ISSN 2045-5577

ORDER: GOD'S, MAN'S AND NATURE'S

The Causal Autonomy of the Special Sciences

Peter Menzies and Christian List[1]

1. Introduction

The systems studied in the special sciences are often said to be causally autonomous, in the sense that their higher-level properties have causal powers that are independent of those of their more basic physical properties. This view was espoused by the British emergentists, who claimed that systems achieving a certain level of organizational complexity have distinctive causal powers that emerge from their constituent elements but do not derive from them.[2] More recently, non-reductive physicalists have espoused a similar view about the causal autonomy of special-science properties. They argue that since these properties can typically have multiple physical realizations, they are not identical to physical properties, and further they possess causal powers that differ from those of their physical realizers.[3]

Despite the orthodoxy of this view, it is hard to find a clear exposition of its meaning or a defence of it in terms of a well-motivated account of causation. In this paper, we aim to address this gap in the literature by clarifying what is implied by the doctrine of the causal autonomy of special-science properties and by defending the doctrine using a prominent theory of causation from the philosophy of science.

The theory of causation we employ is a simplified version of an "interventionist" theory advanced by James Woodward (2003, forthcoming a, b), according to which a cause makes a counterfactual difference to its effects. In terms of this theory, it is possible to show that a special-science property can make a difference to some effect while the physical property that realizes it does not. Although other philosophers have also used counterfactual analyses of causation to argue for the causal autonomy of special-science properties,[4] the theory of causation we employ is able to establish this with an unprecedented level of precision. It permits us to identify necessary and sufficient conditions for the causal autonomy of a higher-level property, and to show that these are satisfied when causal claims about higher-level properties have a special feature we call *realization-insensitivity*. This feature consists in the fact that the relevant claims are true regardless of the way the higher-level properties they describe are physically realized. Our findings are consistent with those of other philosophers, for example Alan Garfinkel (1981), who have noted the realization-insensitivity of higher-level causal relations as a distinctive feature of the special sciences and have suggested that this feature ensures their independence from lower-level causal relations.

---

[1] This paper presents an application of earlier technical results in List and Menzies (forthcoming).

[2] For a very clear statement of this view see Broad (1925); and for historical background see Kim (1992) and McLaughlin (1992).

[3] An important statement of the non-reductive physicalist position can be found in Fodor (1974).

[4] See, for example, Crane 2001, Horgan 1989, Le Pore and Loewer 1987. In a recent unpublished paper, Raatikainen (2006) has independently developed a similar analysis of mental causation in terms of Woodward's interventionist theory of causation.

Our discussion proceeds as follows. In section 2, we clarify what it means to say that the causal powers of special-science properties are independent of those of their underlying physical properties. In section 3, we present a simplified version of the theory of causation as counterfactual difference-making. In section 4, we employ this theory to specify the conditions under which an instance of a higher-level, special-science property can have causal powers not possessed by the instance of the physical property that realizes it. In section 5, we compare our results with Garfinkel's discussion of the indispensability of higher-level causal explanations in the special sciences, and argue that his insights can be systematized in our framework.

We discuss the causal autonomy of the special sciences in the context of non-reductive physicalism rather than British emergentism. The emergentists' views are difficult to interpret because of their unfamiliar terminology and philosophical preoccupations; non-reductive physicalism is a more familiar framework for our purposes. To be sure, this framework has been criticized by several philosophers, most notably Jaegwon Kim (1998, 2005), who has argued that its commitment to the causal efficacy of higher-level properties makes it inherently unstable. A defence of non-reductive physicalism against these arguments is not the topic of this paper. We have discussed it elsewhere (List and Menzies forthcoming; Menzies forthcoming), and the technical results stated in section 4 below draw on this work.

2. Causal Autonomy

Non-reductive physicalists believe that even if the higher-level properties of special-science systems are not identical to lower-level physical properties, they nonetheless supervene on them, meaning that there can be no difference in a special-science property without an accompanying difference in some physical property.[5] For convenience, we focus on a concrete instance of such a supervenience claim, namely the relationship of mental or psychological properties to physical properties, though the morals of this paper can be generalized to other special-science properties. The non-reductive physicalist about the mind maintains that mental properties are not identical to physical properties – notably because of multiple realizability – but they nonetheless supervene on physical properties. The relevant physical properties vary from one supervenience claim to another. In the case of mental properties, they are usually taken to be neurophysiological properties.[6] Accordingly, the non-reductive physicalist holds that any two individuals, actual or possible, who are duplicates with respect to their neurophysiological properties will be duplicates with respect to their mental properties. When an individual instantiates a particular mental property by virtue of instantiating a subvenient physical property, it is customary to say that the instance of the physical property *realizes* the instance of the mental property.[7]

[5] More precisely, the supervenience thesis should be understood as a contingent, global supervenience thesis: any world that is a minimal physical duplicate of the actual world is a duplicate with respect to special-science properties and relations. A *minimal physical duplicate* of the actual world is a world that has the actual world's physical entities, physical properties, and laws, *and nothing else*.

[6] Intentional mental properties are often thought to have wide content in the sense that their content implicates the existence of objects and properties outside the skin of the subjects of those properties. If this is so, intentional properties do not supervene on neurophysiological properties. Here it is necessary to bracket the issue about wide content because of limitations of space. Thus we focus on non-intentional properties and on intentional properties whose contents can be specified, perhaps somewhat artificially, in a narrow way.

[7] Throughout this paper we understand an *instance of a property* to consist in an object's instantiating the property at a certain time. (Usually, the reference to time will be tacit.) We also refer to a property-instance as a *state*. We understand the identity conditions of property-instances or states to be such that

What does it mean to say that special-science properties, in particular mental properties, have causal powers that are independent of those of their physical realizers? To answer this question, we must clarify two terminological issues. First, in discussing a property's *causal powers*, one might discuss its forward-looking powers to cause certain states, or its backward-looking powers to be caused by certain states, or a combination of both. We focus on forward-looking powers, as they are most relevant to our present concerns. Secondly, in talking about the causal powers of properties, one might refer to *properties* or *property-instances*. Which is appropriate depends on whether one is discussing type-causation or token-causation. Since we are concerned with token-causation, we focus on the causal powers of property-instances or states. We say that the state Fa, an instance of the property F, has the *forward-looking causal power* to produce the state Gb, an instance of the property G, just in case Fa can in suitable conditions cause Gb. The causal powers of properties can then be understood in terms of generalizations about the causal powers of their instances.[8]

What is involved in affirming or denying the claim that certain special-science properties, say mental properties, are autonomous and independent of those of their physical realizers? It is easier to begin with what is involved in the denial of such claims. The following thesis, we believe, captures the view of many philosophers who deny the causal autonomy of mental properties:

> *The Physical Determination of the Causal Powers of Mental States*: For all mental properties M and physical properties P, if an instance of property M is realized by an instance of property P, then the causal powers of the M-instance are a *subset* of the causal powers of the P-instance.

The formulation of this thesis in terms of subsets allows for the special case in which the causal powers of the mental state are identical to those of its realizing physical state. So, for example, the instances of M and P may have the same causal powers to produce in identical conditions the behavioural effects B_1 and B_2. But equally, the causal powers of the instance of M may be a proper subset of those of the corresponding instance of P: perhaps the instance of M has the power to produce B_1 and B_2 under certain conditions, while the instance of P has the power to produce B_1, B_2 and B_3 under the same conditions.

Now the assertion of the causal autonomy of mental properties is best viewed in terms of the denial of this determination thesis. The autonomy thesis can thus be stated as follows:

> *The Causal Autonomy of Mental States*: For some mental property M and physical property P, where an instance of property M is realized by an instance of property P, the causal powers of the M-instance are *not* a subset of those of the P-instance.

If the thesis is true, as we seek to show, some mental states have causal powers that are not causal powers of their realizing physical states. This claim is controversial and denied by many philosophers. Jaegwon Kim (1998), for example, affirms a version of the Physical Determination of Mental States for functionally defined mental properties. His Causal Inheritance Principle states that if a mental state is functionally defined in terms of its causal role then its causal powers must be identical with those of the physical state that realizes that causal role. Similarly, Sydney Shoemaker

---

two property-instances or states are identical only if the corresponding properties are identical. We do not construe property-instances as tropes or abstract particulars.

[8] For example, the statement that the property F has the forward-looking power to cause G can be understood as meaning that instances of F can in suitable conditions cause instances of G.

(2001) argues for something like the Physical Determination of Mental States in connection with functionally defined properties. Indeed, he defines the notion of realization in terms of this thesis. He writes: "In general, then, property X realizes property Y just in case the conditional powers bestowed by Y are a subset of the conditional powers bestowed by X… Where the realized property is multiply realizable, the conditional powers bestowed by it will be a proper subset of the sets bestowed by each of the realizer properties" (2001: 78-9). Although Kim's and Shoemaker's theses are restricted to functionally defined properties, our arguments below refute not only the unrestricted Physical Determination thesis but also the restricted ones accepted by Kim and Shoemaker.

3. Difference-Making Causation

Philosophical discussions of the causal autonomy of the special sciences often invoke intuitive principles about causation. Since intuitions can be misleading, especially when chosen selectively, a better procedure is to base the discussion on a fully developed, well-motivated theory of causation. For this purpose, we turn to the interventionist account of causation developed recently by James Woodward (2003, forthcoming a, b). This theory has gained increasing support from philosophers of science as providing an instructive account of causal concepts in science. More generally, the interventionist framework forms the basis of a productive research programme for studying causation in philosophy, computer science and psychology.[9]

Many theories of causation link the concept of causation with that of making a difference. Woodward's interventionist theory falls within this tradition. On his theory, the causal relata are variables, and causation relates changes in one variable to those in another. The simplest version of the theory states that variable X *causes* variable Y just in case if the value of X were to change as a result of an intervention, then the value of Y would change too.[10] Although Woodward presents this simple definition as an account of type-causation, we shall also use it for analysing token-causation, setting aside those situations that require Woodward's more involved account of token-causation. When applied to token-causation, the theory uses variables whose values represent the occurrence or non-occurrence of an event, or the instantiation or non-instantiation of a property by an object at a time.

It is crucial that the changes in the causally related variables occur by virtue of a hypothetical, if not actual, intervention on the cause variable. Changes in one variable may be accidentally correlated with changes in another without any causal relation between them. For example, decreases in barometer readings are correlated with onsets of storms, but this correlation is due to these phenomena being the effects of a common cause – drops in atmospheric pressure. However, if the changes in the barometer reading were brought about by an intervention, say by an experimenter fixing the reading of the barometer, the correlation with the onset of a storm would disappear. This is the central difference between correlations and causal relations: a genuine causal relation is robust to interventions that change the values of the cause variable.

Woodward's interventionist theory bears some resemblance to manipulability theories that state that a causal relationship exists when a human agent can bring about one event by manipulating another. A crucial difference, however, lies in Woodward's definition of an intervention. Very roughly, an *intervention* on one

[9] For other works that employ this interventionist framework, see Gopnik and Schulz 2007, Hitchcock 2001, Pearl 2000, and Spirtes, Glymour, and Scheines 1993.

[10] This simple version of the theory assumes causation to be deterministic and so does not cover probabilistic causation. It is also not intended to cover more complicated cases of pre-emption and overdetermination. Moreover, Woodward defines two causal concepts: the concept of a total cause and of a direct cause. The two concepts coincide in the simple cases we discuss here.

variable X with respect to another variable Y is an idealized experimental manipulation that causes X to change its value in such a way that all other variables that previously were causally relevant to X no longer influence it; and in such a way that any change in Y can only come about through the change in X. Thus an intervention could be the result not only of a human action, but also of a "natural experiment".[11]

Further, on the interventionist theory, causal claims have an implicit contrastive structure, which can be made explicit using "rather than" constructions. So the standard form of a causal claim might be represented: the contrast between X's taking the value x rather than x' causes the contrast between Y's taking the value y rather than y'. When the variables involved are many-valued, it can be indeterminate which of the possible values of X and Y are being contrasted with their actual values, but since our focus is on cases involving binary variables, the relevant contrast is always clear.[12] We suggest that in the binary case the difference-making condition for causation can be adequately expressed in terms of a pair of counterfactuals, where x, x' and y, y' are the possible values of the variables X and Y, respectively:

> *Truth conditions for difference-making causation*: X = x makes a difference to Y = y if and only if (a) $X = x \mathrel{\Box\!\!\rightarrow} Y = y$; and (b) $X = x' \mathrel{\Box\!\!\rightarrow} Y = y'$.

These counterfactuals must be understood according to an interventionist, non-backtracking reading. A backtracking counterfactual involves reasoning from an outcome to earlier events and then forwards again, as, for example, when one reasons that if the barometer reading were low, then this would mean that the atmospheric pressure is low, which in turn would mean that the storm is going to occur. Evaluated as a backtracking conditional, the counterfactual "If the barometer reading were low, then the storm would occur" is true. By contrast, a non-backtracking counterfactual is evaluated by supposing that its antecedent is made true by an intervention, which breaks any existing relationship between the antecedent and its causes. When the barometer reading is set to some value through an intervention, one cannot infer back from this value to the value that the atmospheric pressure must have had, and thus the counterfactual "If the barometer reading were low, then the storm would occur" is false under the non-backtracking reading.

Generally, the counterfactuals are to be understood according to the standard possible-worlds semantics, developed by Lewis (1973), which defines their truth conditions in terms of a similarity relation between possible worlds. The similarity relation is represented by an assignment to each possible world w of a system of spheres of worlds centred on w, subject to standard constraints.[13] The idea is that the

---

[11] This notion of an intervention does not burden the theory with the anthropocentric implications of manipulability theories. For instance, the theory implies that a causal relation can exist between two variables independently of whether any human agent does or could carry out an intervention on the variables. Another difference from orthodox manipulability theories is that it does not attempt to reduce causal concepts to non-causal ones. The notion of an intervention is defined in terms of causal concepts, which means that the definition of causation in terms of interventions is non-reductive. Woodward argues persuasively that the definition is nonetheless informative in virtue of having many non-trivial implications regarding causal relationships.

[12] To be sure, this is an artificial restriction because many causal variables, even in everyday life, are best seen as many-valued. But the simplifying assumption that causal statements involve just binary variables makes the questions we discuss more easily tractable. More generally, for X to be a cause of Y, there must exist at least two different values of X, x and x', and two different values of Y, y and y', such that under an intervention that changes X from x to x', the value of Y changes from y to y'.

[13] *Nestedness*: For any two spheres S and T, either S is included in T or T is included in S. *Weak centring*: w is contained in every sphere. *Exhaustiveness*: There is a largest sphere containing all relevant possible worlds. *Limit assumption*: For any world w and any proposition P, there is a smallest

smaller a sphere is around w, the more similar to w are the worlds in it. Now a counterfactual conditional P $\Box\!\!\rightarrow$ Q is true in world w if and only if Q is true in all the closest P-worlds to w. Figure 1 shows a situation in which the counterfactual P $\Box\!\!\rightarrow$ Q is true in the world w at the centre of the system of spheres. The set of P-worlds is represented by the region with diagonal lines, the set of Q-worlds by the larger region that includes the set of P-worlds.

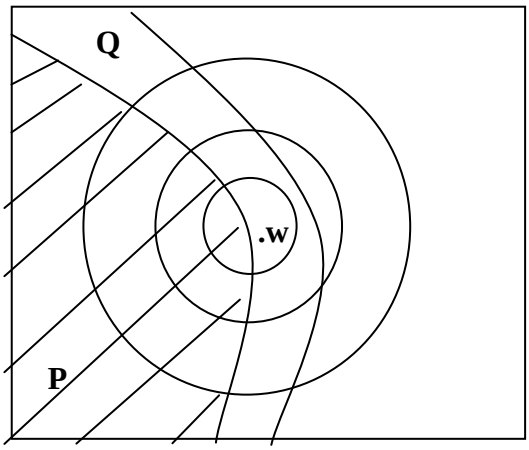


Figure 1

As discussed in more detail in List and Menzies (forthcoming), our semantic framework diverges from Lewis's in one respect: we adopt a weaker centring requirement than Lewis by allowing the smallest sphere around w to contain more than one world. Lewis, by contrast, requires it to contain only w. A relaxation of Lewis's strong centring requirement is essential for our purposes. Strong centring implies that whenever P and Q are true in some world so is P $\Box\!\!\rightarrow$ Q. So, whenever X=x and Y=y are true in the actual world, clause (a) of our truth-conditions for difference-making causation would automatically hold as well, which would trivialize this condition. For the difference-making account of causation to work, clause (a) must be non-trivial: it must rule out insufficiently specific candidate causes. In particular, the counterfactual formulation must allow that even if the candidate cause and effect are both exemplified in the actual world, the smallest sphere around it also contains some other worlds exemplifying the candidate cause.

4. Proportional Causation and Realization-Insensitive Causation

In this section we employ the difference-making account of causation to determine whether the thesis of the Physical Determination of the Causal Powers of Mental States is true, drawing on earlier work in List and Menzies (forthcoming). Recall that the thesis states that if an instance of a mental property M is realized by an instance of a physical property P, then the causal powers of the M-instance are either identical to those of the P-instance or a proper subset of it. If this thesis is true, then its negation, the thesis of the Causal Autonomy of Mental States, must be false. Correspondingly, if the first thesis is false, then the second must be true.

Many philosophers of mind believe that the Physical Determination thesis is true. Their acceptance of the thesis, we think, stems from a more general preference for lower-level, physical causal variables over higher-level, special-science variables: it is the physical properties and states that do all the causal work, it is assumed, and

---

sphere around w containing some P-world, called the *smallest* P-*permitting sphere around* w. The *closest* P-*worlds to* w are defined as the P-worlds within the smallest P-permitting sphere around w.

properties and states supervening on them derive whatever causal efficacy they have from the underlying physical ones.[14] Moreover, these philosophers seem to assume that the Physical Determination thesis is an *a priori* truth. Kim (2005), for example, says it is an analytic truth. But this is mistaken, as we show in this section. If it is a fact about the world that every mental state derives its causal powers from its realizing physical properties then it is an empirical fact about the world. Notwithstanding this point, *a priori* conceptual knowledge can shed light on this issue. Knowing what the concept of causation entails, we are in a better position to understand the precise meaning of the Physical Determination thesis, and so in a better position to determine whether it is true or false in the light of our empirical knowledge about the world.

So we propose to interpret the Physical Determination thesis in terms of the difference-making account of causation. It is convenient to evaluate the thesis by considering a schematic example. Suppose one of us – say Peter – has an intention to signal a taxi (an instance of mental property M) and he waves his arm (an instance of behavioural property B); and suppose that his intention (the M-instance) is realized by some neural state (an instance of neural property N_1), but it could also have been realized by other neural states (say, instances of the neural properties N_2, …, N_n). What is the cause of Peter's waving his arm – the mental state Ma (where "a" refers to Peter) or the neural one N_1a? The difference-making account of causation permits it to be the case that the mental state Ma, and not the neural state N_1a, is the cause of the behavioural outcome Ba. Thus it may be the case that both counterfactuals (1a) and (1b) are true, but not both counterfactuals (2a) and (2b) are true:

(1a) Ma $\Box\!\!\rightarrow$ Ba.
(1b) ~Ma $\Box\!\!\rightarrow$ ~Ba.

(2a) N_1a $\Box\!\!\rightarrow$ Ba.
(2b) ~N_1a $\Box\!\!\rightarrow$ ~Ba.

This situation might be described by saying that Peter's waving his arm rather than not waving it was caused by his having the mental property M rather than not having it, and not by his having the neural property N_1 rather than not having it.

It is not surprising that this situation could obtain. There are many common situations in which a supervenient state has causal powers not possessed by the subvenient state that realizes it. Consider a familiar example from the philosophy of action. Imagine you have feuded with an irascible neighbour for some time but you decide to break the ice by greeting him. Your neighbour is startled by your saying "Hello" unexpectedly. As it happens, your greeting is rather abrupt and loud. But your neighbour is startled, not because you say "Hello" loudly, but because you simply say it. Here the relationship between your saying "Hello" and saying "Hello" loudly is analogous to the relationship between Peter's having the intention to wave his arm and his being in neural state N_1: the first state of each pair supervenes on the second. However, it is the supervenient state, not the subvenient one, that does the causal work. The difference in the states' causal status is reflected in terms of the difference in the truth values of the following pairs of counterfactuals:

(3a) You say "Hello" $\Box\!\!\rightarrow$ your neighbour startles.

[14] A striking manifestation of this general preference is the Exclusion Principle, used by Kim in his Exclusion Argument for the conclusion that non-reductive physicalism is committed to epiphenomenalism about mental properties and states. One version of the principle states that if a physical cause exists for a physical effect, that excludes any mental cause for the same effect. For an evaluation of this argument, see Menzies (forthcoming) and List and Menzies (forthcoming).

(3b) You don't say "Hello" $\Box\!\!\rightarrow$ your neighbour doesn't startle.

(4a) You say "Hello" loudly $\Box\!\!\rightarrow$ your neighbour startles.
(4b) You don't say "Hello" loudly $\Box\!\!\rightarrow$ your neighbour doesn't startle.

Both counterfactuals (3a) and (3b) are true, whereas not both (4a) and (4b) are true. In particular, counterfactual (4b) is false because in some of the closest worlds in which you don't say "Hello" loudly, such as those in which you say it normally, your neighbour still startles. The same point can be put in terms of contrastive focus: your neighbour startled rather didn't startle because you said "Hello" rather than didn't say it, and not because you said "Hello" loudly rather than didn't say it so.

Stephen Yablo (1992) has argued for a similar conclusion on the basis of what he calls a proportionality constraint on causation. Yablo claims that causes must be *proportional* or *commensurate* with their effects in the sense that a cause must have the right degree of specificity to account for its effect – a cause cannot be underspecific or overly specific. So, citing Peter's having the neural property N_1 as the cause of his waving his arm, or your saying "Hello" loudly as the cause of your neighbour's being startled, does not satisfy the proportionality constraint since these states are more specific than is required to account for their respective effects. They are overly specific precisely because they erroneously suggest that Peter would not have raised his arm if he had not had the neural property N_1, or that your neighbour would not have startled if you had not said "Hello" loudly.

Yablo states his proportionality constraint in terms of a metaphysical framework of event essences. By contrast, we agree with those philosophers who suggest that the idea of causal proportionality is described more satisfactorily in terms of the contrastive character of causation (Craver 2006; Woodward forthcoming a and b). In particular, we suggest that the proportionality constraint can be expressed in terms of a pair of counterfactuals having the structure of the (a) and (b) counterfactuals above. The function of the counterfactuals is to ensure that the candidate causes are of the right degree of specificity. The function of the (a) counterfactual is to rule out candidate causes that are not specific enough to account for the change in the effect variable, while the function of the (b) counterfactual is to rule out candidate causes that are too specific to account for this change.[15] In this way, the contrastive, counterfactual account of causation, proposed above, captures the idea of proportionality as well as that of difference-making.

Returning to the example about Peter's waving his arm with the intention of signalling a taxi, we can readily imagine conducting experiments the results of which confirm (1a) and (1b) and disconfirm (2b). Such experimental evidence would establish the falsity of the Physical Determination thesis, since it would establish that a mental state has a causal power not enjoyed by the physical state that realizes it. In demonstrating the falsity of the Physical Determination thesis, it would thereby demonstrate the truth of the Causal Autonomy thesis.

We emphasize again that whether some state satisfies the counterfactual conditions that constitute the difference-making causal relation is a completely empirical matter. Facts about the world determine which is the right level of causation. They determine which type of variable, a higher- or a lower-level one, is such that variation in its value can bring about a variation in an effect variable. In the example at hand, the higher-level variable is the source of causal influence. But in other circumstances lower-level variables can constitute the right level of causation. For example, suppose your interactions with your agitated neighbour have a different history. Suppose you have been getting along fine with him, but on one occasion you startle him because of the loudness of your greeting. In other words, he startles

[15] For more details about the functions of the two counterfactuals, see List and Menzies (forthcoming).

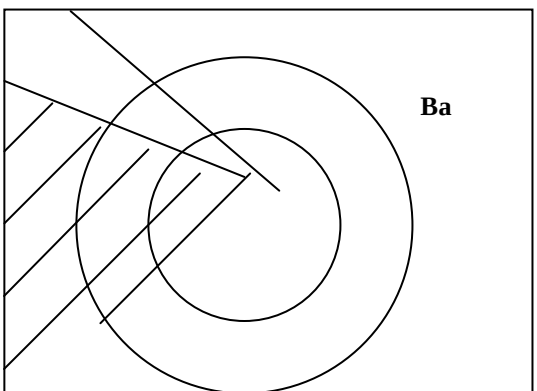
because you say “Hello” loudly, not because you simply say “Hello”. In these circumstances, the lower-level variable is the proportional cause of the effect. Likewise, suppose that what is to be explained in the example of Peter’s waving his arm is a change of fine-grained motor control rather coarse-grained behaviour. This change might be explicable only in terms of a variation in his neural states and not in terms of a variation in his mental states. In this case a neural state would be the proportional difference-making cause of the effect. Generally, the right level of causation is determined by the contrast to be explained and by the empirical facts about which variables can be varied in such a way as to account for the given contrast.

As we have just seen, the Physical Determination thesis is not *generally* true. One might wonder, nonetheless, whether it is not true for the most part, or true more often than not. One benefit of formulating the difference-making conception of causation in terms of counterfactuals is that it makes this question logically tractable.

Returning to the case of the mental state Ma, the neural state N_1a and the behaviour Ba, one can prove that the causal powers of Ma are a subset of those of N_1a – i.e., that if Ma causes Ba, then N_1a causes Ba – *only under very special conditions*. To state this result, call a causal relation between Ma and Ba *realization-sensitive* if Ba fails to hold in all those Ma-worlds that are closest ~N_1a-worlds (i.e., where Ma has a different realizer from the actual one). The result is the following:

> *Entailment Result* (List and Menzies forthcoming): If Ma causes Ba, then N_1a causes Ba if and only if the causal relation between Ma and Ba is realization-sensitive.

Rather than prove this result here, it is more instructive to describe a situation that exemplifies the result. So consider the situation represented in Figure 2. As before, the concentric spheres represent sets of more and more similar worlds to the actual world; the innermost sphere contains the actual world, labelled w, and the other worlds deemed maximally similar to it. The set of N_1a-worlds is represented by the region with diagonal lines, the set of Ma-worlds by the larger region that includes the set of N_1a-worlds. The shaded region represents the set of Ba-worlds. In this situation, it is easy to see that Ma causes Ba. First, since Ma holds throughout the innermost sphere, that sphere picks out the closest Ma-worlds, and since Ba also holds in it, counterfactual (1a) is true. Second, since Ba does not hold in any ~Ma-worlds, it fails to hold in all the closest ~Ma-worlds and thus counterfactual (1b) is true. Further, the causal relation between Ma and Ba is realization-sensitive: since Ba does not hold in any ~N_1a-worlds, it follows *a fortiori* that it does not hold in any of the closest ~N_1a-worlds that are Ma-worlds. And finally, N_1a does indeed cause Ba: counterfactuals (2a) and (2b) can easily be verified to be true.



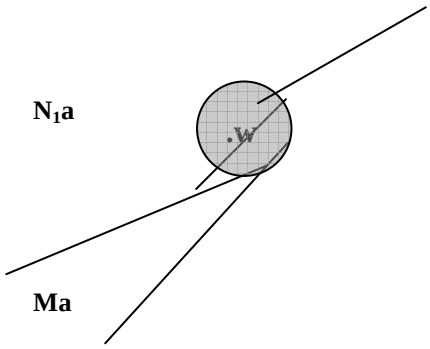


Figure 2

In view of this result, it is open to the defender of the Physical Determination thesis to argue that the thesis is sometimes true even if not always true. It is important to note, however, that the conditions under which the counterfactual pair (1a)-(1b) implies the pair (2a)-(2b) are very special. Figure 2 illustrates this point nicely. Although both Ma and its actual realizing state N_1a are difference-making causes of Ba here, the realization-sensitivity of the causal relation between Ma and Ba means that small perturbations in the way in which Ma is realized would result in the absence of Ba. In other words, if Ma were realized by any neural state other than N_1a (such as N_2a, N_3a, and so on), then Ba would cease to hold. When might we expect these conditions to obtain? If the mental property M were identical to the neural property N_1, then we would certainly expect instances of M to stand in realization-sensitive causal relations with respect to instances of N_1. The fact that M-instances had certain effects when and only when N_1-instances are present would simply reflect the identity of the properties. What other explanations could there be for the realization-sensitivity of higher-level causal relations? It is hard to think of any explanation other than the identity of the properties. But this explanation will not be available if we assume, in keeping with our overarching presupposition, that the higher-level properties are multiply realized by physical properties, and so not identical with them.

At this point it is useful to consider a logically equivalent formulation of the Entailment Result that shows more directly that the Physical Determination thesis is not generally true. In analogy with the earlier definition, call a causal relation between Ma and Ba *realization-insensitive* if Ba holds in some Ma-worlds that are closest $\sim N_1a$-worlds (i.e., where Ma has a different realizer from the actual one). The following proposition is an immediate corollary of the Entailment Result:

> *Downwards Exclusion Result* (List and Menzies forthcoming): If Ma causes Ba, then N_1a does not cause Ba if and only if the causal relation between Ma and Ba is realization-insensitive.

Again let us consider a schematic example that exemplifies this proposition, focusing on the situation represented in Figure 3. As before, the system of spheres represents sets of worlds with greater or lesser degrees of similarity to the actual world, labelled w. The set of N_1a-worlds is represented by the region with diagonal lines, and the set of Ma-worlds by the larger region that includes the set of N_1a-worlds. The shaded region represents the set of Ba-worlds. This figure shows that Ma causes Ba, since Ba holds in all the closest Ma-worlds and fails to hold in all the closest ~Ma-worlds, i.e., counterfactuals (1a) and (1b) are both true. It is also easy to see that this causal relation is realization-insensitive: Ba continues to hold in some, indeed all, of the Ma-worlds that are closest $\sim N_1a$-worlds. Finally, it is easy to see that N_1a does not cause Ba: the counterfactual (2b) is false, since Ba holds in all the closest $\sim N_1a$-worlds.

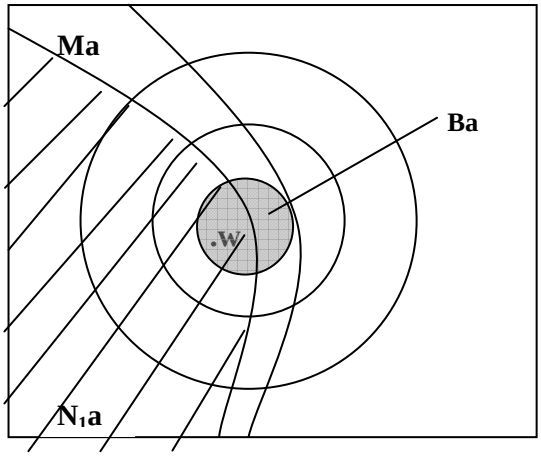


Figure 3

This result is of great significance with respect to the question of whether the Physical Determination thesis or the Causal Autonomy thesis is true. If some mental state Ma stands in a realization-insensitive causal relation to another state Ba, then this mental state has a causal power to bring about a certain effect that does not belong to the set of causal powers of its physical realizing state. Hence, the Physical Determination thesis is false and the Causal Autonomy thesis true in this situation. These logical inferences are depicted in Figure 4, where the proposition in each box logically implies the proposition in the box below it.

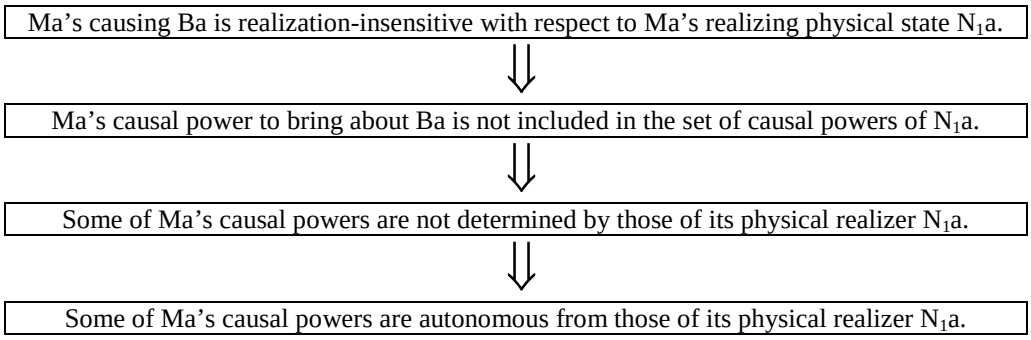


Figure 4

What is the upshot of this discussion? If we have reason to believe that a mental state stands in a realization-insensitive causal relation to some other state, then we are entitled to think that this higher-level causal relation is independent of any lower-level causal relation enjoyed by the neural realizer of the mental state. We have plenty of reason to believe that mental states do indeed stand in realization-insensitive causal relations to other states. Given that a mental state is typically realized in many different ways, we can expect that whatever causal powers it has, it has them independently of the particular way it is realized. In other words, we can expect that a mental state's causal powers do not depend on which of its possible realizers happens to be the actual one.

More generally, there is reason to think that most higher-level causal relations are realization-insensitive in ways that ensure their autonomy. Several philosophers have noted that we intuitively require causal relations to be "insensitive" in the sense that they would continue to hold under perturbations of the actual circumstances. Lewis (1986) was the first to make this observation, but Woodward (2006; forthcoming b) has developed the point in greatest detail. He says that a causal statement is insensitive to the degree that a pair of counterfactuals such as the (a) and (b) counterfactuals above would continue to hold even if the actual

circumstances were varied in certain admissible ways; and that the more insensitive the counterfactuals are, especially the (a) counterfactual, the more willing we are to count the corresponding causal statement as true. As to what counts as admissible variations to the actual circumstances, Woodward says that the changes should not be too unlikely or far-fetched; but more generally, the answer is context-sensitive, depending partly on discipline-specific considerations. We suggest that in the special sciences causal relations are typically required to be invariant under changes to the physical realization of the higher-level properties involved. And we suggest that the required realization-insensitivity of higher-level causal relations is an instance of the more general phenomenon noted by Lewis and Woodward.[16]

5. Garfinkel on the Indispensability of Higher-Level Explanations

In this section we compare the conclusions of the last section with Alan Garfinkel's views about reductionism (1981). Garfinkel argues for very similar conclusions about the indispensability of higher-level causal explanations in the special sciences, although his arguments proceed without the aid of a systematic account of causal explanation. We hope to show that Garfinkel's interesting and original insights have an internal coherence that is explicable in terms of the framework we have developed. Many of his insights follow straightforwardly as consequences from the account of difference-making causation and the formal results described above. This fact indicates a two-way relationship of confirmation: the systematizability of Garfinkel's insights in terms of our framework provides some independent confirmation of them, but the antecedent plausibility of his insights also offers some evidence in support of our framework.

Garfinkel was one of the first philosophers to emphasize the contrastive character of explanation and its importance for many issues in the philosophy of science. He advanced now-familiar arguments for the claim that explanation is relative to a contrast space. Although he focused on explanation, his observations apply equally to causation, as we have seen. Of particular interest is Garfinkel's appeal to the contrastive character of explanation to criticize reductionist claims, such as the claim that psychology is reducible to neurophysiology, the claim that thermodynamics is reducible to statistical mechanics, or the claim that social laws are reducible to principles about the actions of individuals. By *reductionism*, Garfinkel means any doctrine according to which the phenomena in the explanatory domain of one theory are best explained in terms of a lower-level theory. Reductionism is sometimes thought, he says, as an ideal, which is possible "in theory" but not "in practice". In contrast, he argues for the view that reductionism is often impossible even in theory, since the explanations of lower-level theories are simply not good enough to replace the explanations of higher-level theories.

Garfinkel discusses an example of a purported microreduction in which an explanation in terms of the macrostates of a system is eliminated in favour of an explanation in terms of its microstates (1981: 53 ff.). He asks us to imagine an ecological system composed of foxes and rabbits where the population levels of the two species periodically fluctuate: "The explanation [of the fluctuations] turns out to be that the foxes eat the rabbits to such a point that there are too few rabbits left to sustain the fox population, so the foxes begin dying off. After a while, this takes the pressure off the rabbits, who then begin to multiply until there is plenty of food for the foxes, who begin to multiply, killing more rabbits, and so forth" (1981: 53).

Now suppose that a particular rabbit r has been killed. What is the explanation of this? It is plausible to say that the cause of the rabbit's death was that the fox population was high. Can such an explanation be replaced by an explanation in terms of the underlying microstate of the system? The microstate will be an

[16] For further discussion of this issue see List and Menzies (forthcoming).

enormously complex state, specified in terms of the number and location of all the foxes and rabbits, their interactions, and perhaps their physiological states such as their reaction times. Garfinkel argues that an explanation in terms of this microstate is not satisfactory because it does not provide an answer to the question that is implicitly being asked. This is the contrastive question: Why was the rabbit r eaten rather not eaten? The explanation in terms of the microstate contains a great deal of information that is not relevant to this question and does not really answer it. At best, the microexplanation answers the different question: Why did the rabbit get eaten by fox f at place p and time t rather than by some other fox at some other place and time? In other words, the explanations have completely different objects; and so it is inappropriate to try to match one explanation with the object of the other.

Why doesn't the explanation in terms of the complex microstate of the system provide a satisfactory answer to the question of why the particular rabbit was eaten rather not eaten? Garfinkel says that microexplanation does not work because it does not say how things would have to be different in order for the rabbit not to get eaten. For example, if the microstate, specified in terms of the number and location of foxes and rabbits and their interactions at the particular time, had been slightly different, the rabbit would not have been eaten by fox f, but probably would have been eaten by another fox, given that the fox population was so high. An explanation should ideally tell us, Garfinkel argues, how the outcome is sensitive to changes in the conditions. But the microexplanation does not do this, since perturbations in the microstate still leave us with the same outcome.

More generally, a satisfactory causal explanation of the fact that the rabbit was eaten rather than not eaten must, according to Garfinkel, provide some account of "the sensitive aspects of the causal connection". He explains this as follows:

> "We can imagine the space of the substratum as underlying the whole process. We have a complete set of microstates and a principle of microexplanation, V, which explains the microstate Y_0 in terms of X_0:
>
> $$X_0 \rightarrow_V Y_0.$$
>
> The rabbit was eaten by fox f (= Y_0) because it was at a certain place, time and so on (= X_0). For most X_0, this evolution is smooth; small changes in X_0 do not make for qualitative changes. But at certain critical points, small perturbations do make a difference and will result in the rabbit's wandering out of the capture space of the fox. These critical points mark the boundaries of the regions of smooth change. They partition the underlying space into equivalence classes within which the map is stable. The crucial thing we want to know is how this set of critical points is embedded in the substratum space, for that will tell us what is really relevant and what is not. Therefore, what is necessary for a true explanation is an account of how the underlying space is partitioned into basins of irrelevant differences, separated by ridge lines of critical points." (Garfinkel 1981: 63-4)

This account of causal explanation employs ideas and terminology taken from catastrophe theory. But the basic ideas are simple. Suppose we want to explain a contrast that can be represented as the difference in the values of a variable. For convenience, let us call these the *explanandum values*. In the example, the explanandum values are the rabbit's being eaten and its not being eaten. To explain this contrast, Garfinkel tells us that we have to partition the states of the system into equivalence classes. It is implicit in the passage and his subsequent examples that the partition must satisfy certain conditions. One is that the resulting equivalence classes must be such that the laws of the system map all the members of the same class onto the same explanandum value. Accordingly, perturbing the system to change it from one state to another in the same equivalence class will not make for

qualitative changes in outcome. Another condition is that the equivalence classes must be such that the laws of the system map members of different classes onto different explanandum values. Hence perturbing the system to change it from a state belonging to one class to a state belonging to another will make for qualitative changes in outcomes. These changes will mark, in his terms, the boundaries of the regions of smooth transition. In sum, the explanatory partition must be such that "the underlying space is partitioned into equivalence classes within which differences do not make a difference but across which differences *do* make a difference" (1981: 65).

In terms of this account of causal explanation, it is easy to understand his remarks about what counts as a satisfactory explanation of the fact that rabbit r was eaten rather than not eaten. Partitioning the states of the system into the class of states in which the fox population is high and the class of states in which it is not high satisfies, or approximately satisfies, the two conditions.[17] First, the laws of the system map all states in a given equivalence class onto the same explanandum value; and second, the laws map states in different equivalence classes onto different explanandum values. Accordingly, the transition from a state in which the fox population is high to a state in which it is not crosses the boundary between regions of smooth transition. By contrast, the partition of the states into a class consisting of a single microstate X_0 and a class consisting of all other microstates will not satisfy the conditions on an explanatory partition. The states in the second class are not all equivalent from the point of view of the laws of the system. For the laws map the states in this class onto different explanandum values and indeed many of these states will be mapped onto the same explanandum value as the state X_0, reflecting the fact that even if rabbit r was not eaten by fox f it may have been eaten by some other fox.

How does Garfinkel's account of causal explanation relate to the difference-making account of causation? It is not too difficult to see the structural parallels between the two accounts. Suppose that the property-instance Fa is a difference-making cause of the property-instance Gb. Then the following pair of counterfactuals must be true:

(5a) $Fa \mathrel{\Box\!\!\rightarrow} Gb$
(5b) $\sim Fa \mathrel{\Box\!\!\rightarrow} \sim Gb$

The truth of these counterfactuals entails that every closest Fa-world is a Gb-world and that every closest ~Fa-world is a ~Gb-world. Notice that the set of closest Fa-worlds and the set of closest ~Fa-worlds need not belong to the same sphere of worlds in the system of spheres. Indeed, the closest Fa-worlds may be a subset of one sphere and the closest ~Fa worlds a subset of a different sphere. But if we focus on the special case in which they belong to the same sphere, we can see that the truth of these counterfactuals implies the existence of a partition on the common sphere that satisfies Garfinkel's two conditions on an explanatory partition. The possible worlds in this sphere fall into two equivalence classes: one class consists of worlds whose laws map the state Fa onto the state Ga and the other class consists of worlds whose laws map the state ~Fa onto the state ~Gb.[18]

So we have an argument that if a state is a difference-making cause of another, then the first state will be a good causal explanation of the second on Garfinkel's account. The argument in the other direction is even more

[17] It follows from the fact that the explanandum consists of a contrast between two values that the explanatory partition must consist of two equivalence classes. But this need not generally be the case.

[18] If the closest Fa-worlds and the closest ~Fa-worlds do not belong to a common sphere of worlds, the difference-making account of causation is informationally richer, but it is still possible to construct a partition of a suitable set of worlds that satisfies Garfinkel's conditions on an explanatory partition, though the construction is not so intuitively natural.

straightforward. If in all relevantly similar systems, the laws map Fa-states onto Gb-states and ~Fa-states onto ~Gb-states, then it is a simple matter to show the pair of counterfactuals above will be true. In sum, the difference-making account of causation and Garfinkel's account of causal explanation are structurally similar to each other.

This is not the only parallel between our framework and Garfinkel's. Garfinkel argues that many macroexplanations in the special sciences cannot be eliminated and replaced by microexplanations. His justification of this claim is based on the fact that an explanation must have a certain amount of stability under perturbations of its conditions; and that certain structural features of special-science systems ensure the stability of macroexplanations. For example, the successful causal explanation of the rabbit's being eaten in terms of the high fox population rests on a certain stability or resilience of the causal processes in this system: the rabbit was eaten by fox f, but if it had not been eaten by this fox it would have been eaten by another fox. So the causality with which the effect is produced has a strong resilience: "The very fact that the rabbit did not wander into the capture space of fox f makes it likely that it will be eaten by another fox" (1981: 57). When this is true of a system, Garfinkel says, we have a case of "redundant causation". He writes:

> "Systems exhibiting redundant causality have, for every consequent Q, a bundle of antecedents (P_i) such that:
>
> 1. If any one of the P_i is true, so will be Q.
> 2. If one P_i should not be the case, some other will.
>
> Obviously, in any system with redundant causality, citing the actual P_i that caused Q will be defective as an explanation. This will apply to many cases in which P_i is the microexplanation." (Garfinkel 1981: 58)

We think that it is, strictly speaking, misleading in this context to use the term "redundant causation", which is normally used to describe cases of pre-emption and overdetermination. The situation described here is essentially different from one in which multiple causes lead to the same effect, such as when a victim is hit by multiple bullets. We suggest that the phenomenon described by Garfinkel is actually the realization-insensitivity of causal relations. Consider the causal relation between P and Q, in Garfinkel's notation, where P is a higher-level macrostate that is actually realized by a microstate P_i but could have been realized by any of the microstates $P_1,\ldots, P_n$. This causal relation is realization-insensitive just in case Q is true in some of the closest $\sim P_i$-worlds that are still P-worlds, i.e., Q remains true in some of the worlds in which P is realized differently. It is easy to see that the conditions that Garfinkel stipulates for "redundant causation" ensure that the causal relation between P and Q is realization-insensitive in this sense. His second condition ensures that among the closest $\sim P_i$-worlds there are some that are P-worlds; and his first condition ensures that all these closest $\sim P_i$-worlds that are P-worlds are ones in which Q is true.

What is the point of these remarks? It is that Garfinkel has arrived at essentially the same conclusion reached in the last section: the key to understanding the ineliminability of macrostate causal explanations is the realization-insensitivity of the causal links they invoke. In the last section, we argued that when an upper-level macroexplanation rests on a realization-insensitive causal relation, it cannot be replaced by a lower-level microstate explanation. Garfinkel says essentially the same thing. The causal explanation of the death of rabbit r in terms of the high fox population cannot be replaced by an explanation in terms of the actual microstate in which it is eaten by fox f because even if the high fox population were realized by some other microstate, it would still be true that the rabbit would have been eaten, if

not by fox f then by some other fox. The Downwards Exclusion Result, described in the last section, implies that the macroexplanation conveys essential contrastive, difference-making information that is not conveyed by the microstate explanation.

In conclusion, the results of this paper recapitulate many of the conclusions reached less formally by Garfinkel. There are, indeed, some striking similarities between our approaches. First, both approaches attach special significance to the contrastive character of causation or causal explanation in establishing the right level of causation or causal explanation. Secondly, both approaches provide accounts of causation or causal explanation in terms of a partition of states or possible worlds into classes satisfying certain constraints concerning the lawful mappings of one state onto another. Finally, both approaches justify and explain the ineliminability of higher-level causation or causal explanations in terms of the fact that they are realization-insensitive. The framework developed in this paper, involving the difference-making account of causation and the Entailment and Downwards Exclusion Results, is especially advantageous in showing how these common ideas can be systematized in a coherent way.

References

Broad, C.D. (1925) *The Mind and Its Place in Nature*. London: Routledge & Kegan Paul.

Crane, T. (2001) *The Elements of Mind*. Oxford: Oxford University Press.

Craver, C. (2007) *Explaining the Brain*. New York: Oxford University Press.

Fodor, J. 1974. "Special Sciences, or the Disunity of Sciences as a Working Hypothesis", *Synthese*, 28, pp.97-115.

Garfinkel, A. (1981) *Forms of Explanation*. New Haven: Yale University Press.

Gopnik, A. and Schulz (2007). *Causal Learning: Psychology, Philosophy and Computation.* New York: Oxford University Press

Hitchcock, C. (2001) "The Intransitivity of Causation Revealed in Equations and Graphs", *Journal of Philosophy* 98: 273-99.

Horgan, T. (1989) "Mental Quausation". *Philosophical Perspectives*, 3, [\pp.47-76.

Kim, J. (1992) "Downward Causation" in A. Beckermann, H. Flohr, and J. Kim eds., *Emergence or Reduction*. New York and Berlin: De Gruyter.

Kim, J. (1998) *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. Cambridge, MA: MIT Press.

Kim, J. (2005) *Physicalism or Something Near Enough*. Princeton, NJ: Princeton University Press.

Le Pore, E. and Loewer, B. (1987) "Mind Matters", *Journal of Philosophy*, 84, pp.630-642.

Lewis, D. (1973) *Counterfactuals*. Oxford: Blackwell.

Lewis, D. (1986) "Postscripts to 'Causation", *Philosophical Papers: Volume 2*. Oxford: Oxford University Press.

List, C. and Menzies, P. (forthcoming) "Non-reductive Physicalism and the Limits of the Exclusion Principle".

MacLaughlin, B. (1992) "The Rise and fall of British Emergentism", in A. Beckermann, H. Flohr, and J. Kim eds., *Emergence or Reduction*. New York and Berlin: De Gruyter.

Menzies, P. (forthcoming) "The Exclusion Problem, the Determination Relation, and Contrastive Causation" J. Hohwy and J. Kallestrup (eds.), *Being Reduced: New Essays on Reductive Explanation and Special Science Causation*. OUP.

Pearl, J. (2000) *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Raatikainen, P. (2006) "Mental Causation, Interventions, and Contrasts", unpublished manuscript, University of Helsinki.

Shoemaker, S. (2001) "Realization and Mental Causation" in C, Gillett and B. Loewer, eds., *Physicalism and Its Discontents*. Cambridge: Cambridge University press.

Spirtes, P., Glymour, C. and Scheines, R. (2000) *Causation, Prediction and Search*. Cambridge, MA: MIT Press.

Woodward, J. (2003) *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Woodward, J. (2006) "Sensitive and Insensitive Causation", *Philosophical Review* 115: 1-50.

Woodward, J. (forthcoming a) "Cause and Explanation in Psychiatry: An Interventionist Perspective".

Woodward, J. (forthcoming b) "Mental Causation and Neural Mechanisms", in J. Hohwy and J. Kallestrup (eds.), *Being Reduced: New Essays on Reductive Explanation and Special Science Causation*. Oxford: Oxford University Press.

Yablo, S. (1992), "Mental Causation", *Philosophical Review* 101: 245-280.