Published in: K. Talmont-Kaminski and M. Milkowski (eds.) *Regarding the Mind, Naturally: Naturalist Approaches to the Sciences of the Mental* (Cambridge Scholars Publishing).

CAN THE MENTAL BE CAUSALLY EFFICACIOUS?

PANU RAATIKAINEN

1 Introduction

A key problem in contemporary philosophy of mind is the conflict between two viewpoints: On the one hand, there seems to be a convincing argument against strong physicalism, i.e. the view that mental states or properties can be identified with certain physical states or properties (e.g. brain states); this is the argument from multiple realizability. On the other hand, there is a certain line of reasoning, based on causal considerations, which seems to show that mental properties must be identical with physical properties: this is the exclusion argument. This tension, and how to best resolve it, is the topic of this paper.

2 Undermining Strong Physicalism: Multiple Realizability

There is an influential argument which seems to undermine the (type) identity theory, the so-called multiple realizability argument: it is suggested that a particular mental kind can be realized by many distinct physical kinds (see e.g. Putnam 1967, Fodor 1968, 1974, Block and Fodor 1972), and therefore, cannot be identified with any particular physical kind.

Putnam first compared mental states to computer programs (Putnam 1960, 1967). Programs can be described e.g. with the help of flow charts. A particular program, say one that multiplies the given (numerical) input with itself, can however be realized in indefinitely many different ways. The early computers used electronic tubes, the later ones transistors, and the contemporary ones use microchips. They could be realized mechanically, for example with the help of cogwheels (like Babbage's planned 19th Century 'Difference engine' and 'Analytic engine'), or utilizing hydraulics, and so on. Such realizations are radically different for their physical properties, and nevertheless it is possible to view them as realizing, in a given situation, the same program.

The program is something much more abstract and general than any particular physical mechanism that realizes is. In more recent terms, Putnam suggested that the relation of a mental state to the physical state that underlies it in a given situation is analogous to the relation between software and hardware: in both cases the former cannot be identified with the latter; the latter realizes the former, but so could many alternative physical realizers. This is the thesis of multiple realizability.

Or, moving to genuine mental states, consider, for example, *pain* (cf. Putnam 1967). Admittedly described in simplifying terribly simplifying terms,[1] pain is the mental state that is caused by tissue damage, and typically brings about wincing, moaning and avoidance behavior. It seems plausible that various different animals – humans, primates, other mammals, perhaps even birds and reptiles – are all capable of experiencing pain. However, it is also clear that these different animals and their brains must have radically different physical-chemical build-up. Therefore, so the argument goes, it would be a mistake to identify the property of being in pain with any particular underlying physical-chemical property, for the latter must vary greatly between different species.

Further, it seems conceivable that, in addition to humans, there could exist extraterrestrial humanoids, and perhaps also androids, who also had conceptually structured representations of their environments and intentions to act, something like our beliefs and desires – even if their physiology was not even carbon-based but rather, say, silicon-based, or whatever. Again, it seems plausible that the essence of a particular belief

[1] In particular, we ignore here completely the qualitative, or phenomenal, aspects of pain. Because of these very aspects, it is not really likely that pain could be given an exhaustive functional analysis along these lines. Such an analysis is much more plausible for propositional attitudes such as belief and desire; these are, on the other hand, more holistic in nature and it is difficult to give any snappy, illuminating example based on them. Therefore, I only use this simplified example.

or desire comes down to its causal-functional role – its appropriate relation to perceptual input, other mental states, and action – and that a human being and an alien or a robot might have (in the relevant sense) the same belief or the same desire, despite the fact that their respective physical-chemical states would be radically different. Again, this speaks against the identification of that particular mental state with any specific physical state.

The argument from multiple realizability has convinced many to reject strong physicalism (reductionism, or the type-identity theory). It is one main reason for the popularity of weaker, non-reductive forms of physicalism. The argument has not, though, remained unchallenged.[2] And of course, those who want to use the exclusion argument to defend strong physicalism, must somehow circumvent it.

One very popular response among those sympathetic to stronger physicalism, suggested first by David Lewis (1969) and advocated by Churchland (1986) and Kim (1999), for example, is to grant the possibility of multiple realizability, but to add that it does not rule out reductions of a more local kind, for example reductions relative to species, domain, or structure. That is, it is suggested that it is consistent with the thesis of multiple realizability that human-pain, for example, is reducible to one neuroscientific kind, while elephant-pain reduces to another one neuroscientific kind, and so on.

It can be argued, though, that this reply cannot really save the identity theory. First, it has been proposed that in addition to the kind of multiple realizability discussed above, where the realizing physical properties differ between different species, there may occur a much more radical type of multiple realizability, namely, cases in which the underlying physical state which realizes a certain mental state may be different even in one and the same individual at different times (see e.g. Block & Fodor 1972, Block 1978, Horgan 1993).

Already Block and Fodor (1972) suggested that there is actually even some empirical evidence for such more radical multiple realizability, and referred briefly e.g. to brain plasticity. Endicott (1993) has reviewed such empirical evidence more extensively. Finally, Barrett (forthcoming) discusses in detail a case which apparently provides an example of such radical multiple realizability in empirical psychology.

[2] For different critical responses, see e.g. Funkhauser 2007; Bickle 2008, and references given therein. Kim 1992b, Bickle 1998, Bechtel & Mundale 1999, Shapiro 2000 and Polger 2004, in particular deserve to be mentioned.

In the case of such a massive kind of multiple realizability, the idea of local reduction leads to absurdity (cf. Horgan 2001). At least one key advocate of new wave reductionism, John Bickle, grants this much: "The more radical type of multiple realizability seems to force increasingly narrower domains for reductions to be relativized – at the extreme, to individuals at times. This much 'local reduction' seems inconsistent with the assumed generality of science" (Bickle 2008).

Second, independently of such a radical multiple realizability, relativizing reductions to species would rule out all interspecies psychological generalizations. And arguably such generalizations are commonplace in psychological explanation (cf. Horgan 2001, Pereboom & Kornblith 1991). Consider the following imaginary example:

> If *A* wants to get to the Andromeda Galaxy, and believes that the only way to get there is with the help of teleportation, A probably takes measures to get teleported there.

The generalization talks about desire and belief in general, not about desire-in-human or desire-in-robot – not to mention, not about desire-in-Kirk, or belief-in-Spock, or belief-in-Data.

This simplified but not altogether implausible generalization talks about desires and beliefs *simpliciter*, and not about belief-relative-to-this-or-that-species. And it is easy to imagine that there could be more complicated (and more truth-like) generalizations of this sort. (There are apparently other, more complicated actual examples of psychological properties, not based on any rationality assumptions, and which may occur in interspecies generalizations: e.g. vision seems to be multiply realized by vastly different eye structures; see Weiskopf 2011)

In sum, the thesis of multiple realizability and the conclusion that because of it, strong physicalism must be given up, is still very much alive. Typically, though, the conclusion is not a dualism of any sort, but only a weaker, non-reductive type of physicalism: it is agreed that there is no separate non-physical substance, and that the mental at least supervenes (see Section 9) on the physical. Be that as it may, the validity of the multiple realizability thesis is not our main concern here. My approach in this paper may instead be seen as conditional: If the type-identity theory were not true (whether because of the multiple realizability or of some other reason), would it then follow (because of the exclusion argument) that the mental cannot be causally efficacious? I aim to argue that it does not: the exclusion argument does not force us to choose between the type-identity theory and epiphenomenalism. This obviously leaves open the

ultimate status of the multiple realizability thesis and non-reductive physicalism.

3 Motivating Strong Physicalism: Causal Exclusion

The popularity of strong physicalism among contemporary philosophers is not, as Papineau (2001) clearly demonstrates, just a result of arbitrary fashion: it has rather been characteristically motivated by a certain line of reasoning, which is based on the apparently plausible assumption of "the causal closure of the physical realm" (see below) and the worry that the mental, if it is not physical, would end up being causally epiphenomenal, that is, causally impotent. That is troublesome enough: it would show that we've been enormously wrong in our view of the mind. But if one moreover sympathized the so-called Alexander's dictum, "To exist is to have causal powers" (see e.g. Kim 1992a), the conclusion would be that mental properties and states are not even real. In any case, it is unclear why we should postulate the existence of anything like that, something we cannot really observe, if its presence did not make any difference.

"The causal closure of the physical" means here the assumption that all physical effects are due to physical causes. The argument is then, roughly, that anything that has a physical effect it must itself also be physical. Thus, if the mental is capable of causing physical effects, it must itself be physical. Today the argument is widely called "the exclusion argument" (also known as "the overdetermination argument", or "the causal argument"[3]).

The pioneers of the mind-body identity theory such as Feigl (1958) and Smart (1959) thus proposed that we should identify mental states with brain states, for otherwise those mental states would be "nomological danglers", something that lie outside causal laws and the causal structure of reality, and in particular play no role in the causal explanation of behavior.[4] As Smart memorably put it, mental states are "nothing over and above brain processes".

Armstrong (1968) and Lewis (1966, 1972) argued, first, that mental states are picked out by their causal roles, and, further, that we know that physical states play these roles, and concluded that mental states must be

[3] Kim calls a certain variant of it "the supervenience argument" – in its common form, however, the argument does not lean on supervenience.

[4] The talk of "nomological" (relating to laws) suggests that they still presupposed the regularity view of causation. The worry is, though, more general and independent of this view, as we'll see soon.

identical with those physical states. Lewis (1966) made the assumption of causal closure explicit (under the label "the explanatory adequacy of physics").

Usually, though, the exclusion argument proper is credited to Malcolm (1968), who himself used it to argue that reasons could *not* possibly be causes of action! (It is the question of which premise to give up). Later, Peacocke (1979) and Schiffer (1987), for example, have used the argument. More recently, Kim (1989, 1992a, 1998, 2005) and Papineau (1993, 2001) in particular have pressed the exclusion argument in defense of strong physicalism, and this line of reasoning seems to enjoy some amount of popularity.

In sum, it is fair to say the exclusion argument, or something like it, is essential for contemporary physicalism.

4 The Causal Exclusion Argument

Let us look a bit closer at the exclusion argument. Assume first that strong physicalism, or the mind-body identity theory, is false:

Assumption (*distinctness*):
Mental properties are distinct from physical properties.

However, the following premises are – so the argument goes – apparently indisputable:

To begin with, unless something like the following holds, the physical reality would be mysteriously gappy:

Premise 1 (*the causal closure of the physical*):
Every physical occurrence has a sufficient physical cause.

In other words, as we trace back the causal history of any physical effect, there will never be a need to appeal to anything non-physical. Note that this thesis does not in itself amount to strong physicalism, or the identity theory, but is compatible even with some versions of dualism. It only states that, whatever else, non-physical entities or properties there may exist, the physical realm is "causally closed". Even an opponent of physicalism, such as David Chalmers (see Chalmers 1996, p. 150), may find this principle plausible and even undeniable.

Then again, it seems to be both a common-sense truism, and something that much of psychology presupposes, that mental states such as beliefs and desires bring about bodily behavior. That is:

Premise 2 (*causal efficacy*):
Mental events sometimes cause physical events, and sometimes do so by virtue of their mental properties.

Could not both a mental state and the underlying physical state (e.g. the brain state) be the cause of certain behavior? Now philosophers have reflected on peculiar cases in which an event has more than one cause. In such a case, the event is said to be "overdetermined" by its causes. A standard example is a death caused by several members of a firing squad shooting simultaneously. However, there is wide agreement that such cases of overdetermination are relatively rare coincidences, and that behavioral events cannot regularly be overdetermined in this way:

Premise 3 (*no universal overdetermination*):

The physical effects of mental causes are not all overdetermined.

Add the obvious-looking principle of exclusion:

Premise 4 (*exclusion*):
No effect has more than one sufficient cause unless it is overdetermined;

and the assumption and the premises are arguably inconsistent. Therefore, reductive physicalists conclude, the assumption must be rejected.[5]

Conclusion:
Mental properties must be identical to physical properties.

Let us focus on a concrete example: Assume that John, at time *t*, is sitting in the living room, desires beer and believes the there is some beer in the refrigerator. This is his relevant total mental state *M*. This is followed by the bodily behavior *B*: John walks to the kitchen (to get a beer

[5] One could, alternatively, conclude that mental properties do not cause any physical effects, as e.g. Malcolm did, but most contemporary philosophers find this option unattractive.

from the refrigerator). Obviously, John is also, at time *t*, in a certain physical (neurological state, brain-state, or physical-chemical state, or a micro-physical state[6]) *P*.

The question now is, whether John's mental state *M* can be viewed as the cause of his consequent behavior? Now inasmuch as John's bodily behavior is viewed as a physical occurrence, it seems – because of the causal closure of the physical realm – that it has a purely physical preceding cause; say, the brain state of John at *t*. Now if the latter is the sufficient cause of the behavior (at *t*), and if the effect is not – this seems implausible – overdetermined, it seems that the physical state excludes the mental state as a cause.

As Bennett (2007) emphasizes, the exclusion problem is different from many other problems about mental causation, which claim that the mental is somehow *by its nature* unsuited to cause anything. Rather, the problem is, in a sense, in the physical realm: given that every physical event already has a sufficient physical cause, there is no room for the mental to cause (without overdetermination) anything physical, even if the mental was in principle adequate to work as a cause here. And it is this feature that makes the exclusion problem so difficult.

5 Some Popular Responses (and Their Problems)

Philosophers sympathetic to non-reductive views have certainly attempted to reply to the exclusion argument.

Some have defended the autonomy of the mental by referring to the explanatory practice of empirical science. It has been pointed out that actual explanations in psychology, and in the higher-level special sciences in general, often proceed without any reference to lower-level physical concepts and explanations; such explanations may moreover appear to be causal (see e.g. Baker 1993). The overall observation about actual explanations in the special sciences is worth making, but without some further analysis, it provides little to dissolve the puzzlement caused by the exclusion problem.

Some philosophers differentiate causation and explanation, and propose that although 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.

[6] I'll leave it open what exactly is meant by "physical" here. It would be, though, fair to press reductive physicalists about this.

Jackson and Pettit (1988, 1990), for example, make a distinction between "causal efficacy", which is causation in the full-blooded sense, and the weaker "causal relevance", which applies to higher-level special sciences and their explanations, and is (in Loewer's (2002) words) mere "causation lite". Genuine causation, efficacy, occurs only at the fundamental physical level. Higher-level states and properties may nevertheless be used in explanations, and are in this sense (causally) relevant.

However, such ''solutions'' are not particularly attractive. Would it not be preferable to have an account on which explanations explain by citing genuine causes? And in any case, is this not just epiphenomenalism in disguise? In effect, such responses amount to denying that the mental could truly be causally efficacious. Surely what we really are interested in, and would like to have, is the view that mental properties have causal efficacy in the same standard robust sense that everything else has (cf. the "homogeneity assumption" of Crane 1995).

A particularly popular response in defense of non-reductive physicalism is the view known as Compatibilism. It grants that one effect can have both a metal and a physical cause – they are compatible with each other – without this being a case of overdetermination, in the standard sense of the word. (Shoemaker 2001, Pereboom 2002, Bennett 2003). The advocates of this approach all press in their different ways that the relation between the mental and physical causes of an effect is much more intimate than in the paradigmatic cases of overdetermination, e.g. the relation between two shooters in a firing squad. Even if not type-identical, in any particular situation, the underlying physical state realizes the mental state at stake, and moreover determines it, i.e., the latter supervenes on the former, and so on.

Let us focus on Shoemaker's approach, for it seems to be the most exact and perhaps now also the most popular of these. To begin with, Shoemaker subscribes to the following general idea:[7]

The Causal Inheritance Principle: if mental property M is realized in a system at time t in virtue of physical realization base P, the causal powers of M are identical with, or are a subset of, the causal powers of P.

[7] As it happens, Kim has also sometimes (1992b, 1998) leaned on this idea, although his purpose for it has obviously been quite different: he has aimed to use it to argue *against* multiple realizability!

Now Shoemaker submits that a subset of causal powers simply cannot be excluded, in the spirit of the exclusion argument, by the whole set: causal powers of P cannot compete with and exclude the causal powers of M, because they are (at least in part) just the same causal powers. Therefore, we can stop worrying about the exclusion problem (but see Gillett and Rives 2005).

Now there is certainly something in the general idea, shared by the compatibilists, that the relation between the mental and physical causes of a behavioral effect is much more intimate than in the standard cases of overdetermination (indeed, our second argument below can be seen as a more exact elaboration of this idea; see Section 9).

However, our first argument below (Sect. 8) shows that, in a sense, the causal inheritance principle is not even true: assuming the multiple realizability, the causal profile of the supervening mental property is simply different from that of the underlying physical property; the causal powers of the former are neither identical with, nor a subset of, the causal powers of the latter. Consequently, compatibilism grounded on this principle is also untenable.

6 The Relevance of the Theory of Causation

The exclusion argument has been discussed intensively, but for a long time, there has not been much convergence in the views of the philosophers. Though many want to resist the radical reductionist conclusion, responses have varied greatly (see Bennett 2007; Robb & Heil 2009). However, most often the attempted solutions are based only on vague intuitions about causation and not on any explicit, well-developed theory of causation; or, they lean on some arguably outdated and problematic views on causation.

Sometimes, on the other hand, it is suggested that the whole exclusion problem is redundant and that all that matters is the choice of the theory of causation: it is submitted that given the dependence view of causation, mental causation is no problem at all, whereas if the production view of causation is assumed, mental and other higher-level causation is immediately impossible even without any exclusion argument. (Loewer 2002, for example, seems to think along these lines). Bennett disagrees: "But while I certainly agree that the production view [of causation] is often in the background of discussions of the problem … I do not agree that the problem itself actually requires it. I do not agree that rejecting it makes the issue go away." (Bennett 2008). Kim admits that "Loewer is right… in saying that my thinking about causation and mental causation

involves a conception of causation as 'production' or 'generation'" (2002, 675). Bennett, however, thinks it is "wrong to assume that the pure dependence notion alone would dissolve the problem completely" (*ibid*.)

It is indeed old news (see e.g. LePore & Loewer 1987), that from the perspective of the dependence notion of causation (e.g. counterfactual approach), mental states or properties seem perfectly suitable for causing, for example, bodily behaviour, and from this perspective, indeed seem often to do so. On the one hand, I agree with Bennett that this observation does not, in itself, make the exclusion problem go away, and it does not really tell us where exactly the exclusion argument then goes wrong: that is, it does not provide us with any analysis or diagnosis of the problem. On the other hand, I submit, *pace* Bennett, that an extended analysis based on the best current theory of causation *can* illuminate the problem, and indeed effectively dissolve the problem. The resolution does not, though, fall out trivially from the theory, but requires a little elaboration. This is what we'll do in this paper.

7 Theories of Causation

Let us take a quick look at the different theories of causation. It is now commonplace to divide theories of causation into two broad categories, first, to theories that view causation as *dependence*, or as *difference-making*, and second, to theories according to which causation is some kind of *production*, or *transmission*. According to the former, causes are difference-makers for their effects, in the sense that the cause makes a difference to whether or not the effect occurs. It includes the regularity view (often associated with Hume), and various counterfactual approaches. More developed examples of the latter idea are Salmon's mark transmission account (1984) and Dowe's conserved quantity account (1992, 2000)

The Regularity Theory of Causation

For a long time, the received view on causation, especially among the empiristically-minded philosophers, was the regularity theory: *c* causes *e* if and only if all events of type *C* (i.e., events that are like *c*) are regularly followed by events of type *E* (i.e., events like *e*).

This approach has been criticized for long, and problems have cumulated. The fundamental problem is that it just cannot distinguish truly

causally relevant properties from accidental correlations. For example: A sudden drop in the reading of a barometer is regularly succeeded by the occurrence of a storm. However, it does not follow – *pace* the regularity theory – that the barometric reading caused the storm; rather, a drop in atmospheric pressure caused both the barometric reading and the storm.

Salmon (1971) gave an amusing and vivid counter-example: Start with the fact that John Jones, a male, fails to get pregnant. In the example, John Jones, for some strange reason, has regularly taken birth control pills for an entire year. Finally, it is a fact that:

> All males who take birth control pills regularly fail to get pregnant.

Nevertheless, it would be absurd to consider the use of birth control pills as the cause of John Jones's not getting pregnant.

For such reasons, the regularity theory has generally faded from philosophical currency.

The Counterfactual Theory

Another popular approach in the dependence or difference-making group is the counterfactual approach. Counterfactual considerations can easily solve, for example, the above counter-examples to the regularity theory. The basic idea of counterfactual theories of causation is that the meaning of causal claims can be explained in terms of counterfactual conditionals of the form:

> "If A had not occurred, C would not have occurred".

The best known counterfactual analysis of causation is David Lewis's (1973, 1986, 2000) theory. In terms of counterfactuals, Lewis first defines a notion of causal dependence between events, and then causation in terms of chains of such causal dependence.

Nevertheless, vivid philosophical debate over four decades has made it doubtful whether any theory along these lines could work. Difficulties with so-called "preemption" and "trumping" have proven to be insurmountable problems for Lewis's theory. Lewis attempted to revise his theory to handle them (Lewis 2000, 2004), but it has remained questionable whether even his new theory can really deal adequately with all cases of pre-emption and trumping (see Menzies (2009).

Causation as Transmission

The basic idea of the causation-as-transmission or causation-as-production view, in contrast, is that causation involves objects becoming into contact and exchanging or transmitting something. Although the general intuitive idea is again classic, the first well-developed theory of causation of this sort was Salmon's 1984 Mark Transmission (MT) account. Salmon suggested that we need to change the conceptual apparatus: instead of taking distinct events or facts to be the causal *relata*, Salmon thinks we should try to characterize directly when a process is causal. Salmon proposed that a process is causal if it is capable of *transmitting a mark.* Otherwise, we have a non-causal process, or, a pseudo-process.

However, Kitcher (1989) and others soon argued that Salmon's definition excluded some genuine causal processes, and allowed some clearly non-causal processes. Moreover, important for Salmon was to treat causation as an empirical phenomenon, and this involved avoiding any appeal to counterfactuals. When Salmon realized that his MT theory in fact makes tacit appeal to counterfactual relations, he abandoned it (Salmon 1994), and adopted Phil Dowe's Conserved Quantity approach (Dowe 1992). In this theory, a causal process is defined as a process that transmits non-zero amount of a conserved quantity (for example mass, energy, momentum, charge) at each moment in its history.

However, Dowe's theory has likewise serious difficulties in identifying the causally relevant quantities. As Hitchcock (1995) points out, often, in causal interactions, several conserved quantities are exchanged. For example: A pool cue strikes a cue ball, imparting both momentum and a blue dot of chalk. In the former case, momentum is exchanged. In the latter, matter is exchanged. Yet only the first is relevant to the trajectory of the cue ball. It is unclear how to determine which exchanges are relevant. A natural response would be to rely, again, on counterfactuals. If we had removed the dot, or had changed it from blue to red, for example, the trajectory of the ball would have been the same. But this solution is not open for Salmon and Dowe, who are trying to avoid any use of counterfactuals. Ironically, as Hitchcock notes, counterexamples formulated earlier by Salmon himself against the regularity theory – e.g., John Jones taking birth control pills – can be turned against the Salmon-Dowe approach.

Salmon acknowledged such problems and admitted that they are severe problems for his approach. Dowe has attempted to address these worries in his 2000 book. Hausman (2002) and Ehring (2003), for example, have in

turn presented severe critique against it, and it seems that the majority of philosophers remain skeptical about the success of his replies. Furthermore, for Dowe's approach to apply, it should be possible to translate the causal claims of the special sciences, and even of lots of common physics, to the language of fundamental physics. And it is highly doubtful that this would even be possible for, say, biology, not to mention history, economics or psychology. For these and other reasons, the conserved quantity approach is no longer popular.

More generally, there is not available a single defensible, well-developed theory of causation in the causation-as-transmission or causation-as-production group. The whole tradition is arguably bankrupt. It seems we must look elsewhere.

Interlude: Causation and Contrast-relativity

All the above theories of causation have the further problem that they fail to acknowledge an important character of causation and causal claims: it has now become 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; cf. Dretske 1977; Achinstein 1983; Woodward 1984; Bennett 1988; Lipton 1990; Putnam 1992.

For example, consider the following simple causal claim:

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

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.

The Interventionist Theory of Causation

Recently, a 'manipulationist' or 'interventionist' theory of causation has emerged in the philosophy of science, and it is becoming increasingly popular as a theory of causation. 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). This theory is 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. Furthermore, the interventionist theory also embodies the idea that causal claims are essentially contrastive.

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? Why do we care to distinguish between causal and merely correlational relationships? (cf. Woodward 2003, p. 28) According to 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. Thus, we can try to find a cure for AIDS, or suppress poverty, 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 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 (the theory does not aim to give a reductive analysis of causation, so this is not a problem).

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.

This theory is very promising and attractive, and it seems to be quickly gaining ground as the most popular theory of causation in philosophy. It is also our point of departure in what follows.

Let us now proceed to our main arguments.

8 The First (Proportionality) Argument from Interventionism

Now there is an argument, discovered independently at least by myself (see Raatikainen 2006, 2007, 2010) and Peter Menzies (Menzies 2008; cf. List & Menzies 2009; Menzies & List 2010),[8] which shows that from the interventionist perspective, a mental state can be a genuine cause of a bodily behaviour; and moreover, that – at least in some ways of conceptualizing the situation[9] – the underlying physical state may well fail to be the cause.[10]

Recall our earlier example of John who desires beer and believes that there is some beer in the refrigerator (John's relevant mental state *M*), and consequently walks to the kitchen (the bodily behavior *B*). At the same time, John is in a certain physical state *P*. Let us focus on the belief, and assume that the desire for beer is, for the relevant period, an unchanging

[8] Also Carl Craver (2007, pp. 223-4) briefly sketches what seems to amount to the same argument, giving credit to Eric Marcus (unpublished). In addition, Woodward (2008) puts forward similar ideas (though he is somewhat less unequivocal about his conclusions). Thus, such an argument has been very much in the air.

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

[10] Yablo's (1992) earlier response, which emphasizes that a cause must be *proportional* to its effect, bears some resemblance to our approach, and can be viewed as a predecessor of it, even if it is based more on an essentialist metaphysical view than our argument.

background condition. Let us assume that if John did not have that belief, he would instead go to the closest grocery to buy some beer.

Let us denote the cause variable by X, and the effect variable by Y, and let us suppose, for simplicity, that the following cases exhaust all possible cases:

($X = x_1$): John has the belief that there is some beer in the refrigerator

($X = x_2$): John does not have the belief that there is some beer in the refrigerator

($Y = y_1$): John goes to the refrigerator

($Y = y_2$): John goes to the grocery

Further, we must consider some counterfactual intervention *I* which would change the value of X from x_1 to x_2 (i.e., change John from having the belief to not having it): for example, imagine that Peter, John's roommate, informs John that he has actually drunk all John's beers in the refrigerator; John then gives up the belief that there is beer in the refrigerator. Accordingly, John, instead of going to the refrigerator, leaves for the closest grocery.

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. 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

[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 a purely verbal choice on my part—nothing really hinges on this choice.

intervention, e.g. Peter's hypothetical interference, which changes the value of X, brings about a change in the value of Y.

Next, let us focus on John's underlying physical state *P* (at the same time *t*). Surely it counts as the cause of John's behavior *B*? In fact, this depends vitally on how we set the contrasts and choose the relevant variables.[12] We may let the variable Z (for the alleged cause) to range over a number of different possible, mutually exclusive physical states (brain states, or whatsoever) of John, including *P* above (i.e. the brain state which in this particular case 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 physical state *P*. 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. However, this is still a rather weak conclusion, and should by no means be thought of as suggesting that X and Z are somehow in competition here, in the spirit of the exclusion argument, or that the effect (John's bodily behavior) would be overdetermined.

Situations where several variables are causally relevant in this way to an effect variable are very common. 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 has put 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).

Matters get more interesting, if we focus on the natural "default contrast", as we have already done in the case of belief: in that case, the alternative values of Z would be just ''John has the physical state P'' ($Z = z_1$) and ''John does not have the physical state P'' ($Z = z_2$). This way of choosing the contrast makes the two cases (belief/underlying physical state) also more directly commensurable.

In order to evaluate whether we should now consider John's belief or his physical state (or both) as the cause of his behavior (going to the

[12] Neither Menzies, Woodward nor originally myself (in the earlier papers) sufficiently emphasized the importance of the choice of contrasts here, but we all simply assumed that it is the default contrast that is in action here. This may have resulted in some misunderstanding. Its significance was only explicitly emphasized in (Raatikainen 2010). In personal correspondence, Menzies has agreed on its relevance.

refrigerator), we need to, according to the interventionist approach, consider the following two counterfactuals:

> (1) If John's belief that there is beer in the refrigerator had been 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 physical state *P* had been changed by an intervention to not having that state, he would have gone to the grocery (and not to the refrigerator).

It is quite clear that (1) emerges as true. 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 P', one that is different from P, which can also realize the belief that there is some beer in the refrigerator.

Hence, it is possible that an intervention changes John's brain state from P to P', and John nevertheless goes to the refrigerator and not to the grocery. Hence, (2) apparently comes out as false. And consequently, if we hang onto the default contrast, the variable Z (whether John has P 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:

> (3) John's having the belief (that there is beer in the refrigerator) caused him to go to refrigerator,

or, formally:

> (3') $X = x_1$ causes $Y = y_1$,

is true if and only if:

(i) it is actually the case that $X = x_1$ and $Y = y_1$; and:

(ii) if an intervention were to change the value of X from x_1 to x_2, the value of Y would change from y_1 to y_2 (which amounts to the 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 the 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:

(4) John's having the physical state *P* (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 *P* 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.

Of course, the occurrence of *P* is surely sufficient for the effect, John's behavior, but that does not make it (relative to all natural contrasts) the cause of the latter. Being sufficient condition for the occurrence of something, and being its difference-making cause, must thus be clearly distinguished.

Note that the above argument also rebuts the Causal Inheritance Principle: it is not true that the causal powers of a multiply realizable mental state are the same as, or a subset of, the causal powers of the underlying physical state that realizes it. The causal profiles of these two are simply different.

9 The Second (Supervenience) Argument from Interventionism

There is also another interventionism-based argument, also developed independently at least by myself (Raatikainen 2007, 2010), Shapiro and Sober (2007), and Woodward (2008). Its point of departure is the popular supervenience assumption.

A set of properties *A supervenes* upon another set *B* just in case no two things can differ with respect to *A*-properties without also differing with

respect to their *B*-properties.[13] Both non-reductive and reductive physicalists typically believe that everything – and the mental, in particular – supervenes on the physical as a matter of metaphysical necessity – that the physical facts determine all possible higher-level facts, with metaphysical necessity.

As Bennett (2008) nicely puts it: "Physicalists think that mental events and properties are not truly distinct existences that can be snipped away from their physical bases; the connecting laws simply are not breakable. There is no room for any wedge. That is why the metaphysical necessity of the supervenience claim is of crucial importance to their view."

Now it is essential for the exclusion argument to reflect on whether a mental state M and the physical state P realizing it overdetermine the behavioral effect B or not. From the interventionist perspective, this requires that we consider a causal system which includes a variable for both M and P (and their alternatives). 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 paradigmatic firing squad case of overdetermination).

But if the supervenience thesis is true, that is, it is metaphysically necessary that the facts of the physical level determine the mental level, this is simply impossible, and consequently, the question of overdetermination does not even make sense in this context. And this gives us another independent reason for doubting the whole exclusion argument.

More exactly, let us consider again our example of John. The question whether there is overdetermination involved here requires one to evaluate whether the following three counterfactual conditionals hold or not:

(i) If John had not had the physical state P, but had had the mental state M, then he would have still gone to the refrigerator.

(ii) If John had not had the mental state M, but had had the physical state P, then he would have still gone to the refrigerator.

[13] Philosophers have distinguished several different kinds (e.g. local/global) of supervenience; see e.g. (McLaughlin and Bennett, 2011). Here we focus only on the general idea and ignore the fine details.

(iii) If John had not had either the physical state P or the mental state M, then he would have gone to the grocery.

The effect is, according to the interventionist theory of causation, overdetermined if and only if all these three claims are true.

Now (i) and (iii) are apparently true, but given the supervenience thesis, the antecedent of (ii) does not make sense: it is simply not possible to vary the realized mental state, and simultaneously hold the realizing physical state constant. Consequently, it is no more correct to say that there is no overdetermination involved here than that the effect is overdetermined. And this undermines the whole exclusion argument.

This argument agrees, to some extent, with the intuitive idea behind the compatibilist response, i.e. that 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, or in other paradigmatic examples of overdetermination, but it makes the point exact with the help of a well-developed theory of causation.

10. Objections to the Interventionist Arguments

The above arguments have received quite a lot of attention. Quite often the reception has been enthusiastic, but they have also received some criticism. Let us consider a couple of noticeable objections.

To begin with, Marras and Yli-Vakkuri (2010) argue against our first argument. Connecting their critique to our concrete example involving the thirsty John, they seem to assume that our argument goes as follows:

> Let the variable Z range over a number of different possible, mutually exclusive physical states (brain states, or whatsoever) of John, including *P* above (i.e. the brain state which in this particular case 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 physical state *P*.
>
> However, granting the multiple realizability, there are other possible values of Z different from z_1 (say, z_m, for example) which also realize the same belief. Consequently, not all changes in the

> value of Z result a change in the value of Y (John's behavior). Hence, Z is not causally relevant for Y.

This argument presupposes that, according to the interventionist approach, X is causally relevant for Y if and only if *any* change in the value of X, due to an intervention, results in a change in the value of Y. And as Marras and Yli-Vakkuri correctly point out, this is not in accordance with what the interventionist theory actually says: rather, the condition is this:

> X is causally relevant for Y if and only if *some* changes in the value of X, due to an intervention, bring about a change in the value of Y.

So if the above was our argument, it would indeed be correct to protest and not accept it. But it is not. Rather, as I have tried to emphasize above, the argument only focuses on the default contrast, on whether having P rather than not having P makes a difference, and concludes that it does not. The argument is limited to this setting. With some other contrasts, e.g. if a large number of alternative physical states (represented by z_1, z_2, z_3, z_4, …) are in the chosen contrast class, the situation is admittedly different (see above). It may have been that our early formulations of the argument (Raatikainen 2006, 2007; Menzies 2008; Woodward 2008) were insufficiently clear on all this – simply silently assumed the default contrast – and this has misled people. But all this was discussed explicitly in (Raatikainen 2010) and should be clear by now.

Then again, Baumgartner (2009, 2010) argues that the interventionist theory of causation, far from dissolving the exclusion problem and establishing the causal efficacy of the mental, in fact entails a variant of the exclusion argument. In a nutshell, Baumgartner argues as follows: First, he focuses on the fact that according to interventionism, for X to be causally relevant for Y, it must be possible to change the value of X *independently* of all other variables in the causal system – without changing the value of them – except, obviously, of Y. He then takes it for granted that the physical state P, and the variable Z covering it, is causally relevant for the effect (the bodily behavior B, and the variable Y). Finally, Baumgartner argues that because of *supervenience*, it is impossible to change the variable X (covering the mental state M) independently of the variable Z, and consequently that the mental state M cannot cause the behavior B.

To make the long story short, the fundamental flaw in this argument is that on the one hand, it presupposes the metaphysical supervenience assumption, and on the other hand, it presupposes that it is possible to have a causal system which has a variable for both the supervening mental state or property, and for the realizing lower-level physical state or property. And this, as we have seen in Section 9, is impossible. The combination of X, Z and Y together simply is not a causal system in the standard interventionist sense. (Woodward (forthcoming) discusses this issue and its many ramifications in considerable detail.)

In sum, the interventionist response to the exclusion argument is still defensible, and the existing objections to it can be satisfactorily replied.

11 The Causal Closure Again

Consider now again the two premises of the exclusion argument, namely, The Causal Closure of the Physical, and Exclusion:

(2) The Causal Closure of the Physical: Every physical occurrence has a sufficient physical cause.

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

Note now that from the point of view of our earlier arguments, both these assumptions involve confusing causes with sufficient conditions. There are causes, which are difference-makers; and there are sufficient conditions, which are an entirely different issue and not necessarily causes of any sort. The talk of "sufficient causes" in the exclusion argument conflates these two different things. Hence, I do not think that these two assumptions, as they are formulated in the exclusion argument, are so much false (or true) as mongrels based on a conceptual confusion which fail to make clear sense. Recall that 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.

But what if we revise the premises and write them in terms of difference-making causes? Let us focus on the Causal Closure, because it is likely the more interesting one philosophically. As we have emphasized,

in order to make a causal judgment unambiguous, it is obligatory to fix some contrast class or another.

One natural formulation uses "default contrasts":

1. The causal closure with the default contrast

> If a physical event P_1 has a cause at time *t*,
> then there is a *physical* event R_1 at time *t*
> such that R_1 (rather than not-R_1) causes P1 (rather than not-P_1)

However, our first (proportionality) argument (Section 8) demonstrates that formulated in this way, the principle is false. It is certainly possible to fix the contrast classes differently. Indeed, the following formulation seems to be defensible:

2. *A weaker form of the causal closure*

> If a physical event P_1 has a cause at time *t*,
> then for some contrast class $\{P_2, ..., P_n\}$ for P_1,
> there is at time *t* a physical event R_1 and some contrast class $\{R_1, ..., R_m\}$ for it,
> such that R_1 (rather than $R_1, ..., R_m$) causes P_1 (rather than $P_2, ..., P_n$).

Formulated in this way, the causal closure principle – though possibly true – does not support the exclusion argument. But perhaps the inability to clearly recognize the difference between such distinct forms of the principle explains, in part, why so many philosophers have felt that the principle is, even in its stronger forms, beyond dispute.

References

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

Baker, L R. 1993. "Metaphysics and Mental Causation." In *Mental Causation*, ed. John Heil and A Mele, 75–96. Oxford: Clarendon Press.

Barrett, David. 2012. "Multiple Realizability, Identity Theory, and the Gradual Reorganization Principle." *British Journal for the Philosophy of Science* (Forthcoming).

Baumgartner, Michael. 2010. "Interventionism and Epiphenomenalism." *Canadian Journal of Philosophy* 40 (3): 359–383.

Baumgartner, Michael. 2009. “Interventionist Causal Exclusion and Non-reductive Physicalism.” *International Studies in the Philosophy of Science* 23 (2): 161–178.
Bechtel, William and Jennifer Mundale (1999). “Multiple Realizability Revisited: Linking Cognitive and Neural States.” *Philosophy of Science*, 66: 175-207.
Bennett, Karen. 2003. “Why the Exclusion Problem Seems Intractable, and How, Just Maybe, to Tract It.” *Nous* 37 (3): 471–497.
Bennett, Karen. 2007. “Mental Causation.” *Philosophy Compass* 2 (2): 316–337.
Bennett, Karen. 2008. Exclusion again. In Jakob Hohwy & Jesper Kallestrup (eds.), *Being Reduced: New Essays on Reduction, Explanation, and Causation.* Oxford University Press.
Bennett, Jonathan. 1988. *Events and Their Names.* Indianapolis: Hackett.
Bickle, John (1998). *Psychoneural Reduction: The New Wave.* Cambridge, MA: MIT Press.
Bickle, John. 2008. “Multiple Realizability”, *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2008/entries/multiple-realizability/>.
Block, Ned. 1978. “Troubles With Functionalism”, in: C.W. Savage (ed.), *Perception and Cognition: Issues in the Foundations of Psychology*, Minnesota Studies in the Philosophy of Science, vol. 9. Minneapolis: University of Minnesota Press, 261–325.
Block, Ned, and Jerry A. Fodor. 1972. “What Psychological States Are Not.” *The Philosophical Review* 81 (2): 159–181.
Chalmers, D. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.
Churchland, Patricia. 1986. *Neurophilosophy: Toward a Unified Science of the Mind/Brain.* Cambridge, MA: MIT Press.
Crane, Tim. 1995. “The Mental Causation Debate”, *Proceedings of the Aristotelian Society*, Supplementary Vol. 69: 211-36.
Craver, C. 2007. *Explaining the Brain.* Clarendon Press.
Dowe, P. 1992. “Wesley Salmon's Process Theory of Causality and the Conserved Quantity Theory.” *Philosophy of Science* 59: 195-216.
Dowe, P. 2000. *Physical Causation.* New York: Cambridge University Press, 2000.
Dretske, Fred. 1977. ‘Referring to Events’. *Midwest Studies in Philosophy*, 2, 90-99.
Endicott, R. 1993. “Species-Specific Properties and More Narrow Reductive Strategies”, *Erkenntnis* 38, 303–321.
Feigl, H. 1958. ‘The “Mental” and the “Physical”’. In *Concepts, Theories and the Mind-Body Problem* (Minnesota Studies in the Philosophy of Science, Volume 2), ed. H. Feigl, M. Scriven and G. Maxwell, 370-497. Minneapolis: University of Minnesota Press.
Fodor, Jerry. 1968. *Psychological Explanation.* New York: Random House.
Fodor, Jerry A. 1974. “Special Sciences (or: The Disunity of Science as a Working Hypothesis).” *Synthese* 28 (2): 97–115.

Funkhouser, Eric. 2007. “Multiple Realizability”. *Philosophy Compass* 2/2 (2007): 303–315.
Gillett, Carl, and Bradley Rives. 2005. “The Non-Existence of Determinables: Or, a World of Absolute Determinates as Default Hypothesis.” *Nous* 39 (3): 483–504.
Hausman, Daniel M. 2002. “Physical Causation.” *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 33 (4): 717–724.
Hitchcock, Christopher Read. 1996. “The Role of Contrast in Causal and Explanatory Claims.” *Synthese* 107 (3): 395–419.
Hitchcock, Christopher Read. 1995. “Salmon on Explanatory Relevance.” *Philosophy of Science* 62 (2): 304–320.
Horgan, Terence. 2001. “Multiple Reference, Multiple Realization, and the Reduction of Mind.” In *Reality and Humean Supervenience: Essays on the Philosophy of David Lewis*, ed. Gerhard Preyer and Frank Siebelt. Landham, MD: Rowman and Littlefield: 205-221.
Horgan, Terence. 1993. “Nonreductive Materialism and the Explanatory Autonomy of Psychology.” In *Naturalism: A Critical Appraisal*, ed. Steven Wagner and Richard Warner. Notre Dame, IN: University of Notre Dame Press: 295-320.
Kim, Jaegwon. 1989. “The Myth of Nonreductive Materialism.” *Proceedings and Addresses of the American Philosophical Association* 63 (3): 31–47.
Kim, J. 1992a. “‘Downward causation’ in emergentism and nonreductive physicalism.” In *Emergence or reduction?*, ed. A. Beckermann et al., 19–138. Berlin: Walter de Gruyter:
Kim, Jaegwon. 1992b. “Multiple Realization and the Metaphysics of Reduction.” *Philosophy and Phenomenological Research* 52 (1): 1–26.
Kim, Jaegwon. 1998. *Mind in a Physical World*. Cambridge, MA: MIT Press.
Kim, Jaegwon. 2002. “Responses.” *Philosophy and Phenomenological Research* 65: 671-680.
Kim, Jaegwon. 2005. *Physicalism, or Something Near Enough*. Princeton, NJ: Princeton University Press.
Kitcher, Philip. 1989. “Explanatory Unification and the Causal Structure of the World.” In *Scientific Explanation. Minnesota Studies in the Philosophy of Science Volume XIII*, ed. Philip Kitcher and Wesley C. Salmon: 410–505. Minneapolis: University of Minnesota Press.
Lewis, David. 1966. “An Argument for the Identity Theory.” *The Journal of Philosophy* 63 (1): 17–25.
Lewis, David. 1969. “Review of Art, Mind, and Religion.” *The Journal of Philosophy* 66 (1): 22–27.
Lewis, David. 1972. “Psychophysical and Theoretical Identifications.” *Australasian Journal of Philosophy* 50 (3): 249–258.
Lewis, David. 1973. “Causation.” *The Journal of Philosophy* 70 (17): 556–567.
Lewis, David. 2000. “Causation as Influence.” *The Journal of Philosophy* 97 (4): 182–197.

Lewis, D. 2004. “Causation as Influence”, in Collins, Hall, and Paul (2004), pp. 75–106. Lipton, Peter. 1991. *Inference to the Best Explanation*. London: Routledge.

List, Christian, and Peter Menzies. 2009. “Nonreductive Physicalism and the Limits of the Exclusion Principle.” *The Journal of Philosophy* CVI (9): 182–197.

Loewer, Barry. 2002. “Comments on Jaegwon Kim’s Mind and the Physical World.” *Philosophy and Phenomenological Research* 65 (3): 655–662.

Malcolm, Norman. 1968. “The Conceivability of Mechanism.” *The Philosophical Review* 77 (1): 45–72.

Marras, A. & Yli-Vakkuri, J. 2010. “Causal and explanatory autonomy.” In *Emergence in Mind*, ed. G. and C. Macdonald: 129–138, Oxford: Oxford University Press.

Menzies, Peter. 2008. Exclusion problem, the determination relation, and contrastive causation. In *Being Reduced—New Essays on Reduction, Explanation and Causation*, ed. J. Hohwy and J. Kallestrup: 196–217.Oxford: Oxford University Press.

Menzies, Peter. 2009. "Counterfactual Theories of Causation", *The Stanford Encyclopedia of Philosophy (Fall 2009 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2009/entries/causation-counterfactual/>.

Menzies, Peter & List, C. 2010. “The causal autonomy of the special sciences.” In *Emergence in Mind*, ed. G. and C. Macdonald: 108–128. Oxford: Oxford University Press.

Papineau, David. 1993. *Philosophical Naturalism*. Blackwell.

Papineau, David. 2001. The rise of physicalism. In *Physicalism and Its Discontents*, ed. B. Loewer and C. Gillett: 3–36. Cambridge: Cambridge University Press.

Peacocke, C. 1979. *Holistic explanation.* Oxford: Clarendon Press.

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

Polger, Thomas (2004). *Natural Minds.* Cambridge, MA: MIT Press.

Putnam, Hilary. 1960. “Minds and Machines”. In *Dimensions of Mind*, ed. S. Hook, 148-80. New York: University of New York Press.

Putnam, Hilary. 1967. “Psychological predicates.” In *Art, mind, and religion*, ed. W. H. Capitan and D. D. Merrill: 37–48. Pittsburgh: University of Pittsburgh Press.

Putnam, Hilary. 1992. *Renewing Philosophy.* Cambridge, MA: Harvard University Press.

Raatikainen, Panu. 2010. “Causation, Exclusion, and the Special Sciences.” *Erkenntnis* 73 (3): 349–363.

Raatikainen, Panu. 2007. “Reduktionismi, alaspäinen kausaatio ja emergenssi.” *Tiede & Edistys* 4/2007. [’Reductionism, downward causation, and emergence’, in Finnish.]

Raatikainen, Panu. 2006. “Mental causation, interventions, and contrasts.” Unpublished manuscript. Available via: *Online Papers on Consciousness* (complied by D. Chalmers and D. Bourget): <http://consc.net/online/7.7>.
Robb, David and Heil, John. 2009. “Mental Causation”. In *The Stanford Encyclopedia of Philosophy (Summer 2009 Edition)*, ed. Edward N. Zalta, URL = <http://plato.stanford.edu/archives/sum2009/entries/mental-causation/>.
Salmon, W. 1984. *Scientific Explanation and the Causal Structure of the World.* Princeton: Princeton University Press.
Salmon, Wesley C. 1994. “Causality Without Counterfactuals.” *Philosophy of Science* 61 (2): 297–312.
Schaffer, Jonathan. 2000. “Trumping Preemption.” *The Journal of Philosophy* 97 (4): 165–181.
Schiffer, Stephen. 1987. *Remnants of Meaning.* Cambridge, MA: MIT Press.
Shapiro, Lawrence. 2000. “Multiple Realizations.” *Journal of Philosophy*, 97: 635-654.
Shapiro, Larry & Sober, Eliot. 2007. “Epiphenomenalism – The Dos and Dont’s.” In *Thinking about Causes*, ed. Peter Machamer and Gereon Wolters, Pittsburgh: University of Pittsburgh Press, 235–264.
Shoemaker, Sydney. 2001. “Realization and Mental Causation.” In *Physicalism and Its Discontents*, ed. Carl Gillett and Barry M Loewer, 74–98. Cambridge: Cambridge University Press.
Smart, J. J. C. 1959. “Sensations and Brain Processes.” *The Philosophical Review* 68 (2): 141–156.
Spirtes, Peter, Clark Glymour and Richard Scheines. 2000. *Causation, Prediction, and Search*, 2nd ed. New York: MIT Press.
Weiskopf, D. A. 2011. “The Functional Unity of Special Science Kinds.” *The British Journal for the Philosophy of Science* 62: 233–258.
Woodward, James. 1984. “A Theory of Singular Causal Explanation.” *Erkenntnis* 21 (3): 231–262.
Woodward, J. 2000. “Explanation and Invariance in the Special Sciences.” *The British Journal for the Philosophy of Science* 51 (2): 197–254.
Woodward, James. 2003. *Making Things Happen*. Oxford: Oxford University Press.
Woodward, James. 2008. “Mental causation and neural mechanisms.” In *Being Reduced—New Essays on Reduction, Explanation and Causation*, ed. J. Hohwy and J. Kallestrup, 218–262. Oxford: Oxford University Press.
Woodward, J. 2012. “Interventionism and Causal Exclusion.” Forthcoming.
Yablo, Stephen. 1992. “Mental Causation.” *The Philosophical Review* 101 (2): 245–280.