A Glimpse of the *Secret Connexion:* Harmonizing Mechanisms with Counterfactuals

Stathis Psillos
University of Athens

Among the current philosophical accounts of causation two are the most prominent. The first is James Woodward's interventionist counterfactual approach; the second is the mechanistic approach advocated by Peter Machamer, Lindley Darden, Carl Craver, Jim Bogen and Stuart Glennan. The counterfactual approach takes it that causes make a difference to their effects, where this difference-making is cashed out in terms of actual and counterfactual interventions. The mechanistic approach takes it that two events are causally related if and only if there is a mechanism that connects them. In this paper I examine them both in some detail. After pointing out some important problems that both approaches face, I argue that there is a sense in which the counterfactual approach is more basic than the mechanistic one in that a proper account of mechanisms depends on counterfactuals while counterfactuals need not be supported (or depend on) mechanisms. Nonetheless, I also argue that if both approaches work in tandem in practice, they can offer us a better understanding of aspects of Hume's secret connexion and hence a glimpse of it.

An earlier draft of this paper was presented at the fourth Athens-Pittsburgh conference on *Proof and Demonstration in Science and Philosophy,* in Delphi June 2003. Several portions of it were offered in seminars in the University of California San Diego, Caltech, American College of Thessaloniki, Bogazici University, University of Oslo, University of Gent, and in the workshop of the Metaphysics in Science Group in Athens, June 2003. Participants in all of these events, and other friends and colleagues, have made numerous useful comments and suggestions. Some of them criticised the paper relentlessly. Others were more positive about it. I would like to thank them all wholeheartedly. They are too many to name them all. But it would be inappropriate of me if I did not mention the following: an anonymous reader of *Perspectives on Science,* Ken Binmore, Diderik Batens, Craig Callender, Nancy Cartwright, Paul Churchland, Peter Clark, Olav Gjelsvik, Stuart Glennan, Alan Hajek, Chris Hitchcock, Ilhan Inan, Gurol Irzik, Philip Kargopoulos, Patricia Kitcher, Buket Korkut, Peter Machamer, Vincent Mueller, Daniel Nolan, Panagiotis Oulis, Kostas Pagondiotis, Nils Roll-Hansen, Eric Watkins, Erik Weber and Jim Woodward.

Perspectives on Science 2004, vol. 12, no. 3

1. Introduction

Let me start with a general observation about causation (which I will state without much defense). Though traditionally, *causation* has been taken to be a single, unitary, concept, there has been a tendency, as of late, to question this assumption. The case for there being *two* concepts of causation has been made, quite forcefully, by Ned Hall (forthcoming). He distinguishes between causation as *dependence* and causation as *production.* Hall takes dependence to be *counterfactual* dependence, while he takes the concept of production (*c* produces *e*) as primitive. I am very sympathetic to this distinction but not yet quite prepared to claim that it corresponds to two different concepts of causation. Be that as it may, it is plausible to argue that there are these two broad strands in our thinking about causation. According to the first, to say that *c* causes *e* is to say that *e* suitably depends on *c,* while according to the second, to say that *c* causes *e* is to say that something *in* the cause produces (brings about) the effect or that there is something (e.g., a mechanism) that links the cause and the effect. We may usefully call the first approach *dependence* and the second *productive.*[1] On the face of it, there can be different ways to cash out the relation of dependence. It may be nomological dependence (cause and effect fall under a law), or counterfactual dependence (if the cause hadn't happened, the effect wouldn't have happened), or probabilistic dependence (the cause raises the probability of the effect). Similarly, there may be different ways to cash out the concept of mechanism. It may be that only one single thing, e.g., the transfer of energy, captures mechanistic production. Or it may be that there are a number of distinct things that do this job. Presently, I will not discuss these general issues further. Instead, I want to focus on two current philosophical accounts of causation that seem to be the most prominent, and which exemplify these two strands in our thinking about causation. The first is James Woodward's *interventionist counterfactual* approach; the second is the *mechanistic* approach advocated by Peter Machamer, Lindley Darden, Carl Craver, Jim Bogen and Stuart Glennan.

The counterfactual approach takes it that causes *make a difference* to their effects. This difference-making is cashed out in terms of counterfactual dependence. In particular, Woodward's *interventionist counterfactual* approach takes the relationship among some variables X and Y to be causal if, where an intervention changed the value of X appropriately, the relationship between X and Y would remain invariant *and* the value of Y would change.

1. It can be argued that the dependence approach goes back to Hume, while the productive one goes back to Descartes.

The mechanistic approach takes it that causes *produce* their effects. This production is cashed out in terms of mechanisms: two events are causally related if and only if there is a *mechanism* that connects them. Mechanisms are taken to be complex systems, which are composed of parts, have internal structure or organization and certain spatio-temporal locations. The mechanism has a characteristic behavior in virtue of the properties, dispositions or capacities of its parts as well as in virtue of how these parts are organized and interact with each other. *What* the mechanism is doing (its characteristic activity, its behavior or its output) is caused and explained by the details of *how* it is doing it. These details include the internal workings of the mechanism.

These two approaches (the interventionist counterfactual and the mechanistic) fall under the dependence approach and the productive approach respectively. On the face of it, the two approaches need *not* be in conflict.[2] But, overall, both approaches tend to be monistic. Advocates of each argue that their own approach captures causation much better than their opponents'. For instance, Jim Bogen (2003) claims that the notion of a counterfactual intervention is too obscure to serve as the basis of an account of causation. He thinks that the mechanistic approach has distinctive advantages over the counterfactual approach, the most salient of which is that it avoids counterfactuals. Woodward (2003a, p. 93) argues for a "monocriterial" view of causation, the sole criterion being invariance under actual and counterfactual interventions. And in his (2002), he argues that the concept of mechanism can be fully accommodated within his own counterfactual framework.

In his less skeptical moments, David Hume ([1777] 1975, p. 66) noted: "[E]xperience only teaches us, how one event constantly follows another; without instructing us in the secret connexion, which binds them together, and renders them inseparable." Though there may be other ways

2. There is an important *prima facie* difference. The interventionist counterfactual approach takes the causal relata to be *variables,* whereas the mechanistic approach takes them to be *events* (or processes, which can be seen as sequences of events). Woodward (2003a, p. 112), for instance, insists that causes should be such that it makes sense to say of them that they could be changed or manipulated. Thinking of them as variables, which can take different values, is then quite natural. But as he goes on to note, it is not difficult to translate talk in terms of changes in the values of variables into talk in terms of events and conversely. For instance, instead of saying that the hitting by the hammer (an event) caused the shattering of the vase (another event), we may say that the change of the value of a certain indicator variable from *not-hit* to *hit* caused the change of the value of another variable from *unshattered* to *shattered.* This strategy, however, will not work in cases in which putative causes cannot be understood as values of variables. For an important attempt to show how the relata of the interventionist counterfactual approach can be seen as events, see Kluve (2004, especially pp. 81–2).

to interpret this claim, it seems plausible to say that Hume allowed that causation is an intrinsic relation among events (the *secret connexion*), but that we can only get at some extrinsic marks of it. In modern terminology, if causation is taken to be an *intrinsic* relation, then that *c* causes *e* will have to depend *entirely* on the intrinsic properties of *c* and *e* and the relations between *c* and *e.* In particular, it won't depend on things that happen at other places and in other times (e.g., on regularities, or on the presence or absence of other potential or actual causes).[3] My overall view, then, is that in so far as this secret connexion is an *intrinsic* relation between the causal relata, neither of the two approaches I will be discussing tells us what this relation is. For none of them, though for different reasons, renders causation an intrinsic relation.

Yet, after pointing out some important problems that both approaches under discussion face, I will argue that there is still an asymmetry between them. There is a sense, I will claim, in which the counterfactual approach (a fortiori the *dependence* approach) is more *basic* than the mechanistic (a fortiori the *productive*) one in that a proper account of mechanisms depends on counterfactuals while counterfactuals need not be supported (or depend on) mechanisms. Nonetheless, I will also argue that if both approaches work in tandem in *practice,* they can offer us a better understanding of aspects of Hume's *secret connexion* and hence a *glimpse* of it.

2. Early views

J. L. Mackie's (1974) work on causation is the recent common source of both approaches under discussion. Mackie explicitly appealed to counterfactuals in his definition of the meaning of singular causal statements. He argued that a causal statement of the form '*c* caused *e*' should be understood as meaning '*c* was necessary in the circumstances for *e*', where *c* and *e* are distinct event-tokens. Necessity-in-the-circumstances, he added, should be understood as follows: if *c* hadn't happened, then *e* wouldn't have happened.

Mackie's counterfactuals are not, strictly speaking, true or false: they do not describe, or fail to describe, "a fully objective reality" (1974, p. xi). Instead, they can be reasonable or unreasonable assertions, depending on the inductive evidence that supports them (cf. 1974, pp. 229–30). For instance, in assessing the counterfactual "If this match had been struck, it would have lit," the evidence plays a *double role.* It *first* establishes inductively a generalization. But *then,* "it continues to operate separately in making it reasonable to assert the counterfactual conditionals which look like an extension of the law into merely possible worlds" (1974, p. 203).

3. For more on this issue, see my (2002, pp. 128–9).

So for Mackie, it is general propositions (via the evidence there is for them) that carry the weight of counterfactual assertions. If, in the actual world, there is strong evidence for the general proposition "All *F*s are *G*s" we can reasonably assert that "if *x* had been an *F* it would have been a *G*" based on the evidence that supports the general proposition. Mackie was no realist about possible worlds. He did not think that they were as real as the actual. Hence, his talk of possible worlds was a mere *facon de parler* (cf. 1974, p. 199).

These evidence-based counterfactuals *cannot* ground a fully objective distinction between causal sequences of events and non-causal ones. This created a tension in Mackie's overall project. For although he explicitly aimed to identify an *intrinsic* feature of a sequence of events that makes the sequence causal, his dependence on evidence-based counterfactuals jeopardized this attempt: whether a sequence of events will be deemed causal will depend, on his view, on an *extrinsic* feature, viz., on whether there is *evidence* to support the relevant counterfactual conditional. It is for this reason that Mackie went on to try to uncover an *intrinsic* feature of causation, in terms of a *mechanism* that connects the cause and the effect.

As Hume famously noted, the alleged necessary tie between cause and effect is not observable. Mackie thought, not unreasonably, that we might still *hypothesize* that there is such a tie, and then try to form an intelligible theory about what it might consist in. His hypothesis is that the tie consists in a "causal mechanism," that is, "some continuous process connecting the antecedent in an observed [. . .] regularity with the consequent" (1974, p. 82). Where Humeans, generally, refrain from accepting anything other than spatiotemporal contiguity between cause and effect, Mackie thinks that mechanisms might well constitute "the long-searched-for link between individual cause and effect which a pure regularity theory fails, or refuses, to find" (1974, pp. 228–9).

Mackie's own view was that this mechanism consists in the qualitative or structural continuity, or *persistence,* exhibited by certain processes, which can be deemed causal. There needn't be some general feature (or structure) that persists in every causal process. What persists will depend on the details of the actual "laws of working" that exist in nature. For instance, it can be "the total energy" of a system, or the "number of particles," or "the mass and energy" of a system (cf. 1974, pp. 217–8). But insofar as something persists in a certain process, this feature can be what connects together the several stages of this process and renders it causal.

So early versions of both current views about causation can be found in Mackie's work. In fact, it turns out that the mechanistic view was more central in Mackie's overall approach, since it promised to offer a more objective account of causation and to avoid the notorious context-sensitivity

of counterfactual assertions. Yet, Mackie's attempt to spell out mechanisms in terms of persistence was deeply problematic.[4]

After Mackie, the counterfactual and the mechanistic approaches parted their ways. They were separately pursued and developed by other able philosophers. The *locus* of the standard counterfactual theories of causation is the work of the late David Lewis (1986). Unlike Mackie, Lewis (1973) put forward a *quasi-objectivist* theory of counterfactuals, based on possible-words semantics. For lack of space, I will not review this approach here.[5] I will only make the following point, which is relevant to what follows. Lewis's theory renders causation an *extrinsic* relation between events, since it analyses causation in terms of counterfactual dependence among events and it analyses counterfactuals in terms of relations of similarity among possible worlds. In fact, there is a rather important reason why counterfactual theories *cannot* offer an intrinsic characterization of causation. If causation amounts to counterfactual dependence among events, then the truth of the claim that *c* causes *e* will depend on the absence of causal overdeterminers, since if the effect *e* is causally overdetermined, it won't be counterfactually dependent on any of its causes. But the presence or absence of overdeterminers is certainly *not* an intrinsic feature of a causal sequence.

The *locus* of the standard mechanistic theories of causation is the work of the late Wesley Salmon. Unlike Mackie, Salmon (1984) tried to characterize directly when a process is *causal,* thereby finding the mechanism that links cause and effect. So he took processes rather than events to be the basic entities in a theory of physical causation. Here again, I will not review Salmon's views.[6] A general note, however, is important for what follows. Roughly put, Salmon characterized as causal those (and only those) processes that transmit *marks.* Salmon's original promise was for a theory of causation that does *not* involve counterfactuals. The promise, however, was not to be fulfilled. Central to Salmon's theory was the *ability* of a process to transmit a mark. But the ability is a capacity or a disposition, and it is essential for Salmon that it is so. For he wants to insist that a process is causal, even if it is *not* actually marked (cf. 1984, p. 147). So, a process is causal if it *could* be marked. Counterfactuals loom large! All this is explained in some detail in my (2002, pp. 112–8). But the message is clear: Salmon's original mechanistic approach cannot do away with counterfactuals. In fact, Salmon's appeal to counterfactuals has led some philosophers (e.g., Kitcher 1989) to argue that, in the end, Salmon

4. See my (2002, pp. 108–10) for a discussion of the central problems.
5. See at (2002, chapter 3) for an account of it and of its main problems.
6. See my (2002, chapter 4) for a detailed account.

has offered a *variant* of the counterfactual approach to causation. Salmon has always been very skeptical about the objective character of counterfactual assertions. So as he (1997, p. 18) said, it was "with great philosophical regret," that he took counterfactuals on board in his account of causation.[7]

So far, I have been engaged in stage setting. My focus is the current versions of the counterfactual and the mechanistic approaches. Though more can be said, it seems enough to state that Woodward's development of the counterfactual approach is an attempt to provide an account that avoids the metaphysical extremes and epistemological pitfalls of Lewis's view, and that the mechanistic approaches of Glennan and Machamer, Darden and Craver try to provide an account of causal mechanisms that is more in tune with the epistemic and explanatory practices of scientists who study mechanisms.

3. Counterfactual manipulation

In a series of papers and a book, Woodward (1997; 2000; 2003a; 2003b) offers an account of causation based on the idea of counterfactual manipulation. His theory is *counterfactual* in the following sense: what matters is what *would* happen to a relationship, *were* interventions to be carried out. A relationship among some variables X and Y is causal if, were one to intervene to change the value of X appropriately, the relationship between X and Y wouldn't change *and* the value of Y would change. To use a stock example, the force exerted on a spring *causes* a change of its length, because if an intervention changed the force exerted on the spring, the length of the spring would change too (but the relationship between the two magnitudes—expressed by Hooke's law—would remain invariant, within a certain range of interventions).

Woodward (1997; 2000; 2003a) has analyzed further the central notions of invariance and intervention. The gist of his characterization of an *intervention* is this. A change of the value of X counts as an intervention I if it has the following characteristics:

> a) the change of the value of X is entirely due to the intervention I;
> b) the intervention changes the value of Y, if at all, only through changing the value of X.

The *first* characteristic makes sure that the change of X does not have causes other than the intervention I, while the *second* makes sure that the

7. Dowe's (2000) Conserved Quantity theory aims to free the mechanistic view of causation from counterfactuals. For a discussion of whether this is so, see my (2002, pp. 125–7).

change of *Y* does not have causes other than the change of *X* (and its possible effects).[8] These characteristics are meant to ensure that *Y*-changes are exclusively due to *X*-changes, which, in turn, are exclusively due to the intervention *I*. As Woodward notes, there is a close link between intervention and manipulation. Yet, his account makes no special reference to human beings and their (manipulative) activities. Insofar as a process has the right characteristics, it counts as an intervention. So, interventions can occur "naturally," even if they can be highlighted by reference to "an idealized experimental manipulation" (2000, p. 199).

Woodward links the notion of intervention with the notion of *invariance.* A certain relation (or a generalization) is invariant, Woodward says, "if it would continue to hold—would remain stable or unchanged—as various other conditions change" (2000, p. 205). What really matters for the characterization of invariance is that the generalization remains stable under a set of actual and counterfactual *interventions.* So Woodward (2000, p. 235) notes: "the notion of invariance is obviously a modal or counterfactual notion [since it has to do] with whether a relationship would remain stable if, perhaps contrary to actual fact, certain changes or interventions were to occur." Counterfactuals have been reprimanded on the ground that they are context-dependent and vague. Take, for instance, the following counterfactual: "If he had not smoked so heavily, he would have lived a few years more." What is it for it to be true? Any attempt to say whether it is true, were it to be possible at all, would require specifying what else should be held fixed. For instance, other aspects of his health should be held fixed, assuming that other factors (e.g., a weak heart) wouldn't cause a premature death, anyway. But what things to hold fix is not, necessarily, an objective matter. Or, consider the following pair of counterfactuals: "If Julius Caesar had been in charge of U. N. Forces during the Korean war, then he would have used nuclear weapons" and "If Julius Caesar had been in charge of U. N. Forces during the Korean war, then he would have used catapults." It is hard to see how we could possibly tell which of them, if any, is true.

3.1 Experimental counterfactuals

Woodward is very careful in his use of counterfactuals. Not all of them are of the right sort for the evaluation of whether a relation is causal. Only counterfactuals that are related to *interventions* can be of help. An intervention gives rise to an "active counterfactual," that is, to a counterfactual whose antecedent is made true "by interventions" (1997, p. S31; 2000,

8. There is a *third* characteristic too, viz., that the intervention *I* is not correlated with other causes of *Y* besides *X*.

p. 199). In his (2003a, p. 122) he stresses that "the appropriate counterfactuals for elucidating causal claims are not just any counterfactuals but rather counterfactuals of a very special sort: those that have to do with the outcomes of hypothetical interventions. [. . .] [I]t does seem plausible that counterfactuals that we do not know how to interpret as (or associate with) claims about the outcomes of well-defined interventions will often lack a clear meaning or truth value." In his (2003b, p. 3), he very explicitly characterizes the appropriate counterfactuals in terms of *experiments:* they "are understood as claims about what would happen if a certain sort of experiment were to be performed." Consider a case he (2003b, pp. 4–5) discusses. Take Ohm's law (that the voltage E of a current is equal to the product of its intensity I times the resistance R of the wire) and consider the following two counterfactuals:

> (1) If the resistance were set to $R=r$ at time t, and the voltage were set to $E=e$ at t, then the intensity I would be $i=e/r$ at t.
> (2) If the resistance were set to $R=r$ at time t, and the voltage were set to $E=e$ at time t, then the intensity I would be $i^* \neq e/r$ at t.

There is nothing mysterious here, says Woodward, "as long as we can describe how to test them" (2003b, p. 6). We can perform the experiments at a future time t^* in order to see whether (1) or (2) is true. If, on the other hand, we are interested in what *would* have happened had we performed the experiment in a past time t, Woodward invites us to rely on the "very good evidence" we have "that the behavior of the circuit is stable over time" (2003b, p. 5). Given this evidence, we can assume, in effect, that the *actual* performance of the experiment at a future time t^* is as good for the assessment of (1) and (2) as a *hypothetical* performance of the experiment at the past time t.

For Woodward, the truth-conditions of counterfactual statements (and their truth-values) are not specified by means of an abstract metaphysical theory, e.g., by means of abstract relations of similarity among possible worlds. He calls his own approach "pragmatic." That's how he (2003b, p. 4) puts it:

> For it to be legitimate to use counterfactuals for these goals [understanding causal claims and problems of causal inference], I think that it is enough that (a) they be useful in solving problems, clarifying concepts, and facilitating inference, that (b) we be able to explain how the kinds of counterfactual claims we are using can be tested or how empirical evidence can be brought to bear on them, and (c) we have some system for representing counterfactual claims

> that allows us to reason with them and draw inferences in a way that is precise, truth-preserving and so on.

Recall that Mackie had an evidence-based view of counterfactuals. He thought that counterfactual statements were not, strictly speaking, true or false. Rather, they are *warranted* only when there is evidence for a relevant generalization. Unlike Mackie's, Woodward's view is meant to be realist and objectivist. He is quite clear that counterfactual conditionals have non-trivial truth-values independently of the actual and hypothetical experiments by virtue of which it can be assessed whether they are true or false. He (2003b, 5) says:

> On the face of things, doing the experiment corresponding to the antecedent of (1) and (2) doesn't *make* (1) and (2) have the truth values they do. Instead the experiments look like ways of *finding out* what the truth values of (1) and (2) were all along. On this view of the matter, (1) and (2) have non-trivial truth values—one is true and the other false—even if we don't do the experiments of realizing their antecedents. Of course, we may not *know* which of (1) and (2) is true and which false if we don't do these experiments and don't have evidence from some other source, but this does not mean that (1) and (2) both have the same truth-value.

This point is repeated in his (2003a, p. 123), where he stresses "We think instead of [a counterfactual such as (1) above] as having a determinate meaning and truth value whether or not the experiment is actually carried out—it is precisely because the experimenters want to *discover* whether [this counterfactual] is true or false that they conduct the experiment." So though "pragmatic," Woodward's theory is also objectivist. But it is minimally so. As he (2003a, pp. 121–2) notes, his view: "requires only that there be facts of the matter, independent of facts about human abilities and psychology, about which counterfactual claims about the outcome of hypothetical experiments are true or false and about whether a correlation between *C* and *E* reflects a causal relationship between *C* and *E* or not. Beyond this, it commits us to no particular metaphysical picture of the 'truth-makers' for causal claims."

There are a few delicate issues here to be reckoned with. I will restrict myself to the following: *what are the truth-conditions of counterfactual assertions?* Woodward doesn't take all counterfactuals to be meaningful and truth-valuable. As we have seen (see also 2003a, p. 122), he takes only a subclass of them, the active counterfactuals, to be such. However, he does not want to say that the truth-conditions of active counterfactuals are fully specified by (are reduced to) actual and hypothetical experiments. If

he said this, he could no longer say that active counterfactuals have determinate truth-conditions independently of the (actual and hypothetical) experiments that can test them. In other words, Woodward wants to distinguish between the truth- conditions of counterfactuals and their evidence-(or test) conditions, which are captured by certain actual and hypothetical experiments. The problem that arises is this. Though we are given a relatively detailed account of the evidence-conditions of counterfactuals, we are not given anything remotely like this for their *truth-conditions.* What, in other words, is it that makes a certain counterfactual conditional true?

A thought here might be that there is no need to say anything more about the truth-conditions of counterfactuals other than offering a Tarski-style meta-linguistic account of them of the form

> (*T*) 'If *x* had been the case, then *y* would have been the case' is true iff if *x* had been the case, then *y* would have been the case.

This move is possible, but not terribly informative. We don't know when to assert (or hold true) the right hand-side. And the question is precisely this: when is it right to assert (or hold true) the right-hand side? Suppose we were to tell a story in terms of actual and hypothetical experiments that realize the antecedent of the right-hand side of (*T*). The obvious problem with this move is that the truth-conditions of the counterfactual conditional would be specified in terms of its evidence-conditions, which is exactly what Woodward wants to block. Besides, if we just stayed with (*T*) above, without any further explication of its right-hand side, *any* counterfactual assertion (and not just the active counterfactuals) would end up meaningful and truth-valuable. Here again, Woodward's project would be undermined. Woodward is adamant: "Just as non counterfactual claims (e.g., about the past, the future, or unobservables) about which we have no evidence can nonetheless possess non-trivial truth-values, so also for counterfactuals" (2003b, p. 5). This is fine. But in the case of claims about the past or about unobservables there are well-known stories to be told as to what the difference is between truth- and evidence-conditions. When it comes to Woodward's counterfactuals, we are *not* told such a story.

In light of the above, there are two options available. The first is to *collapse* the truth-conditions of counterfactuals to their evidence-conditions. One can see the *prima facie* attraction of this move. Since evidence-conditions are specified in terms of actual and hypothetical experiments, the right sort of counterfactuals (the active counterfactuals) and only those end up being meaningful and truth-valuable. But there is an important drawback. Recall counterfactual assertion (1) above. On the option pres-

ently considered, what makes (1) true is that its evidence-conditions obtain. Under this option, counterfactual conditionals lose, so to speak, their counterfactuality. (1) becomes a shorthand for a future prediction and/or the evidence that supports the relevant law. If *t* is a *future* time, (1) gives way to an actual conditional (a prediction). If *t* is a past time, then, given that there is good evidence for Ohm's law, all that (1) asserts under the present option is that there has been good evidence for the law.

In any case, Woodward seems keen to keep evidence- and truth-conditions apart. Then, (and this is the *second* option available) some informative story should be told as to what the truth-conditions of counterfactual conditionals *are* and *how* they are connected with their evidence-conditions (that is, with actual and hypothetical experiments). There may be a number of stories to be told here.[9] The one I favor ties the truth-conditions of counterfactual assertions to *laws of nature.* It is then easy to see how the evidence-conditions (that is, actual and hypothetical experiments) are connected with the truth-conditions of a counterfactual: actual and hypothetical experiments are symptoms for the presence of a law. There is a hurdle to be jumped, however. It is notorious that many attempts to distinguish between genuine laws of nature and accidentally true generalizations rely on the claim that laws do, while accidents do not, support counterfactuals. So counterfactuals are called for to distinguish laws from accidents. If at the same time laws are called for to tell when a counterfactual is true, we go around in circles. Fortunately, there is the Mill- Ramsey-Lewis view of laws (see my 2002, chapter 5). Laws are those regularities that are members of a coherent system of regularities, in particular, a system which can be represented as an ideal deductive axiomatic system striking a good balance between *simplicity* and *strength.* On this view, laws are identified independently of their ability to support counter-

9. One might try to keep truth- and evidence-conditions apart by saying that counterfactual assertions have excess content over their evidence-conditions in the way in which statements about the past have excess content over their (present) evidence-conditions. Take the view (roughly Dummett's) that statements about the past are meaningful and true insofar as they are verifiable (i.e., their truth can be known). This view may legitimately distinguish between the *content* of a statement about the past and the present or future evidence there is for it. Plausibly, this excess content of a past statement may be cast in terms of counterfactuals: a meaningful past statement *p* implies counterfactuals of the form "if *x* were present at time *t, x* would verify that *p.*" This move presupposes that there are meaningful and true counterfactual assertions. But note that a similar story *cannot* be told about counterfactual conditionals. If we were to treat their supposed excess content in the way we just treated the excess content of past statements, we would be involved in an obvious regress: we would need counterfactuals to account for the excess content of counterfactuals.

factuals. Hence, they can be used to specify the conditions under which a counterfactual is true.[10]

Let me consider here one relevant thought that is central to Woodward's approach. He takes laws to be relations that remain invariant under (a range of) actual and counterfactual interventions. If this is so, when checking whether a generalization or a relationship among magnitudes or variables is invariant we need to subject it to some variations/changes/ interventions. What changes will it be subjected to? The obvious answer is: those that are permitted, or are permissible, by the prevailing laws of nature. Suppose that we test Ohm's law. Suppose also that one of the interventions envisaged was to see whether it would remain invariant, if the measurement of the intensity of the current was made on a spaceship, which moved faster than light. This, of course, cannot be done, because it is a *law* that nothing travels faster than light. So, some *laws* must be in place before, based on considerations of invariance, it is established that some generalization is invariant under some interventions. Hence, Woodward's notion of "invariance under interventions" (2000, p. 206) cannot offer an adequate analysis of lawhood, since laws are required to determine what interventions are possible.

Couldn't Woodward say that even basic laws—those that determine what interventions and changes are possible—express just relations of invariance? Take, once more, the law that nothing travels faster than light. Can the fact that it is a law be the result of subjecting it to interventions and changes? Hardly. For it itself establishes the *limits* of possible interventions and control.[11] I do not doubt that it may well be the case that genuine laws express relations of invariance. But this is not the issue. For, the manifestation of invariance might well be the *symptom* of a law, without being constitutive of it.[12]

Before I move on I want to address an objection posed to me by a thoughtful anonymous reader. It might be that Woodward aims only to provide a *criterion* of meaningfulness for counterfactual conditionals without also specifying their truth-conditions. This would seem in order with

10. Obviously, the same holds for the Armstrong-Dretske-Tooley view of laws (see my 2002, chapter 6). If one takes laws as necessitating relations among properties, then one can explain why laws support counterfactuals and, at the same time, identify laws *independently* of this support.

11. Woodward (2000, pp. 206–7) too agrees that this law cannot be accounted for in terms of invariance.

12. I take to heart Marc Lange's (2000) recent important diagnosis: either *all* laws, taken as a whole, form an invariant-under-interventions set, or, strictly speaking, no law, taken in isolation, is invariant-under-interventions. This does not yet tell us what laws *are.* But it does tell us what marks them off from intuitively accidental generalizations.

his "pragmatic" account of counterfactuals, since it would offer a criterion of meaningfulness and a description of the "evidence conditions" of counterfactuals, which are presumed to be enough to understand causation. In response to this, I would not deny that Woodward has indeed offered a sufficient condition of meaningfulness. Saying that counterfactuals are meaningful if they can be interpreted as claims about actual and hypothetical experiments is fine (and a step forward in the relevant debate). But can this also be taken as a necessary condition? Can we say that *only* those counterfactuals are meaningful which can be seen as claims for actual and hypothetical experiments? If we did say this, we would rule out as meaningless a number of counterfactuals that philosophers have played with over the years, e.g., the pair of Julius Caesar counterfactuals considered in section 3. Though I agree with Woodward that they are "unclear," I am not sure they are meaningless. Take one of Lewis's examples, that had he walked on water, he would not have been wet. I don't think this is meaningless. One may well wonder what the point of offering such counterfactuals might be. But whatever it is, they are understood and, perhaps, are true. Perhaps, as Woodward (2003a, p. 151) says, the antecedents of such counterfactuals are "unmanipulable for conceptual reasons". But if they are understood (and if they are true), this would be enough of an argument *against* the view that manipulability offers a necessary condition for meaningfulness.

It turns out, however, that there are more sensible counterfactuals that fail Woodward's criterion. Some of these are discussed by Woodward himself (2003a, pp. 127–33). Consider the true causal claim: Changes in the position of the moon with respect to the earth and corresponding changes in the gravitational attraction exerted by the moon on the earth's surface cause changes in the motion of the tides. As Woodward adamantly admits, this claim cannot be said to be true on the basis of interventionist (experimental) counterfactuals, simply because realizing the antecedent of the relevant counterfactual is physically impossible. His response to this is an alternative way for assessing counterfactuals. This is that counterfactuals can be meaningful if there is some "basis for assessing the truth of counterfactual claims concerning what would happen if various interventions were to occur". Then, he adds, "it doesn't matter that it may not be physically possible for those interventions to occur" (2003a, p. 130). And he sums it up by saying that "an intervention on X with respect to Y will be 'possible' as long it is logically or conceptually possible for a process meeting the conditions for an intervention on X with respect to Y to occur" (2003a, p. 132). My worry then is this. We now have a much more liberal criterion of meaningfulness at play, and it is not clear, to say the least, which counterfactuals end up meaningless by applying it.

In any case, Woodward (2003a, p. 132) offers an important warning: "[I]t would be a mistake to make the physical possibility of an intervention on *C* constitutive in any way of what it is for there to be a causal connection between *C* and *E*." I take this to imply that his counterfactual approach provides an *extrinsic* way to identify a sequence of events as causal, viz., that the sequence remains invariant under certain interventions. In an earlier piece, he (2000, p. 204) stressed: "what matters for whether *X* causes [. . .] *Y* is the 'intrinsic' character of the *X*-*Y* relationship, but the attractiveness of an intervention is precisely that it provides an extrinsic way of picking out or specifying this intrinsic feature." So there seems to be a conceptual distinction between causation and invariance-under-interventions: there is an *intrinsic* feature of a relationship in virtue of which it is causal, an *extrinsic* symptom of which is its invariance under interventions.[13] If I have got Woodward right, causation has excess content over invariance-under-interventions. So there is more to causation—*qua* an intrinsic relation—than invariance-under-actual-and-counterfactual-interventions. Hence, there is more to be understood about what causation is.

To sum up. We need to be told more about the truth-conditions of counterfactual conditionals. If Woodward ties too tight a knot between counterfactuals and actual and hypothetical experiments, then it seems that counterfactual claims may reduce to claims about actual and hypothetical experiments (without any excess content). If, on the other hand, Woodward wants to insist that counterfactuals have their truth-conditions independently of their evidence-conditions, then it is an entirely open option that the truth-conditions of counterfactual assertions involve laws of nature.

3.2 Causal inference and counterfactuals

In the last twenty years, there has been an increasing interest in causal inference among statisticians and social scientists and counterfactuals have loomed large in some key attempts to model it. Prominent among them is Rubin's model, which has been advanced by Donald Rubin (1978) and Paul Holland (1986).[14] This model focuses on the discovery of the effects of causes. Suppose, to use a simple example, we want to find out whether taking an aspirin makes a difference to a *specific* subject's relief from headache. We would like to give a certain subject *u* an aspirin in order to see

13. In his (2003a, p. 125) Woodward says "there is a certain kind of relationship with intrinsic features that we exploit or make use of when we bring about *B* by bringing about *A*."

14. See also Holland (1988), Stone (1993), Cox and Wermuth (2001), Maldonado and Greenland (2002) and Kluve (2004).

what happens to the headache episode—let's call the result *Y.* But we would also like, at the same time, to withhold giving aspirin to the very same subject *u,* in order to see what happens to the headache episode—let's call this result *Y'*. The difference, if any, between *Y* and *Y'* would naturally be considered the actual causal effect of aspirin-taking on the headache episode of subject *u.* But this kind of experiment is impossible: the experimenter cannot give and *not* give an aspirin to the *same* subject *u* at the *same* time. Rubin's and Holland's main idea is that an appeal to counterfactuals allows us to make an inference about the causal effect.

Let's consider a population *U* of individuals, or units, $u \in U$. In a typical experiment, the experimenter applies one treatment, say *i,* out of a set of possible treatments *T,* to each unit *u* and observes the resulting responses *Y.* The experimental units are chosen and separated into two groups (the experimental group and the control group) by randomization. To simplify matters, let the treatment set *T* consist of two possible actions (treatment—*t,* and control—*c*). For instance, *t* may be taking the aspirin and *c* may be taking a placebo. Let, also, *Y* consist of two possible responses, e.g., headache relief—Y_t, and headache persistence—Y_c. Though it is crucial that each unit is potentially exposable to any one of the treatments, to each unit *u* just one treatment is *actually* given, i.e., either *t* or *c.* Similarly, for each unit *u,* there is just one response that is actually observed, i.e., either $Y_t(u) = Y(t, u)$ *or* $Y_c(u) = Y(c, u)$. Rubin's model defines the two responses in counterfactual terms. That is, *Y(t, u)* is the value of the response that would be observed if the unit *u* were exposed to treatment *t* and *Y(c, u)* is the value that would be observed *on the same unit u* if it were exposed to *c.* A key assumption of Rubin's model is that both values *Y(t, u)* are *Y(c, u)* are well-defined and determined. In particular, it is assumed that even if subject *u* is actually given treatment *t* and has response *Y(t, u),* there is still a fact of the matter about what the subject's *u* response would have been, had she been given treatment *c.* The task is to figure out the so-called *individual causal effect,* that is the difference

$$(3)\ \tau(u) = Y(t, u) - Y(c, u)$$

which measures the effect of treatment *t* on *u,* relative to treatment *c.*

In each particular experiment, either *Y(t, u)* or *Y(c, u)* (but not both) ceases to be counterfactual. Yet, given that one of *Y(t, u)* and *Y(c, u)* becomes observable, the other *has to* be unobservable. Holland has called a situation such as this "the *fundamental problem of causal inference.*" As he (1986, p. 947) put it: "It is impossible to *observe* the value of *Y(t, u)* and *Y(c, u)* on the same unit and, therefore, it is impossible to observe the effect of *t* on *u.*" Does it follow that figuring out (3) above is impossible?

Suppose that we give treatment *t* to *u* and we observe *Y(t, u).* The ques-

tion then is how could we possibly figure out the value of *Y(c, u)*? Recall that *Y(c, u)* is a counterfactual: the response that would be observed if the unit *u* were exposed to treatment *c* (given that it was in fact exposed to treatment *t* and the observed value was *Y(t, u)*). The important insight of Rubin's model is that when certain assumptions are in place, there are ways to assess counterfactuals such as the above. Here is how we may proceed.

Given that unit *u* got treatment *t,* we may try treatment *c* to a different unit *u'*, which is very much like *u,* except that it was given treatment *c* instead. That is, instead of assessing the counterfactual conditional *Y(c, u),* which is impossible, we assess the factual conditional *Y(c, u')*—the response of unit *u'* if she is given treatment *c*—and claim that this tells *indirectly* what the value of *Y(c, u)* is. If this move is to be plausible at all, we need an assumption of *unit homogeneity.* We need to assume that *u* and *u'* are so similar that the actual response of *u'* to treatment *c* is the same as the response that unit *u would* have to treatment *c.* Under this assumption, we take it that *Y(t, u)=Y(t, u')* and *Y(c, u) = Y(c, u').* Then, the individual causal effect can be calculated, since (3) becomes thus:

(4) $\tau(u) = Y(t, u) - Y(c, u) = Y(t, u) - Y(c, u').$

This is all fine and I am prepared to say that, *modulo* the uniformity assumption, it does tell us something about the individual causal effect. But something strange has happened. (3) involves essentially a counterfactual conditional [*Y(c, u)*]. (4) does not. (4) is indeed measurable, but the counterfactuals are gone. Instead, (4) has two factual conditionals, one for unit *u* who received treatment *t* and another for unit *u'* who received treatment *c.* In a sense, we are still asking: what would have happened to *u,* had we given her treatment *c*? But it also seems that we have now *reduced* this question to two different ones: a) what *does* happen to *u',* if we give her treatment *c*?, and b) assuming unit homogeneity, *Y(c, u')* and *Y(t, u),* what *is* the causal effect of *t* on *u*? These questions involve no counterfactuals. The content of the counterfactual conditional *Y(c, u)* seems exhausted by the joined content of the factual conditional *Y(c, u')* and the unit homogeneity assumption. In other words, the unit homogeneity assumption renders the counterfactual conditional *Y(c, u)* not so much a claim about the *specific* unit *u* but rather a claim about *any* of the homogeneous units. It is because of this fact that the counterfactual becomes testable.

There is another way we might proceed in our attempt to calculate $\tau(u)$. This time, instead of giving treatment *t* to unit *u* and treatment *c* to

(uniform) unit *u*', we give treatment *c* to unit *u* at time t_1 and treatment *t* to the *very same unit u* at a later time t_2. As Holland (1986, p. 948) notes, this move requires another assumption, viz., *temporal stability.* This, he says, "asserts the constancy of response over time." It also requires an assumption of "causal transience," since it implies that "the effect of the cause c and the measurement process that results in *Y(c, u)* is transient and does not change *u* enough to affect *Y(t, u)* measured later" (1986, p. 948). So, if my taking a placebo at time t_1 changes some properties of mine enough to affect my response to taking an aspirin at a later time t_2, the causal effect of taking aspirin on my headache episode ceases to be calculable. Under these assumptions, we take it that $Y(t_{t1}, u) = Y(t_{t2}, u)$ and $Y(c_{t1}, u) = Y(c_{t2}, u)$. If this is so, then the individual causal effect can be calculated, since (3) becomes thus:

(5) $\tau(u) = Y(t, u) - Y(c, u) = Y(t_{t2}, u) - Y(c_{t1}, u)$.

The points made about (4) can be repeated about (5) too. (5) has no counterfactuals and it seems that the content of (3)—which does involve the counterfactual *Y(c, u)—reduces* to the joined content of two factual conditionals [$Y(t_{t2}, u)$ and $Y(c_{t1}, u)$] together with the two further assumptions of causal transience and temporal stability.

I am willing to allow that I may be wrong here. That is, it might be the case that counterfactuals such as the ones we have been discussing *do* have excess content over the joint content of the relevant factual conditionals and the relevant assumptions. Still, what matters is that counterfactual conditionals can be assessed in terms of truth and falsity only when certain assumptions are in place. Those assumptions might fail. If, however, there are reasons to believe they do not, then causal inference seems quite safe. This is really an important achievement of Rubin's model. But we shouldn't lose sight of the fact that these assumptions are characteristics of *stable causal or nomological structures.*[15] Consider *unit homogeneity.* For it to hold, it must be the case that two units *u* and *u'* are alike in all causally relevant respects other than treatment status. If this is so, we can substitute *u* for *u'* and vice versa. This simply means that there is a causal law connecting the treatment and its characteristic effect which holds for all homogeneous units and hence is independent of the actual unit chosen (or could have been chosen) to test it. In effect, this holds for temporal stability too, since the latter is the temporal version of unit homogeneity. It

15. My favorite way to spell out this notion is given by Simon and Rescher (1966). In fact, in showing how a stable structure can make some counterfactuals true, they blend the causal and the nomological in a fine way.

does indeed make sense to wonder what would the value of the voltage in a resistor would have been, if the intensity of the current was I instead of the actual I_0 precisely because Ohm's law provides a stable nomological structure to address this counterfactual. But suppose we wanted to check the counterfactual that had the election taken place at an earlier time, the government would have been re-elected. Here it is obvious that temporal stability cannot be assumed because there is no stable nomological structure to back it up. Law-backed counterfactuals can indeed be assessed precisely because the laws make sure that the required assumptions are in place.[16]

In light of the above, it might not be surprising that according to Pearl (2000, p. 428), who is one of the champions of the counterfactual approach, "the word 'counterfactual' is a misnomer." In the case of individual causal effects, Pearl notes, we are interested in finding out things such as this: "Q_{II}: The probability that my headache would have stayed had I not taken aspirin, given that I did in fact take aspirin and the headache has gone."

It does not matter for present purposes that Pearl formulates the issue in terms of probabilities. What matters is that Q_{II} is a counterfactual claim of which Pearl (2000, p. 249) stresses: "[. . .] [c]ounterfactual claims are merely conversational shorthand for scientific predictions. Hence Q_{II} stands for the probability that a person will benefit from taking aspirin in the next headache episode, given that aspirin proved effective for that person in the past [. . .] Therefore, Q_{II} is testable in sequential experiments where subjects' reactions to aspirin are monitored repeatedly over time."

Nothing said so far is meant to belittle causal inference. Whether or not we view it as involving an ineliminably *counterfactual* element, we can certainly draw safe causal conclusions when the relevant assumptions are fulfilled. Actually, both the advocates of the counterfactual approach (e.g., Holland 1986; Cox and Wermuth 2001) and their opponents (e.g., Dawid 2000) agree that we can get valuable information about the so-called *average causal effect.* This is the average causal effect on the whole population, i.e., the difference between the expected value of responses to treatment t and the expected value of responses to treatment c. Indeed, randomized controlled experiments are important precisely because they let us know about average causal effects.[17] However, to get from the average causal effect in a population to the *individual causal effect* on a specific unit u, we need the further assumption of "constant effect" (Holland 1986, p. 948) or

16. It goes without saying that this causal or nomological structure should be characterized independently of counterfactuals, but as Simon and Rescher (1966) show, this can be done.

17. For some interesting (but manageable) complications, see Kluve (2004, pp. 86–7).

"unit-treatment additivity" (Cox 1986, p. 963). According to this, the effect of treatment t on each and every unit u is the same.[18] Whether this holds or not is a largely empirical matter.

To sum up, the counterfactual approach to causal inference is a big step forward. Yet, its very possibility rests on (and gets its purchase from) certain powerful assumptions (unit homogeneity, temporal stability, causal transience, constant effect). In a sense, these assumptions remove the counterfactual element from Rubin's model. But even if this is not quite right, these assumptions characterize the stable *causal or nomological structure* that needs to be in place in order for the counterfactuals to be meaningful and truth-valuable.[19]

4. Mechanisms

Recent philosophical interest in mechanisms stems from two sources. One is the recognition of the fact that scientists try to identify and understand the mechanisms that causally explain certain phenomena or underlie certain functions, e.g., the mechanism of reproduction, of gene-transmission, of chemical bonding, of face-recognition, etc. The other is a general dissatisfaction with standard philosophical views of causation, which fail to explain, or take account of, the mechanisms by which certain causes bring about certain effects.

4.1 Mechanisms and counterfactuals

Mechanisms are complex systems (or objects) that bring about a certain activity or are responsible for a certain behavior. A thermostat might be a stock example of a mechanism. A conventional thermostat works like an on-off switch. A bimetallic coil tips a small mercury-filled glass bottle. The bimetallic coil is made from two different metal strips that have been sandwiched together and then rolled into a coil. As the temperature changes, the two metals expand differently and the coil winds or unwinds. As it does, it tips the glass bottle and the mercury rolls from one end of the bottle to the other. When the mercury falls to one end, it allows an electric current to flow between two wires and the furnace turns on. When

18. In fact, the constant effect assumption is a consequence of unit homogeneity (cf. Holland 1986, p. 949). For some criticism of this assumption see Cox and Wermuth (2001, p. 68).

19. The counterfactual approach to causal inference has been severely criticized by Philip Dawid (2000) and has been vigorously defended by others (see the discussion that follows Dawid's article). Dawid has a number of important complaints. But the thrust of his critique is that the counterfactual approach relies on untestable metaphysical assumptions and in particular on a hopeless attempt to calculate the value of an unobservable quantity. Dawid's reaction, though invariably interesting, may be too positivistic.

the mercury falls to the other end of the bottle, the current stops flowing and the furnace turns off.

According to Glennan (2002, p. S344):

> (*M*) A mechanism for a behavior is a complex system that produces that behavior by the interaction of a number of parts, where the interactions between parts can be characterized by direct, invariant, change-relating generalizations.

Mechanisms, he adds, "are not mechanisms *simpliciter*, but mechanisms *for* behaviors." For the very same complex system may issue in two different behaviors (e.g., the heart is a mechanism that pumps blood and makes noise). What the mechanism *does* determines its boundaries, its division into parts and the relevant modes of interaction among these parts. Broadly understood, a mechanism consists of some parts (its building blocks) and a certain *organization* of these parts, which determines how the parts interact with each other to produce a certain output. The parts of the mechanism should be stable and robust, that is their properties must remain stable, in the absence of interventions. The organization should also be stable, that is the system as a whole should have stable dispositions, which produce the behavior of the mechanism. Thanks to the organization of the parts, a mechanism is more than the sum of its parts: each of the parts contributes to the overall behavior of the mechanism more than it would have achieved if it acted on its own. Mechanisms can be contained within larger mechanisms.

In his (1996), Glennan took his mechanistic approach to offer a rather robust solution to the problem of counterfactuals. He took laws that are mechanically explicable (in the sense that there is a mechanism that underpins them) to show in "an unproblematic way" how "to understand the counterfactuals which they sustain" (1996, p. 63). The key idea is that the presence of the mechanism (e.g., the thermostat) explains why a certain counterfactual holds, e.g., if the temperature had risen, the furnace would have turned off. Similarly, the breakdown of a mechanism would explain why certain counterfactuals fail to hold. In his more recent work (see (*M*) above), Glennan characterizes the interaction of the parts of the mechanism in terms of Woodward's invariant, change-relating generalizations, that is generalizations that remain invariant under actual and *counterfactual* interventions.

It seems then that there is a tension between Glennan's earlier and later views. According to the earlier view, mechanisms explain via mechanical laws when certain counterfactuals hold. According to the later view, it is certain interventionist counterfactuals that explain (or ground) the laws that govern the interaction of the parts of the mechanism. Consider the

thermostat: it is a mechanical law (ultimately, the law that metals expand when heated) that explains why it is the case that had the temperature been higher, the switch would have closed. But why is this a *law?* Because, had we intervened to change one magnitude (e.g., the temperature), the law (that metals expand when heated) wouldn't change and the other magnitude (e.g., the length of the metal strips in the bimetallic coil) would have changed. The tension is obvious: mechanical laws support counterfactuals and counterfactuals render mechanical laws *laws.* Though I am not sure we are faced here with a vicious circle, I am not sure either *where* it can be broken so that the described relation between mechanism and interventionist counterfactuals can get going.

A central and stable feature of Glennan's views is a distinction between the fundamental laws of physics and what he calls mechanically explicable laws. He notes, quite plausibly, that the fundamental laws of physics are *not* mechanically explicable and claims that "all laws are either mechanically explicable or fundamental, *tertium non datur*" (1996, p. 61). A mechanically explicable law is a law which is underpinned by a mechanism, or as Glennan says, which "is explained by the behavior of some mechanism" (1996, p. 62). He takes it that mechanically explicable laws characterize all the special sciences and "much of physics itself" (1996, p. 50). Glennan agonizes a lot about how exactly to formulate his view of the mechanical explication of laws, but let's leave all this to one side. I want to focus on a possible problem that this distinction creates.

If fundamental laws are *not* mechanically explicable, and if they too support counterfactuals (as they do, I suppose), it is not necessary for the truth of a counterfactual that there is a mechanical explanation of it. So, the presence of a mechanically explicable law (and hence of a mechanism) is not a necessary condition for the truth of a counterfactual conditional. Glennan agrees on this; still, he thinks it is a *sufficient* condition. Even if he is right, his theory is incomplete: if some counterfactuals are true even though a mechanism is absent, then there is more to the link between laws and counterfactuals than Glennan's theory admits. Suppose Glennan is right in taking mechanisms to underpin non-fundamental laws. He also subscribes to some kind of supervenience thesis: the non-fundamental laws supervene on the fundamental laws (cf. 1996, pp. 62 and p. 66; 2002, pp. 346 and p. 352). So on Glennan's view, non-fundamental laws are underpinned by mechanisms *and* supervene on fundamental laws, which are not underpinned by mechanisms.

Here is the problem, then. What is the relation between the mechanisms that realize the non-fundamental laws and the more fundamental laws on which the non-fundamental laws supervene? Glennan does not explain. To be sure, he (1996, p. 66) asserts: "Although the mechanism re-

sponsible for connecting two events may supervene upon other lower-level mechanisms, and ultimately on mechanically inexplicable laws of physics, it is not these laws which make the causal claim true; rather it is the structure of the higher level mechanism and the properties of its parts." But this is hardly an explanation of what is going on. One plausible thought is that the fundamental laws govern the interactions of the parts of the mechanism, which realizes the non-fundamental law. If this is so (as I think it is), then it would be odd to say that the mechanism that explains, say, Ohm's law is ultimately determined (supervenience *is* a kind of determination) by the fundamental laws that govern the interaction of fundamental particles but that these fundamental laws are not (part of) the truth-makers of Ohm's law. Once identified, the mechanism might well have explanatory and epistemic autonomy. But, if supervenience holds, the mechanism does not have metaphysical autonomy. If this line of thought is right, then the following seems reasonable. The presence of a mechanism is *part* of a (metaphysically) sufficient condition for the truth of certain counterfactuals; the fully sufficient condition includes some facts about the fundamental laws that, ultimately, govern the behavior of the mechanism. This, of course, is entirely consistent with the thought that in most practical situations when it comes to asserting the truth of a certain counterfactual, it is enough to cite the mechanism. The rest of the sufficient condition is not thereby rendered metaphysically redundant, but only explanatorily so.

There are two major attractions of Glennan's mechanistic theory. The first is that it is descriptively more adequate than the mechanistic approach of Salmon and Dowe. Both of them characterize interactions in terms of the exchange of conserved quantities. To be sure, they do aim at a mechanistic theory of *physical* causation. Still, this account is too narrow to describe cases of causation among higher-level entities. Consider, Glennan says, "a social mechanism whereby information is disseminated through a phone-calling chain" (2002, p. S346). It is surely otiose and uninformative to try to describe this mechanism in terms of exchange of conserved quantities. As we have just seen, Glennan does not deny that the interactions involved in telephone calls supervene on basic physical interactions. But he is surely right in saying that we would miss something if we tried to *explain* them in those terms. We would lose the fact that higher-level interactions form higher-level kinds. So, Glennan's mechanistic view is broad enough to account for mechanisms at levels higher than physics. The explanatory autonomy of higher-level mechanisms is, I think, a lesson that we should take to heart.

The other attraction of Glennan's mechanistic theory relates to his demand that understanding causal claims requires knowing what their

underlying mechanisms are (cf. 1996, p. 66). In fact, Glennan wants to make a stronger point, viz., that "a relation between two events (other than fundamental physical events) is causal when and only when these events are connected in the appropriate way by a mechanism" (1996, p. 56). I don't think the stronger claim is warranted. But a weaker claim is very plausible. Given its centrality to the positive argument of this paper, which will be advanced in section 5, I will postpone its discussion until then.

4.2 Mechanisms and activities

Machamer, Darden and Craver (henceforth *MDC*) claim: "Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions" (2000, p. 3). On the face of it, the *MDC* characterization of a mechanism is fairly similar to Glennan's. On closer inspection, there is a central difference. *MDC* introduce the concept of *activity* as a means to account for the interaction between the parts of the mechanism and its overall causal efficacy. The *MDC* approach is exciting, especially when it comes to the detailed description and classification of how mechanisms are taken to operate in neurobiology. But for the purposes of this paper, I will examine only the notion of activity. This notion is central to *MDC*'s mechanistic view of causation since, as they say, "activities are types of causes" (2000, p. 6) and "activities are needed to specify the term 'cause'" (2000, p. 8).

As I see it, their view is that an adequate understanding of the concept of mechanism requires an *ontological* shift: we need to accept the existence of activities on top of the usual commitments to entities, properties and processes. This unparsimonious move is recommended on the basis of their claim that mechanisms are "active": "they do things" (2000, p. 5). They think that unless activities are accepted as ontological bed-fellows of entities, properties and processes, mechanisms will be *passive:* things might be done *via* them, but not *because of* them. They also claim that appeals to causal laws, or to invariant generalizations, fail to capture the productivity of a mechanism, which "requires the productive nature of activities" (2000, p. 4).

MDC's "dualism," as they put it, requires that there is a fine distinction between entities (with their properties) and activities. But is there? As is usual in philosophy, we are first given some examples. So cases such as bonding, diffusion, depolarization, attraction and repulsion, etc. are cases of *activity.* But what do all these share in common in virtue of which they are *activities?* What we are told is that "activities are the producers of change" (2000, p. 3). But production is itself an activity. So, we are not

given an illuminating account of that which some things share in common, in virtue of which they are activities.

MDC say the following of the relation between entities and activities: "Entities and a specific subset of their properties determine the activities in which they are able to engage. Conversely, activities determine what types of entities (and what properties of those entities) are capable for being the basis of such acts. [. . .] Entities and activities are correlatives. They are interdependent" (2000, p. 6). It follows that entities and activities are ontically on a par: they determine each other. They say this more explicitly when they claim that "(t)here are no activities without entities, and entities do not do anything without activities" (2000, p. 8).

I think the supposed ontic parity between entities and activities is wrong-headed. First, it's conceivable that there are entities without activities. Indeed, there may be entities capable of engaging in certain activities, but the prevailing circumstances, or the laws of nature, may be such that they *fail* to engage in these activities. (Apropos, if what matters is the ability of an entity to engage in an activity and not the actual occurrence of this activity, then it is clear that *MDC* have to rely on counterfactuals to illuminate the link between entities and activities.) Second, I cannot see how activities can determine what *types* of entities can engage in them. There may well be an open-ended list of types of objects that can engage in some activity, and they may share very little, if anything, in common. Take the activity of *playing*. It's hard to say that it determines what kinds of entities (and what properties) are involved in this activity. Admittedly, this is a case of a highly generic activity and it might be problematic precisely because of this. There are cases of more specific activities, where the activity is performed by certain *types* of objects. It then might *seem* that the activity does determine what types of object can engage in it. An example of such a specific activity might be the activity of *pushing*. It seems that this activity determines that the objects involved in it must have certain properties, e.g., rigidity, bulk etc. But I think appearances are deceptive. Epistemically, we might first classify a certain type of activity and then identify what kinds of objects engage in it. But from this it does not follow that this is the order of ontic dependence too. On the contrary, objects can engage in certain activities *because* they have certain properties and not the other way around.

Consider the activity of chemical bonding. Does this activity determine that entities that engage in it must have a certain electronic structure? Not really. Chemical bonding could not exist without some entities having the right electronic structure. So not only are the latter presupposed ontically for the activity, but also they fully *determine* this activity: the activity of bonding *consists* in the fact that certain entities with certain elec-

tronic structure behave in a certain way when they are in proximity. The dependence of the activity on the properties of entities becomes clear when the activity *fails* to take place. Consider the case were chemical bonding does not take place, e.g., the case of noble gases. There, you have the entities without the activity of chemical bonding precisely because the entities and their properties determine that a certain activity *cannot* take place. The situation is exactly symmetrical when the activity *does* take place.

The conclusion I draw is that activities cannot be ontically on a par with entities. But one may wonder: why should *MDC* want to hypostatize activities? Why isn't it enough to talk in terms of entities and their properties? *MDC* are right in protesting against process-theorists that entities are indispensable in understanding mechanisms. They rightly claim that the program of reducing entities to processes is "problematic at best." But they also want to argue against "substantivalists," that is, those who "confine their attention to entities and properties, believing that it is possible to reduce talk of activities to talk of properties and their transitions" (2000, p. 4). Against them, *MDC* claim that entities and their properties are not enough for the characterization of mechanisms: activities are also required. Now, the substantivalists that *MDC* have in mind take the properties of the entities to be dispositional; they equate them with *capacities* or *active powers.* This is a quite powerful ontology. The friends of active powers would surely protest that given that active powers are granted to entities, talk of activities as *distinct* from these powers is redundant.[20]

MDC offer two arguments for activities on top of capacities. I think they are both problematic. The *first* argument[21] is this: "in order to identify a capacity of an entity, on must first identify the activities in which that entity engages" (2000, p. 4). Even if right, this is irrelevant. It only raises an epistemic point: we cannot know what capacities an entity has, unless we first know what it *does.* From this, it does not follow that activities are ontically on a par with capacities. Nor does it follow that it is not the capacities of an entity that determine what activities it engages in. Quite the contrary. To use their own example, it is *because* aspirin has the capacity to relieve headaches (a capacity which we take it to be grounded in its chemical composition) that aspirin engages in this activity, i.e., headache-relieving. If capacities are granted, then activities supervene on them. And this remains so, even if, from an epistemic point of view, we

20. Consider how Harre (2001, p. 96) understands an active power: "a native tendency or inherent capacity to act in certain ways in the appropriate circumstances." Activities come for free if Harre is right. Note that Harre, too, favors a mechanistic account of causation.

21. The essence of this argument is repeated in Machamer (2003).

need to attend to the (observed) activities in order to conjecture about the capacities.

The *second* argument that *MDC* offer is this: "state transitions have to be more completely described in terms of the activities of the entities and how those activities produce changes that constitute the next change" (2000, p. 5). Here the emphasis is on the *production.* As they explain, activities add the "productivity" by which changes in properties (state-transitions) are effected. But isn't this question-begging? Many would just deny that there is anything like a productive continuity in state transitions. All there is, they would argue, is just regular succession (or some kind of dependence). In any case, the friends of capacities would argue that there is productive continuity in state transitions, but that this is grounded in the natures of the entities engaged in state transitions. If water has the capacity to dissolve salt, and if this capacity is grounded in the natures of water and salt, then all that is needed for the dissolution of salt in water (that is, the activity) is that the circumstances are right and the two substances are brought into contact.

I have a final, but central, objection to *MDC:* they cannot avoid counterfactuals. Counterfactuals may enter at two places. The first is the activities themselves. Activities, such as bonding, repelling, breaking, dissolving etc., are supposed to embody causal connections. But, one may argue that causal connections are distinguished, at least in part, from non-causal ones by means of counterfactuals. If "x broke y" is meant to capture the claim that "x *caused* y to brake," then "x broke y" must issue in a counterfactual of the form "if x hadn't struck y, then y wouldn't have broken." So talk about activities is, in a sense, disguised talk about counterfactuals. The second entry-point for counterfactuals is the characterization of interactions within the mechanism. We have already seen (section 4.1) Glennan insisting that this interaction should be captured in terms of the invariance of the relationships among the parts of the mechanism under actual and counterfactual interventions. *MDC* are not quite clear on what the interaction within the mechanism consists in. Note that it wouldn't help to try to explain the interaction between two parts of a mechanism (say parts *A* and *B*) by positing an intermediate part *C.* For then we would have to explain the interaction between parts *A* and *C* by positing another intermediate part *D* and so on (ad infinitum?).

I take this to be a crucial problem of the mechanistic approach. In a sense, this approach fills in the 'chain' that connects the cause and the effect with intermediate loops. But there is still no account of how the loops interact. Here, it might well be the case that the most general and informative thing that can be said about these interactions is that there are

relations of counterfactual dependence among the parts of the mechanism. Even if we posited activities, as *MDC* do, we would still need counterfactuals to make sense of them, as we have just seen. In any case, if I am right, there is more to causation than mechanisms.

5. *Both* mechanisms and counterfactuals are helpful

We can sum up the central claim of the paper as follows. There is an asymmetry between the two accounts we have been discussing: mechanisms need counterfactuals; but counterfactuals do not need mechanisms. In other words, mechanistic causation requires counterfactual dependence but not conversely. It is in this sense, that the counterfactual approach is more basic than the mechanistic.[22,23]

Recall, however, what was noted at the end of section 4.1, viz., that the *understanding* of causal relations requires understanding of the underlying mechanisms. Is this *really* so?

Imagine a perfectly randomized experiment in which *t* (for treatment) produces higher response than *c* (for control). Has a causal connection been established? If we treat the randomized experiment as a black box, then in so far as it is a good experiment, we have established a causal connection. But what is inside the *black box?* Some might think that without a specification of the mechanism by which the higher response *t* was effected, the causal connection has *not* been established.[24]

This is a delicate issue. As I noted in the end of the last section, establishing the causal status of each part of a mechanism would require finding out (or estimating) its causal effect. And the best way to do this is by non-mechanistic means, and in particular by means of the counterfactual approach outlined in section 3.2. So, there seems to be a genuine asymmetry here. The causal effect can be found out, at least in favorable circumstances, *without* understanding the causal mechanisms, if any, in-

22. As an anonymous reader pointed out, it may be that the causal relation between the parts of the mechanism may not be captured by means of a relation of counterfactual dependence. If this were so, then the symmetry between mechanisms and counterfactuals would be restored. This is indeed possible. Yet, as far as I can tell, the asymmetry between mechanistic accounts and some kind of dependence account of causation (even if this is not counterfactual dependence) would remain.

23. This is not to say, of course, that a counterfactual dependence account of causation is all there is to causation. Though I might have already said enough to persuade you of this, the interested reader should look at Hall (forthcoming) for further discussion.

24. Notably, this is the view of D. R. Cox (1992, p .297). He claims that this was also R. A. Fisher's view. When asked, at a conference, for his view on the step from association to causation, Fisher is reported to have responded: make your theories elaborate (cf. Cox 1992, p. 292).

volved; but the causal mechanisms, even if they are present, cannot be understood without the notion of the causal effect, that is without some notion of (counterfactual) dependence.

But there are at least three things that show how mechanistic considerations can help the counterfactual approach to causal inference. *First,* mechanistic considerations can help testing the stability assumptions (unit homogeneity, temporal stability) that are necessary for the counterfactual inference. I take this to be fairly obvious, so I won't elaborate on it further.

Second, mechanistic considerations can help deal with the endogeneity problem. Briefly put, the problem of endogeneity is this. It might happen that the values taken by the so-called explanatory (or causal) variable, are consequences, rather than causes, of the values of the dependent variable. In a perfectly controlled experiment this cannot happen because the variables that are manipulated are the explanatory variables. But in cases where the research is qualitative, or where an experiment is not possible at all, the counterfactual approach might well fail to solve the endogeneity problem. Consider one of the classic problems of the early twentieth century social science: Max Weber's claim a certain type of economic behavior—the capitalist spirit—was induced by the Protestant ethic. Many social scientists have argued that this claim falls foul of the endogeneity problem. Opponents of Weber's Thesis claimed that the order of dependence goes the other direction: Europeans who already have had an interest in breaking free of the pre-capitalist mode of productions might have broken free of the Catholic Church precisely for that purpose. That is, it was the economic interests of certain groups that caused the Protestant ethic and not conversely. In cases such as this, a controlled experiment is out of the question. Besides, the assessment of intuitively relevant counterfactuals will be, to say the least, precarious. But an understanding of the mechanisms at play can well help resolve the endogeneity problem. These mechanisms, I presume, include a more detailed description of the explosion of the capitalist economic activity in the sixteenth century and of the economic behavior of certain groups, e.g, in Venice and Florence or in England and Holland, which predate the emergence of Protestantism.

The *third* way in which mechanistic considerations can help the counterfactual approach concerns the possible confounders. In a perfectly randomized trial, the problem of confounding variables does not arise. The experimental method itself makes it very unlikely that the explanatory variable is correlated with possible confounders. But in qualitative research, or even when matching techniques are used, it is possible that the explanatory variable is correlated with a confounding variable. Take, for instance, the dependent variable to be participation in demonstrations and the ex-

planatory variable to be the age of the participants. It might well be that a confounding variable (e.g., radicalness of beliefs) is correlated to the explanatory variable and has an influence on the dependent variable. In cases such as this, knowledge of mechanisms can help identify possible confounders and control for them. Conversely, knowledge of mechanisms can explain why the experimenter need not control for some variables (e.g., the color of the eyes of those who participate in demonstrations).

Mechanisms cannot be the surrogate of a careful experiment. But we *needn't* see them as a surrogate. Both counterfactuals and mechanisms can work together to secure some causal knowledge. If we think of an experiment as a black box, then counterfactuals have a key role to play. After all, when certain assumptions hold, they can establish a causal relation. But without some knowledge of the mechanism inside the black box, we won't have *full* understanding of the causal relation. Nor can we solve, at least as effectively, some methodological problems of causal inference.

Using the black box carefully does establish a causal link. *Looking into* the box does offer extra understanding, even if the mechanism does *not,* in and of itself, constitute the causal link. In either case, in so far as this link (Hume's *secret connexion*) is an intrinsic relation between cause and effect, we will get a glimpse of it, but not much more.

References

Bogen, Jim. 2003. "Analyzing Causality: The Opposite of Counterfactual is Factual." http://philsci-archive.pitt.edu/archive/00000797/.

Cox, D. R. 1986. "Comment." *Journal of the American Statistical Association,* 81: 963–4.

———. 1992. "Causality: Some Statistical Aspects." *Journal of the Royal Statistical Society,* A, 155: 291–301.

Cox, D. R and Nanny Wermuth. 2001 "Some Statistical Aspects of Causality." *European Sociological Review,* 17: 65–74.

Dawid, Philip. 2000. "Causal Inference Without Counterfactuals." *Journal of the American Statistical Association,* 95: 407–24.

Dowe, Phil. 2000. *Physical Causation.* Cambridge: Cambridge University Press.

Glennan, Stuart. 1996. "Mechanisms and the Nature of Causation." *Erkenntnis,* 44: 49–71.

———. 2002. "Rethinking Mechanical Explanation." *Philosophy of Science,* 69: S342-S353.

Hall, Ned. forthcoming. "Two Concepts of Causation." In *Counterfactuals and Causation.* Edited by John Collins, Laurie Paul and Ned Hall. MIT Press.

Harre, Rom. 2001. "Active Powers and Powerful Actors." In *Philosophy at*

the New Millennium. Edited by A. O'Hear. Cambridge: Cambridge University Press.
Holland, Paul. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association,* 81: 945–60.
———. 1988. "Comment: Causal Mechanism or Causal Effect: Which is Best for Statistical Science?" *Statistical Science,* 3: 186–8
Hume, David. (1777) 1975. *Enquiries Concerning Human Understanding and Concerning the Principles of Morals.* Edited by L. A. Selby-Bigge. Third revised edition by P. H. Nidditch. Oxford: Clarendon Press.
Kluve, Jochen. 2004. "On the Role of Counterfactuals in Inferring Causal Effects." *Foundations of Science,* 9: 65–101.
Kitcher, Philip. 1989. "Explanatory Unification and Causal Structure." Pp. 410–505 in *Minnesota Studies in the Philosophy of Science,* Vol. 13. Minneapolis: University of Minnesota Press.
Lange, Marc. 2000. *Natural Laws in Scientific Practice.* Oxford: Oxford University Press.
Lewis, David. 1973. *Counterfactuals.* Cambridge MA: Harvard University Press.
———. 1986. "Causation." Pp. 159–213 in *David Lewis's Philosophical Papers.* Vol. II. Oxford: Oxford University Press.
Machamer, Peter, Lindley Darden, and Carl Craver. 2000. "Thinking About Mechanisms." *Philosophy of Science,* 67: 1–25.
Machamer, Peter. 2003. "Activities and Causation." http://philsci-archive.pitt.edu/archive/00000864/.
Mackie, J. L. 1974. *The Cement of the Universe: A Study of Causation.* Oxford: Clarendon Press.
Maldonado, George and Sander Greenland. 2002. "Estimating Causal Effects." *International Journal of Epidemiology,* 31: 422–429.
Pearl, Judea. 2000. "Comment." *Journal of the American Statistical Association,* 95: 428–31.
Psillos, Stathis. 2002. *Causation and Explanation.* Chesham: Acumen.
Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *The Annals of Statistics,* 6: 34–58.
Salmon, Wesley. 1984. *Scientific Explanation and the Causal Structure of the World.* Princeton: Princeton University Press.
———. 1997. *Causality and Explanation.* Oxford: Oxford University Press.
Simon, Herbert A. and Nicholas Rescher. 1966. "Cause and Counterfactual." *Philosophy of Science,* 33: 323–40.
Stone, Richard. 1993. "The Assumptions on which Causal Inferences Rest." *Journal of the Royal Statistical* Society, B 55: 455–66.

Woodward, James. 1997. "Explanation, Invariance and Intervention." *Philosophy of Science,* 64: S26-S41.
———. 2000. "Explanation and Invariance in the Special Sciences." *The British Journal for the Philosophy of Science,* 51: 197–254.
———. 2002. "What is a Mechanism? A Counterfactual Account." *Philosophy of Science,* 69: S366-S377.
———. 2003a. *Making Things Happen: A Theory of Causal Explanation.* New York: Oxford University Press.
———. 2003b. "Counterfactuals and Causal Explanation." http://philsci-archive.pitt.edu/archive/00000839/.