*Philosophy and Phenomenological Research*
Vol. LXXVII No. 1, July 2008


# Response to Strevens*

JIM WOODWARD
*California Institute of Technology*

I am very grateful to the editors of *Philosophy and Phenomenological Research* for giving me this opportunity to respond to Michael Strevens' review of *Making Things Happen* (*MTH*).

I believe that many of the positions that Strevens attributes to me and which serve as a basis for his criticisms rest on misinterpretations of *MTH*. One reason for this is that Strevens reads me with the preoccupations of a metaphysician; another is that he relies heavily, in reporting what he takes to be my views, on restatements of those views within his own terminology and system of concepts rather than on what I actually say. I will cite examples below, but I am also mindful that (for very good reasons) readers of this journal are not going to be interested in a long and tedious list of I-never-said- that's. For this reason and because I lack the space to respond to everything in Strevens' review, I would encourage readers of this exchange to read *MTH* and to form their own assessment of its contents.

I turn now to some more specific comments on Strevens' discussion beginning with his claims about the metaphysical commitments of *MTM*.

## 1. Metaphysics

Strevens presents *MTH* as, among other things, an attempt to provide a metaphysics of causation. He writes, for example, that I hold that "facts about causation *metaphysically depend* on what can be manipulated by what," that, according to *MTH*, direct causation is the (presumably metaphysically) "fundamental causal notion" and so on. He then goes on to criticize *MTH* on the grounds that it is deficient qua metaphysics and lacks adequate metaphysical foundations. However,

* Thanks to Chris Hitchcock, Dan Hausman, Ken Waters, and Jiji Zhang for helpful comments.

the metaphysically loaded language (including phrases like''metaphysically depends'' ) that Strevens uses to describe my views does not occur in *MTH*. These are Strevens' *restatements* of my views, presented as though they straightforward reports of what I say. In fact, I go to some lengths to argue that the one of the attractions of the manipulationist account is precisely its unmetaphysical character—rather than thinking of causal relationships as involving mysterious other worldly entities (relations of necessitation among universals, similarity relations among possible worlds and so on), I urged instead that we think of them simply as relationships that are exploitable for purposes of manipulation and control. I argue that this makes it intelligible why we should care about discovering causal (as opposed to merely correlational relationships) and also helps to illuminate many of the ways in which we learn about and reason about causal relationships. For those who care about metaphysics, this sort of view might be supplemented by any one of a number of different stories about metaphysical foundations but *MTH* does not attempt to provide such foundations.

If *MTH* is not a metaphysical treatise, what is it about? *MTH* was written as a contribution to philosophy of science. It ranges over a number of different topics but the primary focus is *methodological:* how we think about, learn about, and reason with various causal notions and about their role in causal explanation, both as these occur in common sense and in various areas of science. It claims that these issues can be illuminated by focusing on the connection between causation and manipulation (or intervention) and accordingly offers interventionist accounts of various causal notions (including the notions of direct, contributing, total, and actual causation). These accounts are compared with alternative treatments of causation and causal explanation in the philosophical and statistical literature and it is argued that my approach avoids various difficulties and counterexamples that infect these alternatives. As Strevens reports, *MTH* employs representational devices such as directed graphs and systems of equations (rather than devices like first order logic, sets of necessary and sufficient conditions or probability theory that have more traditionally been used by philosophers in discussions of causation) but these are employed in the service of the ends just described rather than as part of any grand metaphysical agenda. Among other things, *MTH* asks questions like the following: given that a directed graph or a system of equations can be used, qua representational device, to represent both patterns of correlations and systems of causal relationships, what conditions have to be met for these devices to accurately represent the latter rather than the former? When we use a directed graph to represent causal relationships, what interpretation should be given to e.g., the arrows (directed

edges) in the graph? If we wanted to try to capture the notion of one events being an actual cause of another within a structural equations or directed graph framework, how would we do so? The overall perspective of *MTH* is what might be described as that of a modeler: pragmatic, piece-meal, and anti-foundational.

It is true that *MTH* presents "definitions" of the various causal concepts mentioned above. Some of these concepts are defined in terms of others and all are taken to be related in various ways to the notion of an intervention, which I also define in *MTH*. It may be that it is the presence of these definitions that leads Strevens to think that I am trying to do metaphysics.[1] However, I also explain how these definitions are to be taken in the opening pages of *MTH* (pp. 7-9): they are definitions in the sense that, say, a mathematician might define the notion of continuity of a function in terms of the notion of an open neighborhood. Such definitions are to be judged by their usefulness for various purposes– in capturing previous usage, in clarifying notions that were previously unclear and distinguishing them from related but different notions, in establishing fruitful connections with other concepts and so on, rather than in terms of whether they adequately capture fundamental metaphysical relationships. In particular, the definitions offered in *MTH* were not intended as claims that concept being defined "metaphysically depends" on the concepts offered in the right hand side of the definition or that the latter are metaphysically more fundamental than the former. So while, for example, I define the notion of *X*'s being a "direct cause" of *Y* in terms of facts about how *Y* would change under an intervention on *X* when other variables are held fixed by independent interventions, this is not intended to be a claim to the effect that direct causal relationships metaphysically depend on facts about what would happen under such combinations interventions. And while I then go on to define the notion of a "contributing cause" in terms of the notion of direct causation (as well as the satisfaction of other conditions), it is no part of my view that the notion of direct causation is the "metaphysically fundamental" notion in causation, any more than the mathematician's definition of continuity commits her to the idea that the concept of an open neighborhood is metaphysically more fundamental than the concept of continuity. This conception of how definitions are to be taken is not eccentric, given the traditions of the literature to which I am attempting to contribute. For example, one finds a similar, non-metaphysical conception of what is involved in giving a definition in books like Judea Pearl's *Causality* (which contains definitions of notions like "causal effect," "direct cause" etc.) and in

[1] Strevens has suggested as much in email correpsondence.

Spirtes, Glymour and Scheines' *Causation, Prediction and Search*. The mere fact that these books contain definitions does not mean that they are projects in metaphysics.

There are many reasons why a metaphysical reading of the definitions that I give seems obviously inappropriate. For one thing, as Strevens notes, all of the causal notions discussed in *MTH* are defined by reference to the notion of an intervention, a notion which I explicitly acknowledge is a causal notion. It would make no sense at all to claim that, e.g., direct causation is "the" fundamental causal notion and then define this in terms of concepts that are already acknowledged to be causal. Instead, as I explicitly say in the opening pages of *MTH*, the treatments of the various causal notions I provide (and the definitions associated with these) are not intended to reduce these notions to something more basic or fundamental but rather to exhibit connections and interrelations with other causal notions, to connect with issues about information about causal relationships are learned, and to contrast with other treatments of causation in the philosophical and scientific literature.

None of this would matter much if Strevens did not go on, after offering a strongly metaphysical construal of *MTH*, to use this construal to launch various criticisms that would not be well motivated if *MTH* was not making metaphysical claims. I agree with Strevens , for example, that the notion of direct causation is a particularly poor candidate for a metaphysically fundamental notion (see below for more on why), but I never claimed that it could be used in this way and, in my view, it does not follow from its inappropriateness as a metaphysical primitive that this notion is not a useful one for the non- metaphysical purposes to which I put it in *MTH*. A similar point holds for Strevens' contention that *MTH* fails to establish "metaphysical manipulationism" and instead at best provides arguments for a deflationary position according to which facts about manipulation merely entail facts about causation and vice versa. *MTH* does not attempt to argue for metaphysical manipulationism and (assuming for the sake or argument that the deflationary position is the only alternative) I'd be quite satisfied if *MTH* "only" established this alternative. I would add, though, that giving a statement of the entailment relations between causation and manipulation is not a trivial or insignificant matter and, in my view, can be genuinely illuminating for many purposes, even if this contributes nothing to metaphysics.

Lurking in the background of this back and forth about the alleged metaphysical commitments of *MTH* are a number of more substantive and interesting issues. *MTH* assumes that there are many worthwhile things that can be said about causation and causal explanation without doing metaphysics, and that concepts like direct causation and devices

like directed graphs can be useful even if they have no deep metaphysical significance. I suspect that many of Strevens' criticisms derive at bottom from disagreements with me about these deeper issues: that is, he interprets *MTH* as a book about metaphysics because he thinks that metaphysics should play a central role in any book about causation. This view may be correct, but it needs to be argued for, rather than simply assumed.

## 2. Actual Causation

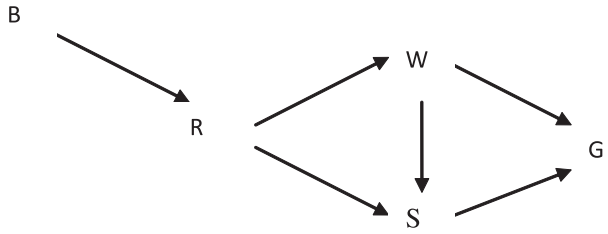
As Strevens says, *MTH* contains an account of "actual causation," as when we claim that one particular event was "a" or "the" actual cause of another. This account is presented within a structural equations or directed graphs framework and is not original with me: it draws heavily on prior work by Pearl and Halpern and by my colleague Chris Hitchcock. Here is the core of the account, reproduced for latter reference:

> AC: $X = x$ is an actual cause of $Y = y$ iff
>
> (AC1) The actual value of $X = x$ and the actual value of $Y = y$.
>
> (AC2) There is at least one route (directed path) $R$ from $X$ to $Y$ for which an intervention on $X$ will change the value of $Y$, given that the other direct causes $Z_i$ of $Y$ that are not on this route have been fixed at their actual values. (It is assumed that all direct causes of $Y$ that are not on any route from $X$ to $Y$ remain at their actual values under the intervention on $X$.)

My discussion of actual causation occupies a peripheral role in *MTH* (12 pages in a 400 page book, hedged with various qualifications acknowledging that what I say cannot be the whole story about actual causation). Most of *MTH* is about notions of causation and causal explanation that are more type-like or population level or that have to do with the explanation of generalizations or repeatable phenomena, rather than particular events. I focus on these latter notions because, like most philosophers of science, I think that they are the more important notions in science. Contrary to what Strevens' review might seem to suggest, the adequacy of my account of actual causation matters very little for most of what I say in the rest of the book.

That said, I don't understand Strevens' purported counterexample to this account. The central problem is that it is unclear what the causal graph associated with his example is. Moreover, Strevens does not rely on the condition **AC** but rather on his own informal gloss which talks about how "the effects of switching… propogate through the system"

and so on. As best I can tell the causal graph associated with Strevens' purported counterexample is this:[2]




Here $B$ is a variable measuring whether or not the broadcast occurs, $R$ measures whether the resolute rebel shows herself, $W$ measures whether the wavering rebel pushes the button, $S$ measures whether the resolute rebel fires the rocket and $G$ whether the general dies. I emphasize that this graph follows immediately from my characterization of direct causation **DC,** assuming that I have correctly interpreted Strevens' description of the example—one of the attractions of **DC** is that it tells you how to construct such graphs**.** By contrast, although Strevens draws some graph—like figures in his review (not directed graphs as ordinarily understood), we are given no account of the rules governing their construction.

The actual values of the variables $B$ and $G$ are that the broadcast and the general's death actually occur, in conformity with AC1. Moreover, in conformity with AC2, there is a directed path or route from $B$ to $G$ (the path $B \rightarrow R \rightarrow W \rightarrow G$) such that when we fix variables that are off this path (in this case, there is one such variable $S$) at their actual values (in this case, the launching of the rocket by the resolute rebel does not occur), an intervention that changes whether the broadcast occurs changes whether the general dies.

Thus the broadcast *does* count as an actual cause of the death of the general according to **AC**, contrary to what Strevens claims (and in accord with intuition). If Strevens has in mind some other causal graph I encourage him to write it down and show how the application of **AC** yields the result he claims.

I will add that although I don't think that **AC** runs afoul of Strevens' example, Hitchcock, forthcoming and Hall, forthcoming do contain genuine counterexamples. My view is that as long as **AC** correctly captures or models one set of considerations that influence actual cause judgment, progress has been made, even if there are other considerations that are left out of the account. (This is part of what I meant earlier when I said that *MTH* adopts a modeler's perspective.)

[2] It was unclear from Strevens' description whether there should be an arrow directly from $B$ to $W$, but adding this arrow makes no difference to what I say below: the broadcast still comes out as the actual cause of the general's death.

Of course, an alternative possibility is that **AC** is on the wrong track entirely and hence should be abandoned.

### 3. Actual Causation, Potential Causes and Robustness

Strevens' review raises interesting questions about the role of information about actual (as opposed to merely potential causes) in explanation. Although most of the discussion of explanation in *MTH* is not about the explanation of individual events, I adopt the view that the explanation of individual events only involves citing information about the factors that are actual causes (as characterized by **AC**) of those events. (''Explanation'' here and in what follows means ''causal explanation,'' as it does throughout *MTH*) Thus in a scenario in which assassin one kills the victim but a backup assassin would have killed the victim if assassin one had not acted, it is the action of assassin one that explains the victim's death. This contrasts with the alternative view that information about merely potential but non-actual backup causes as well as its actual causes is relevant to the explanation of an event. On this view, the fact that the backup assassin two would have shot if assassin one had not, is explanatorily relevant to (part of the explanation of) the victim's death, as of course is the actual shooting by assassin one. This second view—call it *causal potentialism*—may be Strevens' view and he also attributes it to me. (Indeed, he even claims that ''it is this doctrine more than any other that distinguishes explanatory manipulationism from rival causal approaches to explanation,'' which is certainly news to me.) At the same time, he criticizes me for not providing arguments for this position and for not addressing issues about robustness that he thinks are raised by it.

The reason why I give no argument for causal potentialism is that *MTH* does not adopt this view. (Although perhaps it should have—see below.) Strevens does not make it clear why he supposes that *MTH* is committed to causal potentialism, but judging from his remarks on explanatory depth, he may be influenced by what I think is a mistaken analysis (which he takes over from Jackson and Pettit, 1992) of an example that I discuss in *MTH* concerning the explanation of the behavior of a gas.[3] Compressing greatly, I argued that if we wanted to

[3] Another possibility that is that Strevens assumes that my emphasis on the role of what I call what-if- things-had-been-different questions in explanation entails causal potentialism. I don't think there is any such entailment. Claiming, as I do, that e.g., the explanation of the field created by a the charge a long a wire works by conveying information about what the field would have been if the shape of the wire had been different does not imply anything in particular about the explanatory status of back-up causes. More generally, claiming as I do that, in a situation in which there are no back-up causes, information about what would happen under changes in the value of a variable cited in an explanation is relevant to its explanatory import does not imply that citing non-actual backup causes (in situations in which they are present) is similarly explanatorily relevant.

explain why the gas is at one pressure rather than some alternative pressure, it was better to appeal to other macroscopic variables like its volume and temperature, rather than following what I called the microscopic strategy of attempting to trace the trajectories of the individual molecules making up the gas. I intended this as a claim about the choice of level of explanation but Strevens seems (mistakenly, in my view) to assimilate this example to cases like that of the backup assassin. That is, he apparently thinks one should analyze the example in terms of the idea that if that actual set of trajectories had not occurred, another set of trajectories, consistent with the initial thermodynamic constraints, would have instead and would have led to the same outcome, just as the assassination example is understood in terms of the claim that if the actual assassin does not act, the backup will. Thinking of the gas example in this way, it might seem that recognizing the superiority of the macroscopic strategy requires the endorsement of causal potentialism.

Although the matter deserves more attention than I can give it here, I think this is a mistaken analysis (and it is not the analysis presented in *MTH*). There are no back up causes in the gas example, ready to operate to produce the same outcome if the actual causes of the new pressure are not operative. The reason why, on my account, it is unsatisfying to cite the actual molecular trajectory to explain the final pressure of the gas doesn't have to do with considerations involving non- actual back up causes but rather has to do with the of the fact that citing the actual trajectory leads to an explanation that is at the wrong grain or level in the sense that it doesn't accurately represent the dependency relationships present in the example. When we cite the volume and temperature of the gas to explain its new pressure, we are citing straightforward causes of the new pressure, not assigning an explanatory role to potential causes.

Because *MTH* does not endorse causal potentialism, there is also no need from my point of to provide the account of robustness that Strevens thinks should accompany this position.

There is however a separate and more interesting question: *should* an interventionist like me adopt causal potentialsm? Strevens observes (and I agree) that information about potential back up causes is certainly relevant to manipulation and control—so why not take such information to be explanatorily relevant within the interventionist account? In support of this position one might note that historians frequently make claims about back up potential causes to advance historical understanding: e.g., conditions were such that if the assassination of the Archduke had not caused the outbreak of war, some other event would have.

Strevens seems to find causal potentialism a puzzling position, at least within a framework like mine (he writes, "back up causes,

although …not actual causes nevertheless help to explain the event they did not cause,'' apparently meaning to suggest that there is something paradoxical about this) but I don't see the problem. Of course it is incoherent to combine the view that non-actual back up causes explain with the view that anything that explains is an actual cause (thus obliterating the distinction between actual and back up causes—the ''philosophical own goal'' that Strevens worries that I may have committed) but causal potentialism need not take this form; it can retain the distinction between actual and merely potential causes and say that both are explanatorily relevant. (That is, one cites both in the explanation while of course also identifying which are the actual causes and which are the merely potential ones.) As nearly as I can see, someone favoring a broadly interventionist treatment of causation and explanation can quite consistently adopt such a view—and perhaps I should have done so in *MTH*. Again, though, this isn't to say that there is anything inconsistent about *rejecting* causal potentialism within an interventionist framework—it is just that acceptance of this doctrine may be better motivated.

## 4. Interventions

As Strevens notes, the notion of an intervention is crucial to my treatment of causation: on my view, causal claims have implications for what would happen if various interventions or combinations of these were to occur. Although Strevens connects the notion of an intervention to ''God's descending and directly tweaking the relevant factor'' and to Lewisian ''small miracles,'' the notion is introduced and motivated in *MTH* in a non-theological, non-metaphysical and much more down to earth way. An intervention on a variable $X$ is always defined with respect to a second variable $Y$ and can be thought of as an ideal experimental manipulation (of a sort that might be realized in a randomized experiment) of $X$ for the purposes of determining whether $X$ bears one or another kind of causal relationship to $Y$. As I explain in *MTH*, this means that the manipulation of $X$ should be such that various confounding possibilities (e.g., that the intervention $I$ affects $Y$ via a causal route that does not go through $X$, so that $Y$ changes under the manipulation but not because $X$ causes $Y$) are excluded. Providing a characterization of the notion of an intervention that works properly for this purpose turns out to be a non-trivial matter. In order to save space, I will not reproduce my characterization here, but the reader is referred to the conditions **IV** and **IN**, p 98, *MTH*.

Strevens claims that this characterization is ''implicitly relativized to the variables in a causal network.'' I assume that he means that

relativization is to the variables used to characterize the causal relationships in the system in which the intervention occurs. (Note that this claim is different from and much stronger than the observation that an intervention $X$ is always defined with respect to a second variable $Y$ which corresponds to the putative effect. Strevens is claiming that the intervention is relativised to *all* of the variables in the system in which the intervention occurs.) Strevens further claims that this relativization has various consequences that are highly unwanted from my point of view.

Consider, to use Strevens' example, a system in which fancy water consumption $W$ and heart disease $H$ are correlated but only because they are joint effects of a common cause—consumption of salty food $S$. Suppose that we are working with a representation (the $WH$ representation) of this system that includes the variables $W$ and $H$ but does not include the variable $S$. Strevens claims that it is a consequence of my characterization of an intervention that "increasing the amount of water you drink will count as an intervention relative to the salt—free network" (that is, the $WH$ representation). Moreover, according to Strevens, under such an intervention, the chances of heart disease will go up, and hence it follows from my account that $W$ causes $H$. According to Strevens, I make sense of all this by relativising the notion of causation itself to a set of variables: I claim that $W$ causes $H$ with respect to the variable set $WH$ but not with respect to the larger variable set $WHS$ since increasing water consumption will no longer count as an intervention with respect to this larger variable set.

I say more about "variable relativity" below, but a look at **IN** makes it clear that there is no explicit or obvious relativizaton to a variable set of the sort that Strevens has in mind. In particular, **IN** is formulated in terms of requirements that concern the relationship between the intervention variable $I$ and "*other (contributing) causes*" of $Y$, the putative effect variable, and not in terms of the relationship between $I$ and the other causes of $Y$ that are in some particular variable set **V** or that are known to the experimenter. In other words, the intervention must be uncorrelated with *all* potential confounders, not just with all confounders that happen to be in some variable set such as the one we use to describe the system in which the intervention occurs. Thus, contrary to what Strevens claims, to count as an intervention on $W$ with respect to $H$, the manipulation of water consumption must not be correlated with $S$. There is no such thing, from my point of view, as the manipulation counting as an intervention with respect to $W$ and $H$ but not with respect to $W$, $H$ and $S$. Indeed, it was precisely to avoid consequences like those described by Strevens (that $W$ causes $H$ with respect to one variable set but not with respect to another) that I very deliberately elected *not* to relativize the notion of an intervention to a variable set.

How might such a (non-relativized) intervention on $W$ with respect to $H$ be accomplished? As I explain in *MTH*, one way is by means of a randomized experiment in which subjects are assigned different levels of expensive water consumption, independently of whether or not they consume salty foods. Presumably under such an intervention $W$ and $H$ will not be correlated and hence (according to the way I characterize various causal notions) $W$ will not count as a cause of $H$ (in any sense). In fact, this illustrates one of the main virtues of a randomized experiment (which the notion of an intervention is meant to capture): when you successfully carry out such an experiment you remove correlations between the putative cause and effect that are due to *all* potential confounding causes, even those that are unknown or unobserved or "invisible," and not just confounders that are in some particular variable set.

I take such considerations to show that in developing an interventionist account of causation, one must use a non-relativized notion of intervention. I also take it, however, that Strevens thinks that I am not entitled to such a non-relativized notion of intervention. If I have understood him correctly, he thinks that it is a consequence of other things I say that causation itself is "relative" to a variable set.

*MTH* does *not* make this claim (at least when interpreted in anything like what Strevens intends). I conjecture that Strevens is led to this misinterpretation by two portions of my discussion: my treatment of how causal *judgment* is influenced by choice of variable set and my definitions of the various causal notions, including in particular direct causation.

## 5. Interventions and Circularity

I take up both of these issues below, but first I want to address some issues concerning "circularity" in my characterization of interventions. As I explicitly acknowledge in *MTH*, this characterization is nonreductive in the sense that it makes heavy use of causal notions; not only must an intervention $I$ on $X$ with respect to $Y$ cause $X$ to assume a certain value, but $I$ must bear certain relationships to other causes of $Y$ that are independent of $X$ and so on. I argued that this sort "circularity" is less objectionable than many philosophers have supposed. I will not repeat these arguments except to note that one of my claims was that the interventionist account is not epistemically viciously circular in the sense that to determine whether $I$ counts as an intervention on $X$ with respect to $Y$, one has to already establish or know whether $X$ causes $Y$. Determining whether $I$ is an intervention on $X$ with respect to $Y$ requires background causal knowledge but this is knowledge about *other* causal relationships besides whatever causal relationship may or may not hold between $X$ and $Y$.

Strevens contests this last claim, contending that my account *is* epistemically vicious. He writes:

> As you can immediately see, it is impossible to determine whether a manipulation is an intervention on X relative to V without knowing about the causal pathways that connect the members of V, which is to say, without consulting a causal network for V.

I think this claim is mistaken for several reasons. First, as already explained and contrary to what Strevens suggests in the above quotation, the notion of an intervention is not relativized to the entire variable set **V** characterizing the system in which the intervention occurs. Second, the characterization **IN** makes it clear that the conditions for whether a manipulation $I$ counts as an intervention on $X$ with respect to $Y$ do not concern *every* feature of the system of causal relationships in which $I$ occur*s* but only *some* of those features: the relationship between $I$ and other causes of $Y$ that are off the $I \rightarrow X \rightarrow Y$ route and so on. As long as the conditions in **IN** are met, it doesn't matter what other causal relationships hold in this system.

Finally and most fundamentally, one can't move in the way that Strevens does from the *characterization* of an intervention to what you have to *know* in order to know you have carried out an intervention. These are two very different issues. In particular, as explained above, if I carry out an appropriately designed randomized experiment, I can know that I've performed an intervention on $X$ with respect to $Y$, even though I don't know what the other causes of $Y$ are or various other facts about the network of which $X$ and $Y$ are a part. Again, this is one of the main reasons why randomized experiments are methodologically attractive: they allow you to perform an intervention (and to know that you have done so) without requiring that you have detailed, specific knowledge about the relationships among the other variables in the system of interest.

## 6. The Relativity of Causal *Judgment* to a Variable Set

Consider a pair of examples, the first originating with McDermott, 1995, and discussed by Collins, 2000. In the first, a ball moves toward a glass window but a solid brick wall is in the path of the ball. A fielder catches the ball before it reaches the wall. The second example is just like the first except that the wall is replaced by a second fielder who would have caught the ball if the first fielder had missed it. People seem to disagree in their judgments about whether the fielder in the first scenario prevented the window from breaking, but Collins claims and I agreed in *MTH* that we are more likely to judge that the first

fielder prevented the window from breaking in the second situation than we are in the first situation. (I intended this as an empirical claim about our practices of causal judgment,[4] although I also think it is a plausible normative claim about the judgments that we should make in these situations.) Following Collins, I traced this difference in judgment to differences in what we regard as serious possibilities in the two situations. One's judgment that, in the first situation, the fielder did not prevent the window from breaking, is connected to one's judgment that it is not a serious possibility that if the fielder had missed the ball, the ball would somehow have passed through the wall and then have shattered the window. Given that this is not a serious possibility, whether the window breaks does not depend on whether the first fielder stops the ball and hence we do not take the first fielder's action to have prevented the breaking of the window. By contrast, to the extent that one judges, in the second situation that the first fielder prevented the window from breaking, this is connected to one's judgment that it is a serious possibility that if the first fielder had missed the ball the second might have done so also. (Just to clarify: in the actual situation it is stipulated that the second fielder would have caught the ball if the second fielder missed but as a matter of empirical fact fielders are substantially less reliable at stopping balls than brick walls.) I suggested that this difference is connected to (or can be reflected in) how we choose to model or represent the two situations in terms of equations or directed graphs—in second scenario, if we think that it is a serious possibility that the second fielder might have missed, we should include a variable corresponding to whether the second fielder catches the ball or not in our representation of the causal structure of the situation. By contrast, we should not include variables corresponding to possibilities we do not regard as serious. Such choices of variables in turn influence the causal judgments we make if we are guided by the various treatments of causation in *MTH*.

As this discussion should make clear, the "variable relativity" that is present in this example involves a relativity of causal *judgment* to which variables are judged to correspond to serious possibilities and hence are included in the representation which is meant to capture or correspond to that judgment. Strevens seems to have inferred from this, however, that I hold that causation itself (in the various forms discussed in *MTH*) is somehow "relative" to a variable set; so that I am committed, for example, to the claim that the fielder prevents the ball from shattering the window relative to one variable set but not relative to another.

[4] If you think this particular claim is mistaken, there are a number of other examples in *MTH* that illustrate the same point.

It seems obvious, however, that it does not in general follow from the fact that a representation of some state of affairs has feature *F* (in this case, a kind of relativity to a variable set) that the state of affairs so represented (the causal facts themselves) must have feature *F*.[5] Indeed, while I think that I understand what it means to say that causal judgment is influenced by (or relative to) a variable set, I have no clear idea what it would mean to say that whether the catch causes (in the sense of being a total or contributing cause, as characterized in *MTH*) the window not to shatter is somehow relative to a representation.[6]

Here is how I proposed (*MTH,* pp. 90-91) that we should think of the fielder examples (and other similar examples involving considerations having to do with what is a serious possibility). There are facts (" out there in the world" if you like) about what will happen to some variables when one intervenes on others or combinations of others—facts about patterns of counterfactual dependence or, as I sometimes call them, dependency relationships. For example, in the first (fielder/wall) situation, there are empirical facts (facts about patterns of counterfactual dependence) concerning what would happen to the window if the fielder had missed, and the ball had struck the wall, facts about what would have happened if the ball had somehow passed through the wall and so on (As explained earlier, I do not claim that these facts about counterfactual dependence are metaphysically primitive or fundamental—I take them to be just ordinary empirical facts.[7]) When we make causal claims or construct causal explanations, we attempt to represent some features of these facts, but our representations are, in various ways, partial or incomplete in the sense that they

---

[5] Suppose for the sake of argument that (contrary to what I think) that this entailment does hold: that (a) causal judgment has the features under discussion (it is influenced by judgments of serious possibility etc) entails (b) the conclusion that causation itself is "variable relative." If Strevens holds that (a) entails (b), he can only avoid (b) by rejecting (a). But (a) seems fairly plausible, just as a matter of empirical fact, as the fielder example shows. So if he wants to avoid (b), a better strategy it to reject the claimed entailment.

[6] This claim about the variable relativity of causation itself (or some particular concept of it) should be distinguished from the claim (a) that it is true in the two fielder situation but false in fielder/wall situation that the catch prevents the window from breaking. I'm happy to endorse (a) and it does makes the causal facts depend *on certain features of the world*—that balls are often missed by fielders but do not pass through brick walls etc—in a way that is inconsistent with certain standard accounts of causation. However, it invites endless confusion to describe (a) as saying that causation is "variable relative."

[7] In the case of dependency relations that hold in the special sciences and in common sense there will be scientific explanations, formulated in terms of deeper theories for why these dependency relations hold, but the generalizations and initial conditions to which these explanations appeal are (in my view) ordinary scientific facts, not metaphysical ones.

never represent all true dependency relationships in the world. One way in which they are incomplete is that we represent dependency relationships only among a limited set of variables, not all possible variables, and this in turn affects (or is reflected in) the causal judgments we make. For example, our causal judgments in the wall case don't reflect facts about what would have happened if the ball had passed through the wall , since this possibility is not serious. Note what this does NOT say: It does not say. e.g., that if it is not a serious possibility that the ball penetrates beyond the wall, there is then no fact of the matter about what would happen if it were to so penetrate. There are such facts, but they just don't get reflected in our causal judgments.

## 7. The "Variable Relativity" of the Various Causal Notions

I turn now to a more explicit look at the various causal notions characterized in *MTH* (including direct causation) and what these imply about "variable relativity." Let me begin with a correction/concession. In *MTH* two causal notions—direct causation and contributing causation—are defined "with respect to a variable set **V**," largely because the focus of my inquiry was on how one might connect these notions with certain features of directed graphs and sets of equations which of course involve some particular set of variables. In my view, this "with respect to" locution does not imply that either direct causal or (more importantly) contributing causation are "variable relative" in the sense Strevens has in mind—indeed, it didn't occur to me that anyone would draw this implication. Nonetheless I now see that my choice of language was potentially misleading, and that I should have been more explicit in spelling out just what the "with respect to" aspect of my characterizations does and does not imply, and what the (unrelativized) facts are that lie behind the features of representations on which I focus. I am grateful to have the opportunity to try to make all this clearer that I did in *MTH*.

Suppose I start with a single variable $X$—this is the sole member of my variable set $\mathbf{V_1}$. Consider a second variable $Y$ that is not in $\mathbf{V_1}$ . In *MTH,* I define a notion of total cause as follows:

> (**TC**) $X$ is a total cause of $Y$ if and only if there is a possible intervention on $X$ that will change $Y$ or the probability distribution of $Y$.

It should be obvious from **TC** that whether $X$ is a total cause of $Y$ does not depend on whether I am operating with $\mathbf{V_1}$ or some other variable set. All that matters is whether it is (i) true or (ii) false that under some intervention on $X$, the value of $Y$ would change. Of course, in case (i), we won't be able to represent the fact that $X$ is a total cause of $Y$ using

just $\mathbf{V_1}$, but *MTH* does not claim (and I see no reason to take this observation to show) that the notion of total causation is itself somehow relativized to a variable set, so that we are forced to say that $X$ is not a total cause of $Y$ relative to $\mathbf{V_1}$ but is a total cause relative to an expanded variable set that includes $Y$. (Note that what is defined in **TC** is a notion of total causation, not a notion of total causation with respect to variable set **V**.) Observe also that if $X$ is a total cause of $Y$, and I include $X$ and $Y$ in my variable set, so that $X$ is represented as a total cause of $Y$, then adding additional variables to the variable set will never have the consequence that $X$ is represented as not a total cause of $Y$—that is, even at the level of representation it is not true that adding variables to the variable set has the consequence that $X$ will be represented as a total cause of $Y$ with respect to one variable set but represented as not a total cause of $Y$ with respect to an expanded variable set. The representation of total cause relations is in this sense "conserved" under the addition of new variables.

Turning now to the notions of direct causation and contributing causation, matters are (as I acknowledged in *MTH* but perhaps not as fully as I should have) more complicated, but again I do not see that my characterization of either of these notions leads to the kind of radical relativization to a variable set that Strevens has in mind. Consider first the notion of direct causation. Intuitively, $X$ is a direct cause of $Y$ if there is a causal relationship between $X$ and $Y$ that is not mediated by some third variable $Z$. In a directed graph representation, a direct causal relationship between $X$ and $Y$ is represented by an arrow drawn directly from $X$ to $Y$. This informal characterization as well as the more precise technical definition of direct causation in *MTH* provides one sense in which it seems correct to say that the notion of direct causation is "relative to a variable set." The sense is this: $X$ may count as a direct cause of $Y$ relative to some variable set **V** but not with respect to some expanded variable set **V'** which contains a variable $Z$ which is causally intermediate between $X$ and $Y$. It seems to me that any notion of direct causation that is connected to a directed graph representation will need to be "variable relative" in this sense—it is not as though there is some other useful characterization of direct causation that does not have this feature.

It is important to be clear, however, about exactly what follows from this. Suppose I begin with variable set **V** in which $X$ is represented as a direct cause of $Y$. This corresponds to a perfectly objective set of facts: that if I fix the other variables in the variable set **V** at some value, there is an intervention on $X$ that will change the value of $Y$. Suppose now I move to the expanded variable set **V'** in which $Z$ is causally between $X$ and $Y$—i.e., $X$ is represented as a direct cause of $Z$

which is in turn represented as direct cause of $Y$. This too corresponds to an objective set of facts which are *consistent* with the facts previously described: that if I fix the value of this new variable $Z$, there is no intervention on $X$ that changes the value of $Y$. (These facts themselves are not variable relative or model relative except in the innocuous sense that they concern what will happen if one intervenes on one set of variables and not others.)

It is true that the representation of direct causal relationships is not "conserved" under the addition of new variables and this marks an important difference with the representation of total cause relationships. But if $X$ is represented as a direct cause of $Y$ in variables set **V** (which also means that it will count as a contributing cause) and not so represented when a new variable $Z$ is added, it still will be true (given the way that I define the notion of contributing cause) that $X$ will count as a contributing cause of $Y$ in the new expanded variable set that includes $Z$. That is, the representation of contributing causal relationships *is* conserved under the addition of new variables . Within a directed graph representation, arrows between variables can disappear as we add new variables, but a parallel claim is not true of the representation of contributing and total causal relationships.

As remarked above, my characterization of contributing cause in the condition (**M**) (p. 59) does define the notion of contributing cause "with respect to a variable set **V**." I now see that this was potentially confusing. A better way of putting matters would have been to say that the condition **M** characterizes what it is for $X$ to be correctly represented as a contributing cause of $Y$ with respect to **V**. Understood in this way, **M** says is that $X$ is "correctly represented as a contributing cause of $Y$ with respect to **V**" if there is a chain of direct causal relationships (a directed path) leading from $X$ to $Y$[8] and if when one fixes variables that are off that path at some value, an intervention on $X$ changes the value of $Y$. One can then go on to say that $X$ is a contributing cause of $Y$ *simpliciter* (in a sense that isn't relativised to any particular variable set **V**) as long as it is true that there exists a variable set **V** such that $X$ is correctly represented as a contributing cause of $Y$ with respect to **V**.

Let me add that even in the case of direct causal relationships, acknowledgment of "variable relativity" does not have the consequences

[8] Recall that once one fixes the variable set **V** it is an objective matter, characterizable in terms of what happens under interventions on the variables in **V**, what the direct causal relations in **V** are. $X$ is correctly represented as a direct cause of $Y$ with respect to **V** just in case these claims about what will happen under interventions are correct. The notion of a correct representation of contributing cause relationships is understood similarly.

Strevens supposes. Consider again the example in which $S$ = consumption of salty foods is a common cause of $W$ = fancy water consumption and $H$ = heart disease, and suppose that respect to this variable set all causal relationships are direct. Contrary to what Strevens suggests, it simply does not follow from my characterization that with respect to the variable set that includes only $W$ and $H$, $W$ counts as a direct cause of $H$. $W$ would only be a direct cause of $H$ with respect to this variable set if, when I fix other variables besides $W$ and $H$, (in this case there are no such variables so this condition is trivially satisfied), an intervention on $W$ will be associated with changes in $H$ and this is not the case.

What follows from these observations about the status of claims about direct causation? I argued in *MTH* that the notion of direct causation is useful for many purposes, including the representation of what will happen under combination of interventions and for formulating connections between causal and probabilistic claims. The notion is also assumed whenever one employs direct graphs or structural equations to represent causal relationships, which is a very common scientific practice. I agree with Strevens, however, that it is hard to see how direct causation can be "metaphysically fundamental" (to the extent that I understand this notion). As Strevens says, this is perhaps most obvious when one is dealing with a system or processes whose development in time is continuous such as the trajectory of a billiard ball. Given such a system and some variable set **V** with respect to which certain relationships in the system are represented as directly causal, we can, if we wish, always move to a finer grained variable set (e.g., by adding temporally intermediate variables) in which those relationships are not represented as direct. What I infer from this is that any representation of direct causal relationships in such a case is going to be partial or incomplete, in the sense that it will leave some things out or fail to represent some causal relationships, not that the representation is false or useless. That is, the original representation may be true or accurate regarding what it does represent even if it does not represent everything.

To illustrate, consider Strevens example of the initial break in a game of pool. Suppose that my hitting the cue ball ($H$) causes this ball to strike ($S$) a collection of balls at the center of the table in a certain way, causing a series of collisions with the result that the eight ball to drops into the corner pocket ($D$). Relative to this variable set, the direct causal relationships will be represented by $H \rightarrow S \rightarrow D$ and, as far as it goes, this representation is (we may suppose) true or accurate: fixing $H$, intervening to alter $S$ will alter whether $D$ occurs but intervening to alter $D$ will not alter $S$ and so on. At least in this respect, I can usefully employ a representation that represents direct causal

relationships to describe what happens in a system that evolves continuously. Of course there are other aspects of this situation that cannot be represented (or at least usefully or helpfully represented) by directed graphs of the sort that I employ in *MTH*: for example, the fact that the evolution of the system of balls is itself a continuous process.

Strevens observes in this connection that:

> Woodward's entire causal apparatus, and his notion of direct causation in particular, is founded on the supposition that causal networks represent less of causal reality than is actually out there.

I agree but don't see this as pointing to a fatal deficiency in the use of directed graphs or claims about direct causation to represent causal relationships. All of the representational devices used in science and common sense with which I am familiar are like direct graphs in being useful for some purposes (and for the representation of certain kinds of structures) but not others. All such devices are partial or incomplete in the sense of not representing *everything* that is "actually out there".[9] Perhaps the aspiration of the metaphysician of causality is to find a form of description that represents "all" of "causal reality" in a complete, non- partial way that is untainted by any purpose- relative human concerns (i.e., the sort of description that God would produce, if only He existed) but this isn't my project.

## 8. Variable Choice

A final point about the variable relativity of causal representation that does not emerge as clearly as it might in Strevens' discussion is this: It does not follow (and *MTH* does not claim that) that once one gives up on the idea of including all possible variables in a representation, any choice of variable set is as good as any other. On the contrary, I assumed in *MTH* that one can formulate rationally defensible non-arbitrary considerations or guidelines about which variables to employ in representing different sorts of situations—considerations that will

[9] Lest this seem unduly dismissive, let me add that I think that Strevens is entirely right to raise the general issue of the representational limitations of directed graphs—this deserves more attention than it has hitherto received. Here are some limitations in addition to those mentioned by Strevens: directed graphs represent that certain functional relationships exist but don't tell us what those relationships are, they don't (in the form employed in *MTH*) represent general structural features about possible causal relationships among classes of variables (cf. Tenenbaum and Niyogi, 2003), and they don't represent the difference between deviant and default values of variables in the sense of Hitchcock, forthcoming. Arguably some of these limitations can be addressed by enriching the directed graph representation rather than simply abandoning it, as the last two papers suggest.

disqualify some possible representations although they may not always pick out a uniquely best one. To the extent that there are such considerations, they will serve to rule out some wilder forms of relativity in causal representation. Moreover, contrary to the impression that the reader is likely to form on the basis of Strevens' review, a number of different suggestions about such considerations governing variable choice can be found in *MTH*, although I would be the first to concede that there is much more to be said on this topic.

## References


Collins, J. 2000. Pre-emptive Prevention. *Journal of Philosophy* 97: 223–34.

Hall, N. Forthcoming "Structural Equations and Causation."

Hitchcock, C. Forthcoming. Prevention, Preemption, and the Principle of Sufficient Reason. *Philosophical Review*.

Jackson, F. and P. Pettit 1992. In Defense of Explanatory Ecumenism. *Economics and Philosophy* 8: 1–21.

McDermott, M. 1995. "Redundant Causation." *British Journal for the Philosophy of Science* 46: 523–44.

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

Tenenbaum, J. and Niyogi, S. 2003. "Learning Causal Laws." *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society*.