if causal claims involve such contrastive presuppositions, it is no longer clear how the causal completeness of the physical world should be understood. Considered in isolation, a physical event does not have a determinate cause, whether physical or not. The event only has the cause relative to this or that contrast class. Moreover, some choices of contrast classes may make it very difficult to maintain that the cause must be physical.

Consider again our earlier example of John's desire for beer and his trip to the refrigerator. The causal claim, with the relevant contrasts made explicit, is something like:

> (a) John's belief that the nearest place to get beer is his refrigerator, rather than a belief that it is the grocery or the supermarket (or whatever), *caused* John to go to the refrigerator, rather than to the grocery or the supermarket (or wherever).

John's belief that there is beer in the refrigerator appears to be just the right kind of cause with respect to these contrast classes.

In the case of John's brain state, on the other hand, it is not even clear what the relevant contrast class would be. We do not have in mind any obvious class of alternatives. The

multiple realizability also causes problems because *B* and *B´*, for example, cannot count as genuine alternatives. How can we rule out such mock alternatives without appealing to the fact that they realize the same belief – which seems to make the concept of belief primary?

However, even if such problems can be avoided, the problem of mental causation and the exclusion problem may turn out to be more apparent than real. The causal claim involving brain states (or events), with the relevant contrasts made explicit, would be of the form:

> (b) John's brain state B_1, rather than brain state B_2, or …, *caused* him to go to the refrigerator, rather than to the grocery or the supermarket (or wherever).

One may now notice that the contrast class, at least for the cause, in (b) is different from (a), and, accordingly, that (a) and (b) are not in fact incompatible, even if we assume that John's going to the refrigerator is not causally overdetermined: John's belief is the relevant cause for the contrast class of (a), but perhaps his brain state B_1 may be the relevant cause with respect to the different contrast class of (b).

If the proposed contrastive account of causation is correct, it looks as if the whole problem of mental causation is largely a consequence of a too naive and simple-minded conception of causality. From the perspective taken here, the problem dissolves. Causation may simply be a more interest-relative notion than it has traditionally been taken to be. But this does not mean that once the relevant contrasts have been chosen, the truth-value of a causal claim is not an objective matter. It certainly is. The point is simply that there are different, interest-relative ways of choosing the contrasts.

References

Achinstein, Peter 1983: *The Nature of Explanation.* New York: Oxford University Press.

Bennett, Jonathan 1988: *Events and Their Names*. Indianapolis: Hackett.

Bennett, Karen 2006: 'Mental Causation'. *Philosophy Compass*, forthcoming.

Block, Ned and Jerry Fodor 1972: 'What Psychological States Are Not'. *Philosophical Review*, 81, pp. 159-181.

Collingwood, R.G. 1940: *An Essay on Metaphysics*. Oxford: Clarendon Press.

Crane, Tim 2001: *Elements of Mind*. Oxford: Oxford University Press.

Dretske, Fred 1977: 'Referring to Events'. *Midwest Studies in Philosophy*, 2, pp. 90-99.

Fodor, Jerry 1968: *Psychological Explanation.* New York: Random House.

Garfinkel, Alan 1981: *Forms of Explanation.* New Haven: Yale University Press.

Gasking, Douglas 1955: 'Causation and Recipes'. *Mind*, **64**, pp. 479-487.

Hitchcock, Christopher 1996: 'The Role of Contrast in Causal and Explanatory Claims'. *Synthese*, 107, pp. 395-419.

Kim, Jaegwon 1989: 'The Myth of Nonreductive Physicalism', reprinted (1993) in his *Supervenience and Mind*. Cambridge: Cambridge University Press, pp. 265-284.

Malcolm, Norman 1968: 'The Compatibility of Mechanism and Purpose'. *The Philosophical Review*, 78, pp. 468-482.

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.

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.

.