Comments on Woodward, *Making Things Happen*

Michael Strevens

Draft of February 2008

I welcome the chance to revisit my review of *Making Things Happen* and the book itself (Woodward 2003; Strevens 2007) in the light of Woodward's response to the review (henceforth "Woodward's reply"). In what follows I hope to provide a surer footing for some of the claims made in the original review, to address the new definition of unrelativized causation proposed in Woodward's reply, and to expand my own appreciation of the manifold aims of the book, while expressing some doubt as to whether they can be realized by a single system of definitions. I have more to say than I can reasonably fit into this piece; I hope that my silence on some matters will not be interpreted as a concession. (In particular, I am fairly sure that what Woodward has to say about explanation implicitly commits him to what he calls "potentialism".)

1. Preemption

In a preemption scenario, there is an actual cause and a backup cause of the effect of interest; had the actual cause not occurred, the backup cause would have anyway ensured that the effect occurred. My review argued that Woodward's manipulationist account of singular causal claims, encapsulated in the condition AC (*Making Things Happen*, p. 77), will run into trouble in some cases where the backup cause is also an actual cause, incorrectly judging like some varieties of the counterfactual account that the actual cause is not in fact a cause. After giving a perhaps overly colorful and underly diagrammable illustration of this assertion, I claimed that similar counterexamples could be

constructed from far more austere materials – and more importantly, in the light of Woodward's worries in §2 of his reply, from material more conducive to the box and arrow representation of causal relations. Let me make good on this promise.

You have two variables, *X* and *Y*. The first, *X*, can take two values: 0 or 1. The second, *Y*, can take three values: 0, 1 or 2. When *Y* is either 1 or 2, it is sufficient to cause the event of interest, *Z*'s taking a value *z*, which I will call *e* for short. When *Y* is 0, it has no causal consequences. When *X* takes value 1, it is sufficient to cause *Y* to take value 2; when it is 0, it has no causal consequences. The gross structure of these causal relationships is captured by the directed graph shown in figure 1.

Figure 1: Setup for preemption

Suppose that neither *X* nor *Y* has any effects unless a certain background condition obtains; you can think of this as, say, the electricity's flowing. It is the beginning of such a flow that will trigger the sequence of events in which you are interested, a sequence that ends in the occurrence of the event *e* (i.e., *Z*'s taking on the value *z*). You can represent the electricity explicitly if you like; it makes no difference to the argument.

In the beginning, *X* has value 1 and *Y* has value 0. (*Z* has some value other than *z*, so *e* has not yet occurred.) The electricity is switched on. *X*'s being 1 causes *Y* to take the value 2, which in turn causes *e*.

One more thing: when the electricity is flowing, *Y*'s 0 state is inherently unstable. Had *X* not flipped *Y*'s value, *Y* would have spontaneously taken on the value of 1 and thus caused *e* anyway. You therefore have a case of preemption: *X*'s being 1 was a cause of *e*, but even if *X* had not been 1, *e* would have occurred.

What does the manipulation account say about *X*'s causal status? In making its judgment, it looks to the consequences of picking a causal path between *X* and *Z*, and of switching *X* to 0 (by an intervention, of course) while keeping all

variables off the causal path at their actual values. If there is a path for which this switching of *X* has the result that *e* does not occur, then *X*'s state is an actual cause of *e*; otherwise, not.

Now of course in my example there is only one path between *X* and *Z*, and no variables that do not lie on that path. In particular, the variable that serves as a backup cause, *Y*, lies on the path. Thus when applying AC, the value of *Y* is not held fixed, but is allowed to vary as the value of *X* is manipulated. But if *Y* is allowed to vary, then the value of *X* does not matter: once the current begins to flow, *e* will occur whether *X* is 0 or 1. Thus Woodward's condition AC will wrongly judge that *X*'s state is not a cause of *e*.

The manipulationist has a possible comeback, namely, that it is intuitively correct to hold that *X*'s state is not a cause of *e*. I am confident that the scenario can be specified in such a way as to avoid this worry: stipulate that *Y*'s being 1 causes *e* in a very different way from *Y*'s being 2 – perhaps because different fundamental laws apply in each case. Then the scenario is much like that in which I drop a vase off the roof and you smash it with a baseball bat on the way down: your swinging the bat counts as a cause of the vase's breaking even though the vase would anyway have broken when it hit the ground a moment later.

2. The Question of Relativity

Intervention Relativized In *Making Things Happen*, Woodward defines a notion of contributing causation that is a "three place" relation, connecting a cause, an effect, and a set of variables: one event or variable is a contributing cause of another only relative to a set of variables. He then goes on to define an intervention in terms of relations of contributing causation. For example, if a variable *I* is to count as an intervener on a variable *X*, for the purposes of assessing whether *X* is a contributing cause of *Y*, then among other things, *I* must not be a contributing cause of *Y* via some causal pathway that does not go through *X*. As I have just stated this necessary condition, it makes no explicit reference to a variable set. But because it invokes facts about contributing

causation, and these facts are by definition relativized to a variable set, in its underlying structure the relation of intervention is after all relativized to a variable set. When Woodward says that *I* must not be a contributing cause of *Y* by some non-*X* involving pathway, this is shorthand for: *I* must not be a contributing cause of *Y relative to V* by some non-*X* involving pathway. What determines the identity of *V*? In my review I assumed that it is the same set of variables relative to which the question whether *X* is a contributing cause of *Y* is being asked. So: when Woodward says in his reply that "there is no explicit or obvious relativization to a variable set" in his definition of intervention (§4), he speaks the literal truth, but misleadingly: once you follow up the consequences of his network of definitions, it turns out that there is an implicit and perhaps not-so-obvious relativization to a variable set.[1]

Where Woodward and I have a genuine disagreement, I think, is not on the question whether the property of being an intervention is formally relativized to a variable set, but on whether this has substantial and unwelcome material consequences, and in particular, whether the relativization makes the facts about contributing causation " relativistic" in the more loose and popular sense that what causes what depends on your perspective (more exactly, on the variable set singled out by your perspective).

Or at least, that is our disagreement about the definitions proposed in *Making Things Happen*. In his reply (§7), Woodward proposes a substantial amendment to the *Making Things Happen* definitions: he removes the relativization of contributing causation to a variable set, and so removes the relativization of intervention, total causation (see note 1), and so on, to a variable set. There is no longer any prospect whatever of relativism, radical or otherwise. In what follows I will explore the workings of the new definitions, arguing that they suffer from a serious deficiency also found in the old definitions.

Contributing Causation Derelativized Woodward's technique for derelativizing contributing causation is straightforward:

1. For the same reason, any other relation that Woodward defines partly in terms of the facts about interventions is also relativized to a variable set, including the relation of total causation, notwithstanding his claims to the contrary.

1. Retain, as a means to an end, the relativized relation originally called contributing causation. Give it a new name to avoid confusion: Woodward calls it *representation as a contributing cause*. I will call it *relativized causation*.

2. Define contributing causation as follows: X is a contributing cause of Y if there exists some variable set relative to which X is a *relativized* cause of Y.

This yields what I called for in my review: an account of unrelativized intervariable causal claims.

The new definition of contributing causation might strike you, on the surface, as possibly too liberal. Contributing causation requires relativized causation only with respect to a single variable set. What if X is a relativized cause of Y with respect to just a single variable set, and further, a variable set that strikes us as rather meager or deficient? What if, that is, there is a variable set relative to which X is a cause of Y, but this causal relation disappears if you add to or otherwise augment the set to make it a richer representation of causal reality? Surely we would take this as a sign that X is not really a cause of Y. But Woodward's definition of unrelativized causation will count it as such.

Woodward is aware of the problem, but holds that it is illusory on the grounds that relativized causation is in a certain sense *monotonic*: if X is a relative cause of Y with respect to a variable set V, then it is also a relative cause of Y with respect to any superset of V. Adding variables to a set can expose causal relations that were previously hidden, but it cannot hide causal relations previously exposed.

This is a neat solution to a problem articulated in my review. Searching for an unrelativized notion of contributing causation, I pointed to a strategy with obvious appeal: count X as an (unrelativized) cause of Y if it is a relativized cause with respect to *all* the variables there are, that is, with respect to all of causal reality. Woodward cannot avail himself of this strategy, however, because no one of his causal representations is capable of representing all of causal

reality, even in a bounded locale.[2] The unrelativized definition of contributing causation proposed in his reply nevertheless manages, in a straightforward and appealing way, to do justice to the intuition that the relativized causation that matters is causation with respect to all of causal reality, without requiring that all of causal reality be represented. It is surely the right move for Woodward to make – but it is viable only if relativized causation is indeed, as Woodward claims, monotonic. I will point to a reason to doubt monotonicity.

Is Relative Causation Monotonic? My claim is that, in violation of monotonicity, adding variables to a variable set can sometimes make relativized causal relations disappear.[3] Here is an overview of the argument:

1. Adding variables to a variable set can sometimes make relativized causal relations appear (as monotonicity allows).

2. A variable's counting as an intervener depends on the *non-existence* of certain relations of relativized causation.

3. Thus (from (1) and (2)), variables may lose their status as interveners as variables are added to the variable set.

4. A variable's status as a relativized cause requires the existence of an intervener with respect to which a certain further condition is satisfied. If a variable loses its status as an intervener, then, other variables may lose their status as relativized causes.

5. Thus (from (3) and (4)), variables may lose their status as relativized causes as variables are added to the variable set.

The argument depends, you will note, on the assumption that facts about relativized causation are to be assessed with respect to facts about relativized

2. This is because Woodward's causal graphs contain as an artifact "shortest causal links", but in worlds like ours, there are no genuine shortest links; the graphs therefore omit whatever lies "inside" such links (see p. 243 of my review).

3. This talk of appearance and disappearance is of course figurative: the relative causal relations do not change. When I say that adding a variable to a set makes a relativized relation disappear, I mean that the relation holds relative to the original set but not to the augmented set.

intervention – which is to say, that the relations of causation that appear in the definition of an intervention are to be interpreted as relations of relativized causation for the purposes of defining relativized causation. This assumption will be examined below.

As stated, the argument does not establish definitively that relativized causation is not monotonic – it might be that things are set up so that the possibility envisaged in the argument never actually arises. Let me therefore give an example in which it does arise, illustrating along the way the truth of the premises.

I will start by showing that stripping a variable of its status as an intervention can destroy a relativized causal relation (as in (4) above). This is simply a recap of a scenario sketched in the original review. Suppose that consumption of salty food causes an increase in the consumption of bottled water. Suppose also that it causes an increase in the probability of heart disease. I want to use Woodward's definition to determine whether the consumption of bottled water causes heart disease. Looking to the criteria laid down by the definitions, I ask myself what would happen if I used an intervention to raise someone's consumption of bottled water. If the answer is that their chance of heart disease increases, then bottled water consumption counts, for the manipulationist, as a cause of heart disease.

An inept experimenter might manipulate their subjects' water consumption by feeding them salty food. Because the salty food increases the risk of heart disease, the experiment will find a correlation between water consumption and heart disease. If the experimenter persists in interpreting the salt-driven manipulation of water consumption as a genuine intervention, they will conclude, mistakenly, that drinking bottled water causes heart disease.

This cautionary tale illustrates the following conditional fact about relativized causation: if the salt-driven manipulation of water consumption were to count as an intervention relative to one variable set, but not to count as an intervention relative to a strictly larger variable set – so that adding one or more variables to the original set removed its status as an intervention – then water consumption would count as a cause of heart disease relative to the smaller variable set but not relative to the larger variable set, in violation of

monotonicity. Could such a situation ever arise?

Intuitively, what disqualifies salty food as an intervention on water consumption, for the purposes of determining the causal relationship between water and heart disease, is that salty food has an independent effect on heart disease. It is the existence of a causal link between salty food and heart disease, a link that does not go by way of water, that shows that the administration of salty food is what you might call a *spurious intervention* on water consumption.

A problem exists for Woodward's monotonicity thesis if this link only shows up relative to larger variable sets, for then the salty food intervention will count as spurious only relative to larger variable sets, and so the causal relation between water consumption and salty food will exist with respect to small variable sets but will disappear as variables are added.

My next topic, then, will be the circumstances under which relativized causal links appear when variables are added to a variable set (see (1) above), and in particular, the question whether a link between salty food consumption and heart disease might not exist relative to a small variable set, but might appear when variables are added to the set.

Augmenting a variable set will reveal new relations of relativized causation when a variable X has both a positive and a negative effect on another variable Y, and the two effects cancel out. Suppose, for example, that consumption of salty food causes both red wine drinking and hardening of the arteries. Hardening of the arteries increases the chance of heart disease, but red wine drinking (so we all hope) decreases it. The relevant causal relationships in such a world are as shown in figure 2.

Imagine a world under the jurisdiction of a more benevolent god than in the previous example, a world in which the healthful effect of red wine is sufficiently strong to counterbalance exactly the effect of artery hardening. Apply Woodward's criterion for relativized causation between salty food consumption and heart disease relative to two variable sets. The first set contains just two variables, namely, the variables whose causal relationship is under investigation: salty food consumption and heart disease. Woodward's criterion for causation relative to a variable set V is as follows: you select a causal path between putative cause and effect, holding all variables that are in V but not on the path

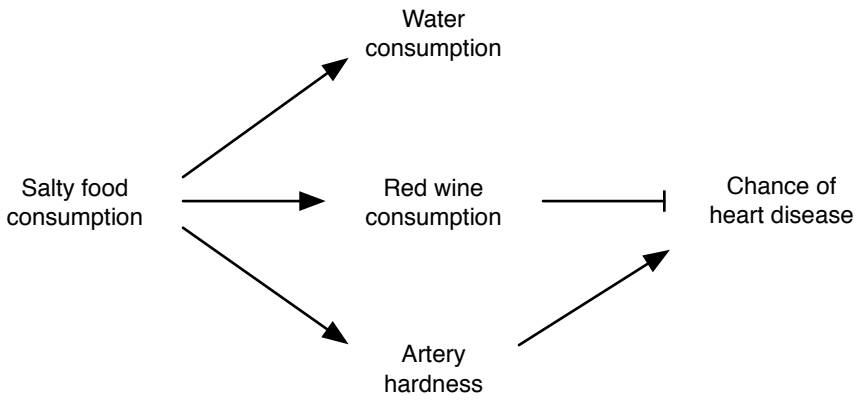

Figure 2: Consuming salty food promotes heart disease along one causal pathway (arrow) but inhibits it along another (bar). Water has nothing to do with heart disease.

to some particular value, and then ask what would happen if you intervened on the putative cause. If the value of (or the probability distribution over the values of) the putative effect variable would change as a result, then there is indeed a causal relationship between the two variables. If the test fails for all causal paths, there is no causal relationship. In the case at hand, since only the putative cause and effect are in the variable set, no variables will be held fixed. The question is simply: if you alter your level of salty food consumption, will your chance of heart disease change? Because of the balance between the effects of wine and hardened arteries, the answer is no. Thus relative to this variable set, salty food does not cause heart disease.

Next, apply the criterion relative to a variable set that contains the two additional causal factors shown in figure 2, red wine consumption and artery hardness. You now have variables to hold fixed. In particular, as part of the test, you might hold red wine consumption fixed while you intervene on salty food consumption. The chance of heart disease will increase, and so salty food consumption will count as a (positive) causal influence on heart disease. (Holding artery hardness fixed will show that it is also a negative causal influence.) In short, salty food is a cause of heart disease relative to a

variable set containing all four variables shown in figure 2 (ignore water), but not relative to a set containing only salty food consumption and heart disease.

The story so far: I have considered two causal scenarios. In one scenario (the second), adding variables exposes salty food consumption's status as an independent cause of heart disease, so undercutting its status as an intervention on water consumption. In the other scenario (the first), undercutting salty food consumption's status as an intervention on water consumption erases a relativized causal relation between water consumption and heart disease. Does this add up to trouble for monotonicity, the thesis that adding variables cannot dispel relativized causal relations? Not yet, as the two scenarios make slightly different assumptions about the causal relation between salty food and heart disease. But they are easily integrated, as follows. Suppose that consuming salty food sometimes causes the consumption of red wine and sometimes causes the consumption of bottled water. Although a salty food eater does not always choose to slake their thirst with red wine, however, they do so often enough that salty food has no net effect on heart disease: the effects of even sporadic red wine consumption balance out the effects of artery hardening.

Consider the effect of using salty food to manipulate water consumption, in order to evaluate the causal relation between water intake and heart disease. The manipulation is not always successful: some episodes of salt consumption will be followed by red wine consumption rather than water consumption. We are interested only in successful manipulations. Thus we will take into account only those cases in which salty food consumption does in fact result in water consumption. But in this subset of cases, there is no red wine consumption; thus, there will be within this subset a correlation between salty food consumption and heart disease. It follows that when salty food is used to manipulate water consumption, there is a correlation between water consumption and heart disease. The spuriousness of this intervention does not show up when it is evaluated relative to a set of variables that does not contain either red wine or artery hardness, because relative to such a set, the causal relation between salt consumption and heart disease does not show up.[4] So relative to such

4. It is important here that any genuine intervention on salt consumption will result in a set of cases in which sometimes water, sometimes wine, is consumed – and thus in which there will

an impoverished variable set, water consumption counts as a cause of heart disease.

This has two related, disastrous consequences for Woodward's account. First, because the causal relation between water consumption and heart disease goes away when further variables are added to the set (exposing the causal link between salt consumption and heart disease, and so the spuriousness of salt consumption as an intervention), relative causation is revealed to be non-monotonic. Second, because Woodward counts one variable as a cause of another provided that there is at least one variable set with respect to which the one is a relative cause of the other, he is committed to counting water as a cause of heart disease.

A Way Out? How might a manipulationist respond to the difficulties I have presented for the derelativization of causation? Perhaps the most Woodwardian reply is as follows.[5]

The proper derelativization of causation will take the form, I have been assuming, of a definition of unrelativized causation citing in its definiens only relativized causal facts. That is, any reference to causation, explicit or implicit, on the definiens side of the definition should be interpreted as a reference to relativized causation. When unpacking the definition, then, everything is relativized: the causal connections invoked by the definition are relativized causal connections; the interventions invoked by the definition of these relativized causal connections are relativized interventions (in the sense explained above), and the further causal connections invoked in the definition of these

---

be no correlation between salt consumption and heart disease.

5. An alternative is to require, in order for one variable to count as a cause of the another, that intervening on one be correlated with a change in the other (holding off-path variables fixed) for *every possible means of intervention*, that is, for any non-spurious choice of intervention variable – rather than for *some* choice of variable. Since there are many interventions on water consumption that have no effect on heart disease, water will not, then, count as a cause of heart disease. Such a move raises an entirely different kind of worry about Woodward's definitions, that they are too conservative: for some X that is (intuitively) a cause of Y and any variable set V, there may be a variable I that counts as an intervention relative to V, but whose manipulations of X are not correlated in any way with changes in Y. Why? There is a *negative* causal link between I and Y that exactly balances the effect of X, but that does not show up relative to V. I will have to leave the question whether there can be such variables for arbitrarily large variable sets to another time.

interventions are more relativized causal connections. In particular, the causal relation in virtue of which salty food consumption stands to be disqualified as an intervention on water consumption – the causal relation between salty food and heart disease – is a relativized causal relation. (I further assume that all relativization is to the same variable set.)

It is, of course, the relativity of this last relation that undermines the monotonicity of relative causation: when enough variables are added to the relevant set that the causal relation between salty food and heart disease appears, the causal relation between water and heart disease disappears. What, then, if the salt/heart disease relation were derelativized? What if, in other words, for the purposes of determining what is and is not an intervention, only absolute, unrelativized causal relations matter? Suppose (though see below) that on Woodward's definitions, salty food is a contributing cause, in the unrelativized sense, of heart disease. Then in an equally unrelativized sense, manipulating water consumption by administering salty food is not a genuine intervention. If these unrelativized facts are the facts about interventionhood that matter for determining the facts about relativized causation, then under no circumstances – with respect to no set of variables – will water be a relativized cause of heart disease. More generally, there is no scenario under which adding variables to a variable set will "hide" relations of relativized causation, and so relative causation will be monotonic after all.

What are the disadvantages, if any, of such a move? There are two. First, it is far from clear that an account of unrelativized causation that takes the form of a definition invoking unrelativized causal facts constitutes a genuine derelativization. The existence of what was to be constructed appears to have been assumed. Second, although salty food may be, intuitively speaking, an absolute contributing cause of heart disease, Woodward is not thereby entitled to count it as such: he must earn the right by showing that his account of unrelativized causation does indeed recognize its existence.

Let me pursue this second point a little further. On Woodward's account, salty food is an (unrelativized) contributing cause of heart disease just in case there is a set of variables with respect to which it is a relativized cause, which is to say, just in case there is a set of variables and a causal path between salty

food consumption and heart disease such that holding the variables off the path constant and intervening in some way on salty food consumption changes the probability of heart disease. Whether this is true depends on what counts as an intervention. One way to manipulate salty food consumption, you might think, is to ensure that a person drinks large quantities of bottled water (since this dilutes the body's electrolytes, producing an appetite – perhaps – for more). Such a manipulation is a legitimate intervention on salty food consumption (for the purposes of examining the connection between salty food and heart disease) only if there is no causal connection between water consumption and heart disease. But it was precisely this connection that was in question.

Expand the definition of unrelativized causation between water and heart disease, then, and you find yourself with a definiens that calls on the facts about unrelativized causation between salty food and heart disease.[6] But expand the definition of unrelativized causation between salty food and heart disease and you find yourself with a definiens that calls on the facts about unrelativized causation between water and heart disease.

Woodward allows that there is a certain circularity in his definitions; he denies, however, that it is vicious, where a vicious circularity would result if for example "the characterization of an intervention on *X* with respect to *Y* itself makes reference to the presence or absence of a causal relationship between *X* and *Y*" (*Making Things Happen*, 104). But this is precisely the sort of thing we have here.[7]

6. This is not quite correct: the definiens takes the form of an existential claim, thus, what is required for the definiens is facts about some intervention or other, not facts about whether water consumption in particular is an intervention. But the circularity I am describing here will, I think, arise for any putative intervener, which is to say that the definition of a variable *I* as an intervention for the purposes of determining *X*'s causal influence on *Y* will, when the definiens is expanded, invoke the facts about *X*'s causal influence on *Y*.

7. Incidentally, in both *Making Things Happen* and his reply, Woodward appears to hold that a system of definitions cannot be viciously circular if it imposes constraints on the defined properties. This is not correct. Consider the following two definitions: a day is a Tuesday just in case it comes after a Monday; a day is a Monday just in case it comes before a Tuesday. Clearly the definitions are viciously circular in Woodward's sense, yet equally clearly, they are inconsistent with certain possible facts about days, for example, yesterday's being Monday and tomorrow's being Tuesday (assuming that no day can be both a Monday and a Tuesday). Like any necessary condition, this one can warrant an inference from one fact about days to another, distinct fact about days: if I already know that yesterday was Monday, I can infer that today is

In the next section, I will nevertheless suggest that there is a way of understanding Woodward's project that makes room for "vicious" circularity.

3. Ultimate Ends

Woodward denies, in his reply to my review, that his account of causation is a metaphysics of causation – he denies that it is an attempt to articulate the nature of causal facts, or equivalently, to articulate the nature of the truthmakers for causal claims. What, then, does the Woodwardian theory attempt to do? I cannot find its aims plainly spelled out in the reply. In this final section, let me therefore advance several suggestions, that is, several possible ultimate goals for an account of causation that are, as best as I can make them, true to the spirit of Woodward's metaphysical demurral.

What I cannot give you is a non-metaphysical goal for a philosophical account of causation that is a plausible interpretation of what Woodward is up to in *Making Things Happen* – because despite his denials, what he is up to is pretty clearly metaphysics. I suppose I had better defend this claim; let me do so in three ways.

First, in *Making Things Happen* Woodward sets up a dialectic that treats his account of causation as belonging to the same genre as other, clearly metaphysical, theories of causation such as David Lewis's counterfactual accounts and Menzies and Price's manipulation account (Lewis 1973, 1986b, 2000; Menzies and Price 1993). Most strikingly, he presents his account as a direct rival to these other accounts (in, for example, sections 2.1, 3.4, 3.6, 3.8), which hardly makes sense if his philosophical aims are fundamentally different from the aims of Lewis and company. In a comparison of the "deep differences" between his theory and Lewis's (p. 136), for instance, Woodward notes, first, that *Making*

Tuesday. The problem with vicious circularity is not that it precludes a system of definitions' constituting a necessary condition, a constraint, on the defined terms, but that it (typically) precludes the system's offering a sufficient condition, by rendering the system ungrounded: if the mooted definitions of Monday and Tuesday were for real – if they exhausted the facts about what days were Mondays and Tuesdays – there could be no matter of fact as to whether today were Monday or not.

Things Happen is offering an account of causation between event *types* (or as he says, variables), whereas Lewis is offering an account of causation between event *tokens*, and second, that *Making Things Happen* is offering a non-reductive account of causation (the notion of a non-reductive metaphysics is of course a familiar one). He does not utter a word about what would presumably be the deepest difference of all: Lewis's goal is entirely different from Woodward's putative unmetaphysical goal. No reader could fail to infer from many passages in *Making Things Happen* that Woodward intends his manipulation account as a replacement for Lewis's and other accounts of causation – which implies that it is intended to do metaphysical work.[8]

Second, Woodward uses his characterization of causation in ways that make sense only if the characterization is understood as articulating the nature of causation. In section 3.3 of *Making Things Happen*, for example, he argues that according to the manipulation account, causation is mind-independent, or objective, on the grounds that the characterization of causation offered by the account makes reference only to counterfactual facts that are "not dependent on human attitudes or beliefs" (118). This argument clearly supposes that the Woodwardian characterization of causation tells us about the nature of causation (indeed, it had better tell us *everything* about the nature of causation, or else we might miss some mind-dependent truthmaker for causal claims).

Third, and most important, is Woodward's explicit statement of his goal in *Making Things Happen*'s chapters on causation: "my aim is to give an account of the content or meaning of various locutions, such as *X* causes *Y*" (p. 38). This, then, is apparently the most familiar of philosophical projects: to capture the meaning of causal language, or the content of causal concepts, with definitions.

In modern times, such a project is invariably interpreted as aiming to provide

8. Let me give you one example among many in the discussion of Lewis. On p. 137, Woodward writes that "To the extent that [Lewis's criteria for similarity between possible worlds] are clear, they often lead to the same results as the manipulationist account, but not always; sometimes, Lewis's rules lead us to insert miracles in the 'wrong' place and generate mistaken evaluations of counterfactuals and hence of causal claims, assuming the connection between counterfactual dependence and causation that Lewis advocates. In such cases, the intervention-based approach is superior." This passage obviously presumes that it is an aim of the manipulation account to give the correct story about the evaluation of causal claims, that is, to give the truth conditions for causal claims.

truth conditions for the sentences or thoughts in question, and therefore as aiming to specify those representations' truthmakers. It may look like semantics, but it is also a kind of metaphysics. Further, this not merely a tradition: it is generally agreed that a word with an explicit definition has as its extension whatever stuff satisfies that definition. If Woodward's causal semantics is a truth-conditional semantics, he is inevitably, unavoidably, ineluctably committed to producing an account of the truthmakers for causal talk, a metaphysics of causal facts, whatever his protestations.

What else might he be up to? His reply cites a passage on pp. 7–8 of *Making Things Happen* in which he tells us he is not merely doing conceptual analysis. A closer reading of these pages shows, however, that he is very much in the neighborhood – he is attempting what Carnap called an explication of causal language, simultaneously an analysis and a tidying up of everyday ways of talking.[9] This hardly absolves his definitions of the role of specifying truthmakers for causal talk.

Could it be, then, that Woodward's semantics is not truth conditional? This interpretation is hardly supported by passages such as that quoted in note 8; nevertheless, I will explore this possibility shortly.

To sum up, there are many compelling reasons to read *Making Things Happen* as offering a metaphysical account of causation; as a consequence no unmetaphysical interpretation of the book will enjoy strong textual support. As Woodward says, however, *Making Things Happen* does contain several passages of anti-metaphysical editorializing; at one point, for example, Woodward writes that his manipulation account is committed "to no particular metaphysical picture of the 'truth makers' for causal claims" (122). I simply cannot square this with what is actually going on philosophically in *Making Things Happen*.[10] The sentence I have quoted, in particular, occurs immediately after the argument for

9. For example, although he says his account "makes recommendations about what one *ought to* mean by various causal . . . claims" (my italics), he then goes on to explain that these recommendations are merely attempts to repair elements of causal talk that are "confused, unclear, and ambiguous" (p. 7).

10. Unless, of course, Woodward means simply that, having analyzed causal claims in terms of a certain kind of counterfactual dependence, he is leaving the further analysis of counterfactual dependence to other writers.

mind-dependence sketched above, an argument which makes sense only on the assumption that *Making Things Happen*'s definition of causation presents an accurate picture of causal claims' truthmakers.

In composing my review, I understood these passages as symptoms of Woodward's vestigial discomfort with the overt metaphysical tenor of his project – the sort of thing that would be quietly removed from the revised edition of *Making Things Happen* – and thought it better to pass over them in silence. It now appears that it is the metaphysics that will be expunged from the revised edition. Let me therefore put aside the whole project of finding a univocal reading of *Making Things Happen*, and explore for a couple of pages the possibility of an unmetaphysical manipulationist account of causation.

* * *

If not metaphysics, then what? At times, Woodward talks as though his "manipulation account of causation" is nothing more than an exploration of certain connections between causality and manipulability. He wants to tell us that wherever you have causality, you have the possibility, in principle, of a certain kind of manipulation, and perhaps vice-versa.

This would certainly constitute a philosophically interesting claim. But it can hardly be all that Woodward intends. He quite self-consciously presents the connections between causation and manipulation in the form of *definitions*, not matters of fact. This has the consequence – perverse on a "matter of fact" interpretation – that they are treated as stipulations rather than as hypotheses, and so are not directly defended.

I do not mean to denigrate the "matter of fact" project: examining the links between causing and manipulating is a worthwhile and (in a certain sense) unmetaphysical undertaking. But Woodward is clearly, in his use of definitions and his repeated claims to be analyzing causal language and thought, after something richer and deeper than this.

In his reply, Woodward tells us that his primary focus in *Making Things Happen* is "*methodological*: how we think about, learn about, and reason with various causal notions" (p. 2); later he compares the role of definitions in his project to their role in work by Pearl and Spirtes et al. (pp. 3–4). Let me see if I

can do these claims justice.

Pearl and Spirtes et al. are attempting to systematize, and indeed to a great extent automate, various kinds of causal inference. They treat causation between variables as a primitive, and introduce a number of postulates about causation that provide the grounding for their algorithms. Their systems can be understood, then, as operating along Euclidean lines: certain axioms are posited regarding the raw material, causality, which remains uninterpreted; further notions such as intervention are defined in terms of the primitives; theorems are then proved, the theorems in this case being chosen for their utility in causal inference.

Woodward's project might perhaps be understood as a contribution to the foundations of this project, along the following lines:

1. Woodward's "definitions" of intervention and various notions of causation are axioms.

2. The aim of the axioms is to capture certain foundational (but not necessarily metaphysical) facts about causation and manipulability that have consequences that (as Pearl and Spirtes et al. have shown) are especially fruitful for the purposes of causal inference.

3. Therefore, Woodward's "definitions" are not definitions in the mathematical sense. Further, they do not provide any kind of causal semantics (conceptual analysis, explication, rational reconstruction, etc.) And they do not constitute a metaphysics of causation.

Such a project is a coherent one, and fits rather well with the anti-metaphysical tenor of Woodward's reply. It is, however, at odds with the various aspects of *Making Things Happen* noted above.

First, providing a formal characterization of a system of causal inference is a project that is quite orthogonal to that of providing a metaphysics of causation, so it is a mystery how, on this interpretation, Woodward could present his manipulation account as a rival to Lewis's counterfactual metaphysics or Menzies and Price's manipulation metaphysics. I suppose it is possible that something in the system of causal inference might be inconsistent with something in

the metaphysics, but as far as I can see, the methods developed by Pearl and Spirtes et al. are compatible with any of the known metaphysical accounts of causation.

Second, for similar reasons, an axiomatization of causal inference is obviously not going to resolve the question of the mind-dependence of causality.

Third, providing inferentially fruitful axioms about causality and intervention is hardly the same thing as providing what Woodward claims to supply in *Making Things Happen*, a semantics for causal language and thought.[11]

You might wonder, though, whether a semantic claim could not be appended to the axioms. Could they not be interpreted, in the spirit of the conventionalist approach to the philosophy of mathematics, and later empiricist approaches to the philosophy of everything, as *implicit definitions*? The result would be a kind of inferential role account of the meaning of causal claims, on which their cognitive significance is exhausted by their place in a system of causal reasoning.

Such an addendum would make sense of Woodward's claims to be doing the semantics of causation. But I think he would be wise to reject it, as it has apparent metaphysical implications, giving you a deflationary, unmetaphysical metaphysics of causation on which there are no real causal connections in the world, merely causal ways of thinking about real world connections. (This deflation was, of course, an explicit aim of the conventionalists and logical empiricists.)

An alternative, originally raised in my review (p. 246), is that causal language has a "two-factor" semantics (Putnam 1975), the two factors corresponding to, roughly, inferential role and truth conditions (or reference). Woodward might claim to be doing the semantics of only the first factor, which would retain a place for causal reality while keeping its nature out of the account.

If Woodwardian manipulationism is to be given this semantic gloss, a further question must be raised: whose concepts are the subjects of the analysis? Is

11. Consider for example p. 38 of *Making Things Happen*: "I have nothing to say about issues having to do with . . . causal inference. Instead my enterprise is, roughly, to provide an account of the meaning or content of just those qualitative causal notions that Pearl (and perhaps Spirtes et al.) take as primitive . . . my project is semantic or interpretive."

this an account of the semantics (or rather, of one component of the semantics) of everyday causal talk, or is it rather an analysis of a highly refined technical language used only in the more exalted temples of causal inquiry?

Some parts of *Making Things Happen* strongly imply that Woodward's topic is the everyday concept. In particular, Woodward refers in several places to the evolutionary history of the concept in question, and in particular the survival value of manipulative know-how. Though the selective pressures in the Carnegie Mellon Philosophy Department are surely ferocious, Woodward must be referring here to events taking place long before the development of science, let alone Bayesian networks.

Yet if Woodward's semantics is an inferential role semantics, it is not clear that Woodward *can* be telling us about the everyday concept. The system of causal reasoning developed by Spirtes et al. is intended to be an augmentation of previously existing techniques. It may resemble everyday causal reasoning in some respects, but it goes well beyond it in others, such as the importance it gives to the calculation of "tetrad statistics".

Perhaps here we can make sense of Woodward's claim – rather puzzling for either a semanticist or a metaphysician – that his account of causal concepts is concerned to "establish[] fruitful connections with other concepts" (p. 3). I have in mind the following train of thought: Our everyday concepts have a certain inferential role as a result of their place in our pretheoretical practices of causal reasoning; further, this role in some sense contributes to their semantics. But the pretheoretical practice of causal reasoning can be improved. This will result in a change in the inferential role of, and thus a change in (one component of) the semantics of, our causal concepts. Woodward's manipulation account, the thought continues, presents this conceptual ideal: it is an account of what inferential role our causal concepts ought to have, and thus what semantics (truth conditions excepted) they ought to have. Those semantics are, concludes the thought, pithily captured by the axioms of the relevant reasoning system.

Suppose that *Making Things Happen* was all along intended to advance the project just described, and not to engage in causal metaphysics. Which of my criticisms of the book, both here and in the original review, stand, and which no longer seem relevant? (Obviously, what I have to say about Woodward's

accounts of singular causation and causal explanation stands; it is what I have to say about what I took to be his causal metaphysics that requires reexamination.)

First, my complaint that Woodward's definitions presume a system of representation, involving directed graphs and so on, that fails to capture naturally (or perhaps at all) certain kinds of causal structure: sustained. If the account is supposed to systematize, indeed materially improve upon, our everyday causal reasoning, it had better have at least the scope of our everyday reasoning (except perhaps where that reasoning is hopelessly confused).

Second, my complaint that Woodward's system does not capture every aspect of our causal reasoning, in particular, our reasoning about the connections between different "levels of causation" (e.g., causation at the biological and the physical level): sustained.

Third, my complaint in section 2 above that Woodward's account does not provide a proper grounding for the fact that water consumption does not cause heart disease, or even worse, that it provides grounding for the "fact" that water consumption *does* cause heart disease: dismissed. On the present interpretation, Woodward's account does not attempt to provide the grounding for such facts. My complaint in the review that such facts look to change depending on the variable set to which they are relativized is dismissed for the same reason.

Fourth, my complaint that Woodward's definitions are viciously circular: sustained and dismissed. Sustained, that is, because they meet the criteria for vicious circularity, but subsequently dismissed (in spite of Woodward's own worries about viciousness in *Making Things Happen*) because this circularity is not a vice in an axiom, insofar as "circular" axioms are quite capable of providing a formal basis for a system of inference (see note 7) – indeed, it is virtually a category mistake to call an axiom "circular". (Nevertheless, I hope that my investigation of the circular structure of Woodward's system of definitions will be useful to philosophers who, in the future, pursue manipulationism as metaphysics.)

This does not mean that circularity has no negative implications for Woodward's project. Very (very!) loosely, the more circular a set of axioms, the weaker their implications, and thus the weaker the inferential machinery that they

underwrite. It is well worth taking a closer look at the power of the inferences licensed by the Woodwardian axioms.

For these purposes, let me divide causal inference into two classes: reasoning from cause to effect, as when we attempt to predict the consequences of some event or action, and reasoning from effect to cause, as when we attempt to infer underlying causal structure from observed phenomena.

Woodward's system provides a foundation for a certain kind of reasoning from cause to effect, I think (and this is after all what matters for a book on causal explanation). It is harder, however, to see how it provides more than token assistance in reasoning from effects to causes. The case of water and heart disease shows why. Given a sufficiently rich set of causal information – for example, a division of causes of water consumption into those that are genuine interventions with respect to heart disease and those which are not – Woodward's axioms allow us to infer new causal relationships. But if we have to start from scratch, knowing nothing or very little about the causal relations between water, salt consumption, and heart disease, the axioms are far less helpful; indeed, it is hard to see how to get started at all, that is, how to get from even an encyclopedic knowledge of correlations to any knowledge at all about causal relations.[12]

If this inferential helplessness in the face of a causal blank slate were simply a reflection of our actual inability to extract causal information from statistical information, then there could be no criticism of Woodward's system: intended to capture our patterns of causal reasoning, it naturally reflects their weaknesses as well as their strengths. But in fact we can do much better than this. Spirtes et al., in particular, have shown how to infer causation from information concerning only correlation.

At the core of Spirtes et al.'s technique lie two principles of special interest to philosophers of causation: the faithfulness condition and the causal Markov

12. Woodward appears to offer a way out of this bind when he writes in his reply that "If I carry out an appropriately designed randomized experiment, I can know I've performed an intervention on *X* with respect to *Y*" (p. 14). But as far as I can see, this is true only on a definition of "appropriately designed" on which ascertaining that I have carried out an appropriately designed experiment raises the same questions as ascertaining that my chosen intervention satisfies the criteria for interventionhood – questions that the axioms cannot answer.

condition. It is these two posits more than any others that allow Spirtes et al. to reason in a principled way from pure correlation to causation. Faithfulness says in effect that positive and negative effects will not, or at least will almost never, exactly cancel out, that is, that you will almost never get cases like my (imagined) relation between salty food and heart disease above. I will not try to state the Markov condition here; suffice to say that it follows from Woodward's axioms together with the principle that you do not (or almost never?) get correlation without appropriately directed causation. Now, Woodward's axioms entail neither (a) that positive and negative effects seldom cancel out, or (b) that there is no correlation without causation. Thus they do not entail either the faithfulness or the causal Markov condition. For this reason, they make for a system for inferring causes from effects that is far less powerful than the system we in fact use (or at least, than we are capable of using).[13]

It therefore seems that Woodward should add the faithfulness and causal Markov conditions to his system. But here is the rub: despite their great heuristic power, they look very much like contingent, not conceptual, truths. Just as clearly as they belong to any characterization of idealized human causal inference, then, they clearly fail to belong to any semantics of human causal thinking.[14] The methodological and the semantic projects come apart.

Let me conclude by considering one other project that Woodward might be pursuing. I have made a distinction between those principles of Pearl's and Spirtes et al.'s inferential systems that might conceivably be regarded as

13. In the main text I consider only the problem of learning about causation in a case in which a putative intervention has been made. I should add that, although a great deal of causal knowledge is gained by way of experiments involving interventions, at least as much is gained through careful statistical reasoning about cases in which nothing remotely approaching an intervention occurs. Take, for example, the analysis of the possible causal link between high exposure to lead and childhood cognitive deficits. Obviously, no clinical trial was conducted in which a group of children were, by way of an intervention, systematically dosed with lead. Rather, the analysis turns on existing cases of lead exposure, cases in which the causes of exposure fall short of the criteria for interventionhood in almost every respect imaginable. There are various ways of tackling such an analysis, beginning with the strategy of – when enough of the right kind of data is available – controlling for confounding causes by finding subgroups in which all likely causes of cognitive deficits other than lead are present to an equal degree. Such strategies are based on assumptions in the same vein as faithfulness and the causal Markov condition.

14. Our use of powerful inferential heuristics creates tremendous problems, I might add, for any kind of inferential role semantics.

conceptual truths, and those that clearly are not conceptual truths. Perhaps Woodward's definitions are intended to capture only the former? If so, then contrary to what I supposed immediately above, the definitions are not themselves axioms of a complete system of inference. They rather supply the conceptual grounding, in some sense, for the axioms.

What kind of grounding might this be? In particular, what grounding needs to be given to the notion of causality that appears in Pearl's and Spirtes et al.'s foundational principles? Clearly, Pearl and Spirtes et al. intend their inferential structure to apply to the same stuff as our everyday claims about causality. You might think, then, that a condition that is both necessary and sufficient for their systems to be well grounded is that the causal terms that appear in the systems have the same reference (more or less) as everyday causal notions. It would follow that the conceptual foundations of the Bayesian networks literature are identical to the conceptual foundations of everyday causal talk, and are made manifest in a truth-conditional semantics for this causal talk, or in other words, by a metaphysics of causal facts.

Now Woodward in his reply of course denies that he is engaged in providing such a thing; his conceptual foundations would then have to be of a different sort. If conceptual foundations are not an outwardly directed, truth-conditional matter, perhaps they are something internal – a matter of coherence or inferential significance. You might, for example, ask that conceptual foundations for causality entail the major principles of causal inference. But as I have argued, Woodward has not succeeded in providing such foundations; quite possibly, he has never intended to.

What is the point, then, of conceptual foundations? For internal purposes, Pearl and Spirtes et al. simply take the causal notions around which their inferential systems revolve as primitive. Is any further grounding needed (apart from an external, truth-conditional link to causal facts)? I do not see that it is. But perhaps Woodward disagrees: perhaps he sees the work of Pearl and Spirtes et al. as somehow in need of additional non-truth-conditional conceptual scaffolding. If so, we need to know what kind of support is required and how Woodward's definitions of causation do the job.

I think that is enough for now; I hope that the discussion in this "counter-

reply" of mine will provoke Woodward to give us a clearer view of the aims of his unmetaphysical "account of causation", and of the conditions under which those aims might be thought to have been achieved.

Acknowledgments

Thanks to Jim Woodward and Brad Weslake for helpful comments.

References

Lewis, D. (1973). Causation. *Journal of Philosophy* 70:556–67. Reprinted in Lewis (1986a).

———. (1986a). *Philosophical Papers*, volume 2. Oxford University Press, Oxford.

———. (1986b). Postscript to "Causation". In Lewis (1986a), pp. 172–213.

———. (2000). Causation as influence. *Journal of Philosophy* 97:182–97.

Menzies, P. and H. Price. (1993). Causation as a secondary quality. *British Journal for the Philosophy of Science* 44:187–203.

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

Putnam, H. (1975). The meaning of 'meaning'. In K. Gunderson (ed.), *Language, Mind and Knowledge*, volume 7 of *Minnesota Studies in the Philosophy of Science*. University of Minnesota Press, Minneapolis.

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

Strevens, M. (2007). Essay review of Woodward, *Making Things Happen*. *Philosophy and Phenomenological Research* 74:233–249.

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