Preamble for the Woodward-Spohn Exchange in *Philosophy of Science*

The authors of the following two discussion notes have been working on causation for about 40 years. Woodward's work culminated in his book *Making Things Happen* (OUP 2003), Spohn's work culminated in chapters 14 and 15 of his book *The Laws of Belief* (OUP 2012), which have almost book-length. *Prima facie*, their accounts look quite similar; the interventionist theory of causal Bayes nets seems to be their common ground. Therefore, it is important to also see their differences. These are explained in the notes, which mutually discuss their theories of causation. The exchange originates from an Author Meets Critic session at the APA meeting in Baltimore in January 2017.

On Wolfgang Spohn's *Laws of Belief*

James Woodward

Department of History and Philosophy of Science

University of Pittsburgh

Abstract: This note compares some features of the account of causation in Wolfgang Spohn's *Laws of Belief* (2012) with the "interventionist" account in Woodward (2003). Despite striking similarities there are also important differences. These include (i) the "epistemic" orientation of Spohn's account as opposed to the worldly or "ontic" orientation of the interventionist account, (ii) Spohn's focus on token- level causal claims in contrast to the primary interventionist focus on type-level claims, (iii) the role of temporal information in accounting for causal asymmetry and (iv) the extent to which causal claims are "frame-relative".

1. Introduction

Wolfgang Spohn's *Laws of Belief* (*LB*) is an extraordinarily rich book, ranging over a vast number of topics, including the representation of belief, rules for belief change, confirmation, laws of nature, probability and alternatives to it, and much else. All of this is discussed within the framework of "ranking theory"—the highly original and fertile framework for doxastic representation that Spohn has developed. *LB* will more than repay careful study by anyone interested in formal epistemology and philosophy of science. In these comments, I focus on Spohn's ideas about causation—a topic to which Spohn has made important contributions and which occupies (at least) 100 or so pages in *LB*.

Before doing so, however, I want to draw attention to one portion of *LB* (or really Spohn's overall corpus): the "laws" of conditional independence described in Chapter 7. These laws, the surrounding results and connections to causation reflect earlier innovative work by Spohn from 1970s and early 1980s which had an important impact on the development of the framework of Bayes nets by Judea Pearl and others. (Pearl's notion of a graphoid encodes some of these ideas.) This is a striking example of a philosopher doing work which ended up contributing importantly to statistics, artificial intelligence, and causal inference—work which still hasn't gotten the attention it deserves in the wider philosophical community.

2. Spohn on Causation

In what follows I will make use (as Spohn does) of the framework of directed graphs for representing causal relationships in which an arrow drawn from the variable X to the variable Y ($X \rightarrow Y$) indicates that X is a direct cause (or parent) of Y. Variables on a directed path originating with X are descendants of X. A standard assumption is that there is a probability

distribution *P* over all of the variables of the graph's variable set **V**. A graph *G* and a probability distribution *P* over **V** satisfies the *Causal Markov condition* (in one standard formulation) if and only if every variable in *V* is independent in *P* of all of its non-descendants conditional on its parents. A probability distribution *P* is *faithful* to a graph *G* iff the only independence relations exhibited by *P* are those that are implied by *G* alone. Taken together these two conditions, Markov and Faithfulness, constrain the graphs that are consistent with the probability distribution *P*—a fact which can be used to guide inference from *P* to causal structure. In some cases, only a single graph will be consistent with *P*; more commonly there will be an equivalence class of graphs consistent with *P*—an observation to which I will return below.

Turning now to Spohn's account of causation, his approach is, as he explicitly says, Humean in spirit. According to Spohn, what we find in the world are non-modal facts about properties possessed by objects or values taken by variables at particular times and locations. The modal element in causation comes in, as it does in Hume, as a result of our "projection" on to the world of our belief-regulating or "epistemic" activities, although on Spohn's view to some extent this projection can admit of a kind of "objectification" (see below). Spohn's account is thus an epistemic theory of causation in the sense that what we have reason to believe and how this changes under additional information is the primary notion from which, when combined with non-modal wordly information, causal notions are constructed. In this foundational respect, I see Spohn's account as contrasting in a fundamental way with the interventionist approach I favor, which instead sees causation as having to do with what happens in the world when we or nature *do* things, where "doing" has to do with the manipulation of worldly items and not just patterns of belief formation. Although this difference is important, it is also very interesting to ask about the extent to which each approach can replicate the results of the other and at which points the two diverge. I discuss both in what follows.

Spohn's point of departure is what he calls "singular" or token-level causation involving particular causal processes—for example, how "my income affects my first son's educational opportunities" (*LB*, 341). Such singular causal claims relate what Spohn calls specific variables with specific temporal locations, such as a variable whose values are possible incomes for Spohn at specific times. These contrast with "generic" or type-level variables such as parental income and causal claims involving them, such as claims about how parental income causally affects children's educational opportunities in general in some population.

In developing his account of singular causation, Spohn assumes a framework of "temporal points" that can be used to specify relations of temporal precedence among values of specific variables. Spohn also assumes that *A* can be a cause of *B* only if *A* is not later than *B*. (As an aside, I note that Spohn does allow for simultaneous causation, although much of his discussion abstracts away from this possibility, assuming instead that if *A* causes *B*, *A* is temporally prior to *B*. My discussion below also adopts this assumption.) Spohn relies heavily on this assumption in his account of causation—it yields an account of causal asymmetry and contributes to his account of what needs to be held fixed or controlled for in assessing whether *A* causes *B*.

The following characterization, which Spohn describes as "preliminary", captures the core idea of his account:

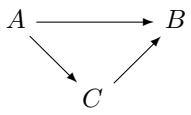
> (**C**) *A* is a cause of *B* iff *A* and *B* obtain, *A* precedes *B*, and *A* raises the metaphysical or epistemic status of *B* given the obtaining circumstances. (*LB*, 352)

What does "raise the metaphysical or epistemic status" mean? As noted above, Spohn's preferred account is in terms of ranking theory—the idea is that *A* provides something like a reason for belief in *B* or boosts the believability of *B*. As Spohn puts it, "causes are a specific kind of conditional reason". This is the epistemic status version of the core idea, and with

reference to which the metaphysical status version is to be explained. As Spohn makes clear, ranks behave in many (although not all) respects like probabilities. Thus in the present context we will not go far wrong if we understand "status raising" in terms of positive probabilistic relevance when the context is probabilistic and something like the obtaining of a necessary and/or sufficient condition when the context is deterministic.

(**C**) is, as I have said, the core idea, but Spohn proceeds by first defining a notion (or notions) of direct cause via a more specific version of (**C**) and then extending this to other varieties of causation. I won't describe his definition in detail—it is complicated—but here are some of the central ideas. First, the "obtaining circumstances" in (**C**) are understood as including the entire past of the effect of the direct cause except of course for the direct cause itself. Second, the definition of direct cause is relativized in two distinct ways—it is relativized to a choice of ranking function and also relativized to what Spohn calls a "conceptual frame" **U** which includes a specification of the variables we employ.

To illustrate how Spohn's account is intended to work, suppose *A* is both a direct cause of *B* and an indirect cause of *B* through a third variable *C*—that is, represented graph-theoretically, we have the following structure:



(Since Spohn's is an account of singular causation, we should think of this graph as representing a token-cause structure relating values of specific variables in Spohn's sense.) According to Spohn's proposal, to capture the idea that *A* is a direct cause of *B*, we hold fixed the entire past before the effect *B*, except for *A* itself. This includes holding fixed *C* at its actual value since (we are assuming) *C* must be temporally prior to *B* if it causes *B*. If, under these conditions, *A* raises the epistemic status of *B* (e.g., is positively relevant to *B* etc.), then

A is a direct cause of *B*. It is worth noting, and Spohn explicitly observes, that an "interventionist" characterization of direct causation looks very similar. Woodward (2003, 55), for example, characterizes a notion of "direct cause" roughly as follows[1].

> (**DC**) *X* is a direct cause of *Y* with respect to some variable set **V** iff there is a possible intervention on *X* that will change *Y* (or the probability distribution of *Y*) when all other variables in **V** are held fixed at some value by interventions.

Applied to the example above, **DC** tells us to fix *C* at some value via an intervention—in the sort of case considered by Spohn its actual value will do—and then determine whether some distinct independent intervention on *A* will change the value of *B*. It will do so if and only if *A* is a direct cause of *B*. As it happens this is also sufficient for *A*'s taking its actual value to qualify as an "actual" or "singular cause" of *B*'s taking its actual value within an interventionist framework, at least on one characterization of "actual cause" (Woodward, 2003).

Despite this, there are, as intimated above and as Spohn is well aware, important differences between the interventionist framework and his own. One important difference is that the interventionist framework does not aspire to provide a reduction. This is a consequence both of the fact that the notion of intervention itself is a causal notion and for other reasons as well. By contrast, Spohn's account is reductionist in aspiration. Roughly speaking, temporal information and assumptions about the relation between temporal and causal order play the role that explicitly causal assumptions play in the interventionist account—in particular, such assumptions tell us what we need to control for or hold fixed

---

[1] In addition to the differences described below, DC differs from Spohn's treatment in that DC is intended to capture a type-level causal notion.

when we ask whether something is a direct (or any other kind of) cause. Spohn's assumptions about the relationship between causal and temporal order also help to solve underdetermination problems that would otherwise imperil his reductionist project. Spohn argues that it is an advantage of his account that provides such a reduction, thus avoiding the "circularity" that infects non-reductionist accounts

3. Some Issues Raised by Spohn's Account

Spohn's appeal to temporal considerations raises several important issues that I can only discuss briefly. First, as Spohn is well aware, there is an obvious epistemological concern with this appeal —as he puts it, "it is impossible for us to know and take into account the whole past of the effect (unless we consider only very small frames)". One conclusion he draws is that "our causal judgment is always bound to be preliminary" (*LB*, 360). He does not intend this merely as a claim about judgments of direct causation. That is, he does not mean merely that our judgment that *A* is a direct cause of *B* may change when we become aware of an additional variable *C* that lies on the only causal route from *A* to *B*, so that *A* is now judged to be an indirect cause of some kind for *B* but still a cause—e.g., perhaps a "contributing cause" in the framework of Woodward (2003). Rather Spohn holds that depending on the stock of variables we have available for characterizing the past of *B*, whether *A* causes *B* in any sense (directly or in some other way) can flip back and forth indefinitely as we consider more or less detailed descriptions of that past. It is only with reference to a properly behaved "universal" frame or a maximally detailed description of the past that such judgments will stabilize. And of course we usually and perhaps always don't have such a description. I will return to this issue below.

A second general issue is this: As Spohn is well aware, we often make (apparently reliable) causal judgments in the absence of the kind of temporal information to which he

appeals and we often make such inferences involving generic, non-specific variables (rather than the "specific" variables and their values which Spohn prefers as causal relata). For example, causal modeling techniques often are applied to cross-sectional data in which we are given information about the distribution of various variables *X*, *Y*, *Z* in a population, perhaps observed at roughly the same time, but with no temporal information that would allow us to causally order these variables. Indeed, as a matter of the way in which they are defined and measured, many variables of interest to social and behavioral scientists may not have the kinds of temporal location that allows for the temporal ordering Spohn wishes to make use of—in other words, the appropriate temporal information may not be available for conceptual reasons, rather than merely as a consequence of our ignorance. As an illustration, consider data on individual health status and wealth at age 50. It is unclear in what sense, if any, one of these variables is temporally prior to the other but we still might investigate their causal relation. Of course it might be claimed that it is always the case that "underlying" such variables are much more fine-grained causal relationships that are fully temporally ordered, but I would again emphasize that such information is often not epistemically accessible to investigators and that causal analysis is often carried out in the absence of such information. Indeed it is common in the causal modeling literature to employ non-recursive causal models which contain causal cycles—the graphical representation is still in terms of directed arrows, but these arrows obviously cannot represent temporal priority[2]. In addition, sometimes the temporal scale on which it is possible to measure changes in variables of interest is much slower then the temporal scale that characterizes the causal dynamics among those variables, so that the former information is relatively uninformative about the latter. This is the case, for

[2] Put differently, We should not confuse the claim that causal relationships come with a direction (reflected in the use of *directed* graphs) with the distinct claim that causal relationships must be *acyclic*.

example, for fMRI measurements, which at present have a temporal resolution of one to two seconds, while the underlying brain dynamics often proceed on much faster scale. Indeed, as Spohn is well aware, many of the standard frameworks for representing causal relationships and for causal inference—structural equations and directed graphs and the representations and inference procedures described by Pearl and in Spirtes, Glymour, and Scheines (2000)—make little or no use of temporal information. Relatedly, in actual applications, the target of such techniques is rarely particular singular causal relations (the causal relationship between Spohn's income and his son's educational attainment or between Jones' smoking and his lung cancer) but rather causal claims that are much more generic—e.g., the causal relation between parental income and son's educational attainment or between smoking and lung cancer in some population[3].

Spohn is, as I have said, fully aware of these points—indeed precisely because standard causal modeling techniques abstract away from temporal information and don't in standard applications yield information about singular causal relationships he regards them as unsatisfactory as a foundation or starting point for a philosophical account of causation. I assume that for Spohn part of the attraction of focusing on singular causal claims is that it may seem prima facie plausible that for each individual causal episode described by a singular causal claim, there is some determinate temporal relationship between cause, effect, and whatever else is going on in the system of interest, even if we don't know this information. This is less obviously true when we move to more "type-level" causal claims.

---

[3] This is not to deny of course that temporal information, when available, can be extremely useful in determining causal direction. I also agree that it would be desirable for causal modeling techniques to make more use of such information when it is available. My point is that in many cases we don't have such information and we don't always need it for successful inference.

It is also worth noting that in taking "singular causation" as his starting point, Spohn inclines to one side of a standard divide in the literature on causation. Philosophers with a more metaphysical orientation tend to regard singular or "actual" causation as primary, and the target on which accounts of causation should focus, as one sees for example in Paul and Hall (2013). Those with a more philosophy of science or methodological orientation, like myself, focus instead on more type-level causal notions, such as those that are the target in standard causal modeling techniques. Although the relation between these two notions has received some philosophical attention, it remains, in my view, underexplored, as do the implications of selecting one of these notions rather than the other as a starting point.

In support of maintaining at least some focus on the latter—the type-level notions—let me make the following observations. First, a great deal of scientific inquiry has one or another sort of type level or generic causal claim as its target. In my opinion it is thus a reasonable project for the philosopher or methodologist to try to understand both these more generic claims and the procedures that are used to establish them (and, when appropriate, to make suggestions for their improvement). Although scientists and engineers are sometimes interested in singular causal claims about particular events (e.g., what caused the Challenger disaster?), more commonly, in both the social and natural sciences, their interest is in causal relationships that are more generic and repeatable: What causes the tides? What is the effect of economic conditions during a U.S. president's first term on his or her re-election prospects?). Standard causal modeling techniques are designed to answer such questions, not to assess the truth of particular singular causal claims, so it is not surprising that they cannot be straightforwardly used to address foundational issues about the latter[3]. That they cannot be

[3] I acknowledge that there have been a number of noteworthy recent attempts to extend the structural equations/ graphical models framework to capture "actual causation"—see, e.g.,

so used may not reflect any fundamental inadequacy in these techniques but rather merely that they are directed at different targets than the assessment of singular causal claims. It also seems to me that such modeling techniques at least sometimes successful in reliably establishing generic causal claims and that they sometimes succeed in the absence of the kind of temporal information to which Spohn's account of causation appeals. It is also worth noting that Spohn does not try to argue that all such applications always yield unreliable conclusions.

A closely related point is this: reliably assessing singular causal claims is often very difficult and in many cases requires a lot of information that may be difficult to get and that may not be required to establish more type level or generic claims. It is one thing to show that smoking causes lung cancer or is causing lung cancer in some population and another matter—much harder—to show of some particular person, Jones, who smoked and developed lung cancer, that his smoking caused his lung cancer. (Indeed even estimating the proportion of subjects in a population whose illness has been caused by exposure to a particular cause, such as smoking, typically requires a great deal of non-statistical "biologic" information that may not be available—a problem well-known to epidemiologists under the heading of estimating "etiological fractions".) In general the problem of inferring "forward" to identify the effects of a cause (e.g., discovering that lung cancer is among the effects of smoking) is very different from the problem of inferring "backwards" from an observed effect to its causes, particularly when the effect is a particular episode (e.g., observing that Jones has lung cancer and identifying what caused this) which is typically what is at issue when establishes a singular causal claim. These two problems (identifying the effects of causes and identifying the causes of some particular outcome) require different kinds of information for their

---

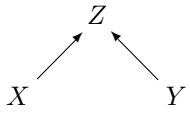
Halpern (2016). But whatever the merits of these attempts, actual causation is not the target notion this framework was originally designed to capture.

solution, a point recognized by a number of methodologists. An additional reason why social scientists and others often focus on more generic or type level claims rather than singular causal claims is thus that in many cases they may have information that enables them to establish the former but not the latter.

4. Causal Order from Non-Temporal Information

In my remarks above, I raised the general issue of the relation between causal and temporal order. In this section I want to very briefly describe some procedures for inferring causal order in the absence of temporal information, both because I think they are of interest for their own sake and because they suggest the possibility of an account of at least some kinds of causal judgment that is less reliant on temporal information than Spohn's. (Spohn says (in *LB*, 341) that he can't imagine how one might proceed in constructing an account of causation without relying on information about temporal ordering—well, here are some possibilities.) I should also say at the outset that what I will describe does not yield anything like a *definition* of causal direction—for one thing the procedures are fallible and don't always work. In this respect they are arguably similar in status to other causal inference procedures which make use of such assumptions as the causal Markov and Faithfulness conditions.

The first example is a very simple one suggested by the inference procedures described in Spirtes, Glymour, and Scheines (2000). It requires both the causal Markov condition and Faithfulness as assumptions. Suppose that $X \not\perp Z$, $Y \not\perp Z$, $X \perp Y$, $X \not\perp Y \mid Z$, where $\perp$ means probabilistic independence, $\not\perp$ means probabilistic dependence and $X \perp Y \mid Z$ means that X and Y are independent conditional on Z. It turns out that given Markov and Faithfulness there is exactly one structure which is consistent with these conditions—namely a so-called unshielded collider structure:



Thus, given the above assumptions, in this particular case dependence and independence information is sufficient to fix causal ordering among the variables in the absence of temporal information. Of course it is not true that in all cases statistical information, even assuming Markov and Faithfulness, will yield a single unique graph or a unique orientation for the edges. But examples like the one above do show that statistical information, in the presence of other assumptions (assumptions to which Spohn is sympathetic), can sometimes settle questions of causal direction/priority even in the absence of temporal information.

Next consider two examples from machine learning (cf. Janzing et al. 2012). Suppose (i) Y can be written as a function of X plus an additive error term that is independent of X: $Y = f(X) + U$ with $X \perp U$. Then if the distribution is non-Gaussian, one can show there is no such additive error model from Y to X—that is no model in which (ii) X can be written as $X = g(Y) + V$ with $Y \perp V$. A natural suggestion is that to determine the correct causal direction, one should proceed by determining which (if either) of (i) or (ii) holds, with the correct causal direction being one in which the cause is independent of the error. Remarkably, this yields results which are accurate in a majority of cases (70 –80%) when applied to real world data in which causal direction is independently known. This success is not magic—although I lack the space for detailed discussion, it has an explanation in interventionist terms. When $X \perp U$ this indicates that there is a source of variation (U) in Y which is independent of the variation in X which is just what we would expect if X causes Y.

Another idea, which does not appeal to considerations about error term independence, is this: Suppose X and Y are statistically dependent. If the causal direction runs as $X \rightarrow Y$, then we should expect that the function f describing this relationship ($Y = f(X)$) will be "independent" of the (marginal) distribution of X—independent either in the sense that we can

observe f apparently changing independently of $p(X)$ and/or $p(X)$ changing independently of f [4] or that f is "informationally" independent of $p(X)$ in the sense of Kolmogorov information. If we find that this is the case and that the corresponding claim is not true of the inverse relationship $X = g(Y) = f^{-1}(Y)$, then we conclude that X causes Y. Again, as an empirical matter, this yields fairly reliable conclusions when applied to real world data in which causal direction is independently known. This idea also has a natural rationale within an interventionist framework: If there is a causal relationship running from X to Y, then we should expect that relationship to be stable or invariant under variations in the value of X or the probability distribution of X but we have no such expectations for the relationship $X = g(Y)$. Informational independence between f and $p(X)$ is one source of evidence that f is invariant in this way[5].

Although, none of these procedures is 100 per cent reliable, I believe they work because they pick up on deep features that causal relationships have—features gestured at above. They provide concrete illustrations of what I had in mind when I said above that there are features of causal relationships that are independent of temporal information that are relevant to causal direction.

---

[4] That is, $p(X)$ is not stationary but the relationship between X and Y described by f remains stable under changes in $p(X)$.

[5] Here is a very simple illustration of the basic idea, although it involves conditional probabilities rather than a functional relationship between X and Y. Suppose that X and Y are statistically dependent. $p(X, Y)$ can the be written as $p(X, Y) = p(X \mid Y)\, p(Y)$ or as $p(Y \mid X)\, p(X)$. Suppose that we find that $p(Y \mid X)$ remains stable under changes in $p(X)$ but $p(X \mid Y)$ does not remain stable under changes in $p(Y)$. Then the strategy under discussion tells us to conclude that the causal direction runs from X to Y.

5. Causal Relativity

Let me next return to an issue that I raised earlier concerning the "relativity" of causal claims. As noted above, on Spohn's treatment, causal claims are 'relative" both to choice of a ranking function and to choice of a "frame" **U** — which for our purposes we can think of as roughly the variables used to characterize the claim and the surrounding system or circumstances. I won't comment on the former issue except to note that according to Spohn the subjectivity associated with choice of a ranking function can be mitigated by means of what he calls "objectification". I do, however, want to comment on the second issue and in particular on Spohn's discussion of the frame-relativity of causal dependence. To save time I will not go into the details of Spohn's definition of one variable causally depending on another (this is a bit different from his notion of direct causation described earlier) except to say that he proceeds by first characterizing direct causal dependence (which is a natural extension of his characterization of direct causation applied to variables) and then characterizes causal dependence as the transitive closure of direct causal dependence. Spohn explicitly acknowledges that both definitions offer only a frame-relative definition of causal dependence (and that this seems prima-facie troubling) writing:

> … direct causal dependence is frame-relative not only due to possible omissions of mediating variables, but also because of incomplete representations of circumstances, and this point carries over to causal dependence as well. This seems to be an unacceptable consequence; there is no such frame-relativity in our ordinary notion of causal dependence, not even in a hidden way. (*LB*, 394)

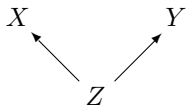
He adds,

> I agree that we have an absolute notion of causal dependence. What I want to argue, though, is that we can gain an understanding of that absolute notion only through the frame-relative one.

His characterization of this absolute notion appeals to the notion of "the universal conceptual frame", comprising "all variables whatsoever". He makes this explicit with the following proposal:

> Explanation 14.39: *Y* causally depends on *X* (simpliciter) if *Y* causally depends on *X* within the universal frame (and relative to a true ranking function on the universal frame). (*LB*, 394)

To get a handle on this, consider the following two structures (*LB*, 396):

(i) 

(ii) 

Spohn says in this case that (i) "reduces to" (ii). He then makes the following assertion (*LB*, 397):

> Assertion 14.42: Assuming universal faithfulness[6], $X \rightarrow Y$ within any given frame means, in absolute terms that *Y* causally depends on *X or* that *X* precedes *Y* and is causally joined with *Y*…

"Causally joined" refers, roughly, to cases in which a dependency holds between two variables *X* and *Y* in virtue of a shared causal ancestor, but neither *X* nor *Y* causes the other.

I am inclined to say that the (non-universal) frame-relative notion (and the arrow that Spohn uses to represent this) is not a notion of causal dependence at all (at least as I would understand the notion) but (at least in a probabilistic setting) something more like a notion of probabilistic relevance that does not disappear when one conditions on certain other variables in one's frame—in other words, it is a variety of conditional dependence. Thus in (ii) the arrow seems to mean simply that *X* and *Y* are statistically dependent or positively correlated (and that *X* precedes *Y*), where this may be because *X* causes *Y* or because *X* and *Y* are joint effects of a common cause, as in (i). The arrow from *X* to *Y* in (ii) disappears in (i) because *X* and *Y* are not dependent conditional on *Z*.

Now it is uncontroversial that probabilistic dependence and independence relations, conditional and unconditional, are always "relative" to a variable set. *X* and *Y* can be unconditionally dependent, independent conditional on a third variable *Z*, dependent again conditional on both *Z* and a fourth variable *W* and so on. But many think that causal relationships are different from relationships of conditional dependence and that we should not expect the former to be variable or frame-relative in quite the same way as the latter.

As an alternative, consider how one might think about examples (i) and (ii) within an interventionist framework. Here the basic idea is that *X* causes *Y* in background circumstances

---

[6] This is the assumption that faithfulness holds for each graph in the process of refining frames on the way to the universal frame. (*LB*, 397)

B iff there is some intervention on *X* that is associated with a change in the value of *Y*. An intervention *I* is an idealized unconfounded experimental manipulation—one way of realizing an intervention is by means of a randomized experiment. (For details, see Woodward 2003). I emphasize that whether *I* counts as intervention is *not* a variable-relative notion—see below.

Suppose one is faced with a situation in which what one knows is represented just by (ii) in Spohn's example—*X* and *Y* are dependent and one does not know about *Z*. An intervention on *X* will nonetheless involve a manipulation of *X* that is independent of *Z*—an intervention on *X* puts *X* entirely under the control of the intervention and manipulates *X* in a way that is uncorrelated with *Z*—this is part of the definition of an intervention and also what is achieved by a properly conducted randomized experiment. One of the virtues of such an experiment is that one does not have to know about the existence of *Z* or how it is related to *X* and *Y* to carry out a random assignment of values to *X*. Roughly all that one has to know is that the randomizing device is causally independent of any factors such as *Z* that might act as confounders for the *X*—*Y* relationship. Suppose that under all such interventions on *X*, *Y* does not change[4]. Then we may conclude, within an interventionist framework, that *X* is not, in one sense, a cause of *Y* (*X* is not what Woodward calls a total cause of *Y*), and, assuming faithfulness, one can also conclude that *X* is also not a direct or contributing cause of *Y*. If an arrow from *X* to *Y* is interpreted to mean that *X* directly causes *Y*, (ii) is, on an interventionist account of direct cause, simply a mistaken representation of the causal facts. Further, if it is also the case that *Y* does not change under any interventions on *X*, and that *X* changes under interventions on *Z*, with *Y* fixed and similarly for *X* and *Y* interchanged, then we may conclude that (i) is the correct structure. Note that, within the interventionist framework, these

[4] Of course this requires that the randomization "works" in the sense of making *Z* independent of *X*. This assumption may be false but one can think of it as involving an ordinary inductive risk.

conclusions are not understood as variable-relative claims—we do not say that (ii) is correct relative to the variables *X, Y*, and (i) correct relative to the variables *X, Y, Z*.

In correspondence, Spohn has suggested that this characterization of the frame-relative features of his view (and the implied opposition between his view and interventionism) is misleading. On his view, it is true in one sense that in order to ascertain with full belief or confidence whether some causal claim is true, we would have to carry out the full extension to the universal frame. However, viewed from another perspective, this just amounts to an expression of fallibilism about the claim in question—an acknowledgement of the logical possibility that the belief might be mistaken. In other words, it is only in the sense that we have to acknowledge the possibility that any of our empirical a posteriori beliefs may turn out to be mistaken that we must also acknowledge the possibility that our causal beliefs about the relation between *X* and *Y* may flip back and forth between dependence and independence as we add more information, only stabilizing when (if ever) we reach the universal frame.

Let me conclude by suggesting a somewhat different perspective on all this. It seems clear, as a matter of empirical fact, that in many cases our beliefs about causal relationships do not exhibit the kind of radical instability (flipping back and forth from dependence to independence) countenanced by Spohn as an in-principle possibility. Putting matters in terms of Spohn's notion of a universal frame, it looks as though we can sometimes reliably figure out from the information we presently have that further information about the (presently unknown) details of the universal frame is unlikely to change our present causal beliefs. This in turn suggests further questions: What features of the procedures by which we form our causal beliefs (and what features of the natural world) sometimes make it possible for us to reach conclusions that are stable in this way? My suggestion is that randomization (and arguably experimentation itself) can be thought of along these lines: thinking in terms of the universal frame, the results of the randomization tell us something that must be true in that frame (that if *Y* is associated with randomized *X*, there is no confounding *Z* in the universal

frame) but not by telling us anything in detail about what is in it. There are other strategies with similar virtues.

References

Halpern, Jospeh. 2016. *Actual Causality*. Cambridge: MIT Press.

Janzing, Dominik, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniusis, Bastian Steudel, and Bernhard Schölkopf. 2012."Information-geometric Approach to Inferring Causal Directions." *Artificial Intelligence* 182-183:1-31.

Paul, Laurie A., and Ned Hall. 2013. *Causation: A User's Guide.* Oxford: Oxford University Press.

Spirtes, Peter, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search*, Cambridge, Mass.: MIT Press.

Spohn, Wolfgang. 2012. *Laws of Belief: Ranking Theory and its Philosophical Applications.* Oxford: Oxford University Press.

Woodward, James. 2003. *Making Things Happen*. New York: Oxford University Press.

Reply

to Jim Woodward's Comments on

Wolfgang Spohn's *Laws of Belief*

Wolfgang Spohn

Department of Philosophy, University of Konstanz

Abstract: The discussion note sees the core difference of the accounts of causation under discussion in the fact that Woodward's account follows an epistemological order, while Spohn's account follows a purely conceptual order. This unfolds in five further differences: (i) starting with type- or with token-level causation; (ii) avoiding or allowing reference to time; (iii) actual/counterfactual intervention vs. epistemic/suppositional wiggling; (iv) circularly conceiving the circumstances of a direct causal relation as the other direct causes or as the entire past of the effect (without the cause) in a circle-free way; and (v) grasping absolute causation directly vs. first grasping model-relative causation.

Woodward started writing on causation and explanation in 1979 (Woodward 1979), culminating in, and by far not ending with, his formidable book *Making Things Happen* (2003). I started writing on causation in 1978 (Spohn 1978, ch. 3), so far culminating in chapters 14 – 15 or 130 pages in Spohn (2012). Our accounts of causation look similar; both seem to be variants of causal Bayes net theorizing (which was anticipated in Spohn (1978, 1980). Both look dissimilar (by Woodward's reliance on structural equations and my reliance on ranking theory), in ways that are of subordinate importance. And both *are* dissimilar in important ways that may be less obvious. Therefore, it may be useful to give an easily accessible description of those differences, which is perhaps an expedient companion to those many pages. Here, I am giving this description from my point of view, in the course of which I shall also respond to various comments by Woodward in his previous discussion note.

The obvious difference is that Woodward states his account in terms of structural equations, or he starts illustrating it with them from the beginning in Woodward (2003, sect. 2.2). The equations generate or are represented by a causal graph, and they are richly amended by statistical methodology, thus resulting in causal Bayes nets. In this way his account is well aligned with current scientific methodology. By contrast, I present my account

in terms of my idiosyncratic ranking theory. This has a reason, of course, which is dear to me. I came to think of ranking theory in Spohn (1983) precisely with the aim of developing an account of deterministic causation—the kind of causation thinkers were thinking about for millennia—in perfect parallel to accounts of probabilistic causation, which at that time seemed superior and more sophisticated. They did so because they could use more adequate notions of (conditional) independence. Ranking theory succeeds in doing so, while the structural equations approach still does not in my view.

The core point here is that the independence notion provided by structural equations is basically logical or functional dependence. Even with counterfactual dependence one does not get anything like the graphoid axioms (cf. Spohn 1978, sect. 3.2, or Pearl 1988, pp. 84ff.), which hold for conditional probabilistic (or ranking) independence. An important symptom of the relevant difficulties is the treatment of symmetric causal overdetermination, which is not satisfactorily dealt with by counterfactual theories and which has provoked at least four different versions of the ominous condition AC2(b) of the structural equations account of actual causation. The most recent one is in Halpern (2016, 23ff.), which differs from his previous ones and also from AC*2 in Woodward (2003, 84). (For a more extensive comparison of the structural equations and the ranking-theoretic approach to causation see Spohn (2010).)

This difference between the structural equations and the ranking-theoretic approach is neglected by Woodward's discussion. Rightly so; it is, in a way, superficial. I have emphasized that ranking theory may be replaced throughout by probability theory and that all definitions and theorems should and can be preserved, thus resulting in my proposal for a probabilistic theory of causation. Likewise, structural equations can be probabilistically amended. And then the difference is largely gone, opening the view on the more basic differences.

There is also a salient difference in style. I recall when first reading Woodward's book that I almost felt like reading ordinary language philosophy. That's praise. I was raised with this kind of philosophy and love it. The book is a most considerate investigation into the conception of causation of laymen and scientists. By contrast, my chapters are largely a formal exercise in rigorous theorizing (and thus much less pleasant to read). Again, one might say that this is a superficial difference while we are united in conceptual analysis.

However, this already moves me to what I take to be our core difference. Woodward's conceptual analysis is very *close to scientific practice* (this is why it is so instructive for scientists of various sorts), while mine is very much *more geometrico*. Or in more substantial terms: I think that Woodward's investigation in effect rather follows an *epistemological order*. At least his discussion note comments on my account of causation distinctly from an epistemological point of view. By contrast, I try to keep to a *purely conceptual order*, without regard to its epistemological consequences. Both orders are legitimate, important, and informative. So I have nothing to criticize about Woodward's order. But they entail quite a number of further differences. In the sequel I would like to explain five such differences:

(1) A first conspicuous difference is that Woodward deals with *type-level* (I say, generic) causation while I treat *token-level* (singular) causation. The reason is simple. The natural and the social sciences are interested in laws or invariances, and if they have causal form, they concern type-level causation. Woodward addresses those sciences. (One should never forget, though, that there are large fields, law and history, for instance, that are primarily concerned with token-level causation.) By contrast, my entry to causal theorizing was causal decision theory, and this is always about single decision situations and hence about token-level causation. Woodward discusses this contrast in sections 2 and 3 of his comments.

The point now is that Woodward is thereby oriented at the epistemological order. At least in the sciences we first try to find out about the causal regularities (even if only statistical ones) and then apply them to the single cases. This agrees with his strategy to explicate actual

causation only in a second step, namely relative to given structural equations or causal laws (see Woodward 2003, 74ff.).

Woodward cites the example that it is statistically overwhelmingly confirmed that smoking is a contributing type-level cause of lung cancer, while it may be very difficult to confirm that Jones' smoking for 30 years caused his lung cancer. Well, that's how the tobacco industry argued for a long time: there is no conclusive evidence in the single case. I would think, if the type-level claim is well confirmed, the token-level claim about Jones is *prima facie* equally well confirmed. Still, the example supports Woodward's case.

However, this cannot be the conceptual order. What else could causal regularities be but generalizations (or perhaps averages) of singular causal claims? In my view, the strict conceptual order can only proceed to first understand the latter in order to understand the former. This is the order I attempt to pursue.

The point is emphasized by the fact that type-level causation is always about causal dependence between (type-level) variables, while token-level causation is about causation between events or singular facts. What is usually overlooked is that there also are token-level variables and causal dependencies between them. In any case, it seems clear to me that causal dependence between (token- or type-level) variables is causation between some realizations of those variables and thus derivative upon causation between facts or events.

This difference is one that I do not only have with Woodward. Since 1978 I spoke about token-level causation and was hence at cross-purposes with the mainstream on probabilistic causation like, e.g., Cartwright (1979). It took me a long time to realize how crucial this difference is. Cartwright acknowledged it only in Cartwright (1989, 9), where she also came to grant priority to token-level causation.

I confess that I have no more to offer to type-level theorists like Woodward than the crude suggestion that statements about causal relations between type-level variables are generalizations or averages of causal relations between the corresponding token-level

variables. This may well be too crude. The type-level theorist has the reverse problem, though. It is quite unclear what his statements tell about the single case. The problem is analogous to the old issue about the relation between statistical probabilities and single-case propensities, which seems as dim as ever.

(2) This point directly entails the next crucial difference (also discussed by Woodward in section 4 of his comments). Since I start with causation between singular facts, I can assume that those facts are temporally located and that causes always precede their effects. (There are difficulties with simultaneous causation, and I take backwards causation to be a *contradictio in adjecto*.) Of course, this *reference to time* considerably facilitates theorizing. For instance, there is only one causal graph respecting the temporal relations between token-level variables and the conditional probabilistic dependencies between them, while the latter alone, without reference to time, are often compatible with many causal graphs.

The problem is, as Woodward rightly states, that on the type level of causal theorizing temporal relations are rarely specifiable. My parents' income precedes my educational status, yet there is no temporal and only a causal relation between parental income and educational status in general. Statistical data reveal a lot of conditional statistical dependencies between type-level variables, and the causal theorist can proceed only from them. However, that's the dictate of the epistemological, not of the conceptual order.

Woodward is right in pointing to the extent to which causal theorizing gets along without reference to time. This is indeed remarkable and shows how far one can get in the epistemological order. From the purely conceptual point of view, however, I don't see why I should burden me with such restrictions. Conceptually, the relation between time and causation has been fundamental throughout history (even though it is contested how exactly to state it).

(3) The story continues. The notion of *intervention* is at the center of Woodward's account of causation (see Woodward 2003, ch. 3, and section 2 of his comments). In the

epistemological order this is perfectly justified. *Actual* intervention produces by far the best and most direct confirmation of causal relations. If I wiggle the variable *X* at will, and the variable *Y* wiggles accordingly, then *Y* obviously causally depends on *X*. (The details are extensively elaborated by Woodward.) However, as he is well aware, actual intervention is rare. Therefore he tries to generalize this epistemological virtue by turning to *counterfactual* interventions, which can be applied to each causal relationship. (Counterfactually, I can also shift the moon and see what happens to the tides.)

I perfectly agree with this move. However, I feel that the advocates of this move— who are many, not only Woodward—grossly underrate its size and risk. In my view, the very rich and confusing literature on conditionals and counterfactuals only allows the conclusion that their truth-evaluability is very dubious or shallow. For instance, do we have a robust realistic notion of a similarity ordering between worlds as we have of tables or electrons? Surely we have the peremptory intuition that at least some counterfactuals are true, and Woodward (2003, 118ff.) makes as strong a case as possible that the narrow, though not well-delineated class of interventionist (non-backtracking) counterfactuals belongs to them. Indeed, if we could start from true causal laws (or structural equations), the counterfactuals derived from them inherit their truth-evaluability. However, in my conceptual order this start is illegitimate, and the doubt about their truth-evaluability reversely spreads to causal laws.

So, if we want to have a general account of the conditional and counterfactual idiom, we must not presuppose its truth-evaluablity and then explain it in some shallow way or find excuses why it often does not work properly. Rather, we must try to find some other analytical starting point—say: expressivism—and then try to explain how some conditionals can emancipate from that starting point and acquire proper truth conditions. At least, that's the strategy I propose in Spohn (2015).

What does this entail for causation? My basic explication is that *A* is a *direct cause* of *B* iff *A* and *B* are singular facts, *A* precedes *B*, and *A* is positively relevant to *B* given the

circumstances. (And then I go on to define causation as the transitive closure of direct causation and defend this step against recent criticism; see Spohn (2012, sect. 14.11 – 13). I shall not substantiate this issue here.) The crucial point is that the only workable notion of positive relevance that I can find is an epistemic one, either in terms of subjective probabilities or of ranks (or possibly of other degrees of belief). Thus my notion of direct causation is thoroughly epistemically relativized. This is our Humean heritage, which we must accept and not reject in my view.

Thus, in my view, we have no choice but starting from this epistemic relativization, I share the intention to go beyond it. Somehow, we have to turn what is relative to our minds into an objective feature of the world. In other words, I adhere to 'projectivism', which Woodward (2003, 118ff.) finds confused and misguided. Indeed, the projectivist is always in danger to talk in perverted ways. This is due to the causal character of the projectivistic metaphor: without projector nothing projected; so without the human mind, no causation in the world. That's absurd and rightly attacked by Woodward. But it is a misunderstanding of projectivism, which does not make any counterfactual claims. How else, though, is the metaphor to be understood? That's an inveterate difficulty. I try to solve it in Spohn (2012, ch. 15) by what I describe as the 'objectivization' of the epistemically relativized causal relations, which is, as far as I see, not subject to Woodward's criticism of projectivism. For me, this is still the most important issue about causation and one at which Woodward and I are clearly at odds. However, it is too large to be further pursued here.

The only point I want to make is that once we move from actual to counterfactual wiggling, we have in effect moved to epistemic or suppositional wiggling, as embodied in my explication. Positive relevance of A to B, epistemically interpreted, says nothing but that variation of degree of belief in A covaries with variation of degree of belief in B (given the circumstances). Hence, the counterfactual theorist is in fact perilously close, in my view, to

the Humean heritage. We have here another point of similarity between Woodward and me paired with disagreement underneath.

(4) This brings me to the next issue. Whatever the relevant kind of wiggling, we must keep the *circumstances* fixed. What does this mean? I argue that the circumstances of *A*'s being a direct cause of *B* consist of the entire history of *B* without *A*. Conceptually this is crucial, since it frees the above explication of conceptual circles; circumstances are thereby explained only with reference to time. But this is epistemological disaster, as Woodward rightly remarks. It seems then that we can never affirm causal relations due to that reference to the entire history. Another facet of my basic theme.

I can offer some consolation. Given my explication and some mild auxiliary assumptions, the circumstances of *A*'s being a direct cause of *B* can be reduced to consist of all the other direct causes of *B*. This was already suggested by Cartwright (1979), and she inferred, wrongly in my view, that we are thus entangled in an inextricable conceptual circle. The reducibility means that the rich (circle-free) and the reduced (circularly defined) circumstances provably yield the same causal relations. So, it suffices to keep fixed only those other causes. This is what we try to do in actual experimentation. Our painful experience, though, is that we were often wrong with our guesses as to those other causes. Conceptually, we are guaranteed to be on the safe side only with the rich circumstances. Epistemologically, however, we do, and have to, proceed without any such guarantees.

(In temporally ordered Bayes nets the same holds. For temporally located token-level variables X and *Y* we can define *X* to be a parent of *Y* iff *Y* probabilistically depends on *X* given all the other variables in the past of *Y*. And then we can prove the causal Markov condition (without auxiliary assumptions), i.e., that *Y* is independent from its past and indeed from all its non-descendants given all its parents.)

(5) I can offer another consolation (which Woodward discusses in section 5 of his comments). We shall see, though, that it is a Greek gift. When I said that the circumstances

consist of the entire past of the direct effect without the direct cause, I didn't literally mean the entire past, but only the entire past insofar as it is represented within the causal model or its set $\boldsymbol{V}$ of variables. This is something in our grip. Clearly, though, it makes my notion of direct causation model-relative. Woodward (2003, 55) makes a similar move when saying that Y directly causally depends on X iff an intervention on X would make a difference for Y provided all other variables in $\boldsymbol{V}$ are kept fixed. (Since he cannot refer to a temporal order in $\boldsymbol{V}$, he must keep fixed all other variables in $\boldsymbol{V}$, not only those in the past of Y.) He thus accepts the model-relativity of direct causation, i.e., of the distinction between direct and indirect causation. (He does not do so, though, for causation *simpliciter*; see below.)

That's no surprise. Of course, the direct/indirect distinction is model-relative. A larger model may spell out the steps mediating what appears to be a direct causal relationship within a smaller model. However, the model-relativity implied by my explication runs deeper (that's the Greek gift). An apparent causal relation may turn out spurious in the larger model or it may show up only in the larger and not yet in the smaller model (where the latter can occur only when faithfulness is violated). In other words: Simpson's paradox cannot raise its head within a given causal model, because no variable in the model is left for further conditioning which could reverse the probabilistic dependencies. But by looking outside the model or enlarging it, we may fully run into the paradox again.

This raises Woodward's suspicion. It appears that I am not talking about causation, after all, but only about conditional dependence. (As far as I know, so-called Granger causality has fallen into disgrace in economics, precisely for the same reason; it is said to provide only a method of forecasting rather than of causal inference. As such it is still appreciated. I find this assessment unjustified, as I am going to explain.)

So, all the conditional dependencies or (partial) statistical correlations within the model cannot prove real causation. This is true. This entails that causation is a *model-transcendent* notion. This insight can hardly be overemphasized. Of course, causation can be represented in

models. But it cannot be ascertained within them; it always refers to things not represented in them. So, obviously my model-relative notion of causation does not capture the model-transcendent or absolute notion, which all causal theorists including Woodward intend. In short, I am missing the topic.

This criticism is well taken. However, this model-transcendence presents a substantial problem, which I would like to discuss in the rest of this note. How do others try to fulfill their intention? Let me look at three attempts:

The original causal Bayes net theory of Spirtes et al. (1993) may appear model-relative as well. However, they do not really say what causation is. In the terms of Glymour (2004), they prefer the 'Euclidean' method of specifying only axioms (causal Markov, minimality, and possibly faithfulness) that a causal model has to satisfy. Those axioms, though, hold only under an important assumption: that the model is *causally sufficient* in the sense that every (direct) common cause of two variables in the model is in the model as well (the exact definition is slightly more complicated; see Spirtes et al. 1993, 45). That's a very strong assumption, and it is clearly model-transcendent. How ever to ascertain it? The Emperor of China may exert all kinds of hidden influences. Who knows? Maybe he is a common cause of two variables in the model? Maybe God is invariably so, as occasionalism originally intended? Of course, Spirtes et al. (1993, 44) are right when saying that we are justified in disregarding such weird hypotheses. Scientists have great, but often failing skills in selecting reasonable models. Still, this assumption is the entry of the unavoidable model-transcendence into their theory.

Spirtes et al. were aware of the strength of that assumption and were not happy with it. Substantial parts of their subsequent theorizing are occupied with how much of causal inference about hidden, latent or unmeasured variables is still possible when the assumption is weakened. This is not the place to go into this. It is clear that these are most interesting ways to get a hold on the model-transcendence, which, however, are bound to be only partial.

Another example is Woodward's notion of an intervention, which is model-transcendent as well. For him, as mentioned, Y directly causally depends on X iff there is an intervention I on X which changes (the probabilities of) Y while keeping the other variables in the model set $\boldsymbol{V}$ fixed. But such an intervention variable I must itself satisfy various, partially causal conditions. This circular characterization of causation and intervention has been criticized, e.g., by Glymour (2004), but is defended by Woodward as something to be expected. This is not my issue. My point is that one of those conditions requires that the intervention variable I is statistically independent of any variable that causes Y along some causal path that does not go through X (this is condition I4 in Woodward 2003, 98). Here, "any variable" must be taken as quantifying also about variables outside $\boldsymbol{V}$. Again, we have a reference to the don't-know-what's outside the causal model. But, of course, this condition is necessary. If there were such a statistical correlation or dependence, the intervention on X would not show X's influence on Y.

If I intervene at will, how could there be such a correlation? This rhetorical question points to an action theoretic justification of Woodward's requirement. Interventions are actions that fall under the competence of decision theory, indeed causal decision theory (CDT), as most philosophers think including me. A basic tenet of CDT (see Spohn 1978, 109f.) is that acts are exogenous, cut off from any causal ancestry. The intervention variable I is really *do* I, as conceived in the *do*-calculus of Pearl (2000, 85ff.) and as emphasized by Woodward (2003, 47) as well. Since the only effect of the intervention I is on the variable X, each correlation I may have with Y must therefore run through X. Thus, Woodward's requirement is automatically fulfilled in the picture of CDT.

However, CDT represents the internal picture of the agent or intervener. She must think that her actions are cut off and uncorrelated in this way. But the external observer may find hidden correlations. Otherwise we could leave it, e.g., to the physician to assign patients to the test group and the control group at will. However, the wisdom of randomized controlled trials

(RCTs) is precisely not to rely on the physician's will because of those possible hidden correlations. And since Woodward takes the observer's external view, he is right in insisting on that requirement.

This brings me to the third way of accommodating the model-transcendence of causation, which is offered also by Woodward in response to my worries. The whole point of the experimental methodology of RCTs is precisely to average out all those model-transcendent factors and thus to deliver reliable average type-level causal hypotheses. Isn't this good enough?

Well, RCTs require a lot of care. The randomization must be carried out properly, one must check for all selection variables potentially built in into the experiment, etc. Good experiments attend to all these things. Given this, there is no doubt that RCTs belong to the most advanced epistemological strategies we have in the social sciences, epidemiology, etc.

However, I don't think that they can fully control what I call the model-transcendence of causation. RCTs can properly randomize over the given population. If well done, they optimize internal validity of the experiment. However, as is well known, this is no guarantee of external validity, of the generalizability of the experimental findings to other populations. Still worse, is the given population, even if we take it to be the entire present mankind, representative for the behavior of all the variables outside the model? It is not even clear what this representativity could mean when the joint distributions of the variables change all the time. In practice we have nothing better to go for. In theory, though, such an assumption of representativity would be a big and presumably not well-founded step. Again we find the opposition between the epistemological opportunities and the conceptual requirements.

Does the original counterfactual analysis of causation, as introduced by Lewis (1973) and refined later on, not tackle this model-transcendence straightaway? It may seem so. This analysis does not mention models at all and immediately refers to possible (grand) worlds, which Lewis takes in the most comprehensive, not transcendible sense. However, the

downside of this procedure is its startling theoretical poverty. Surprisingly, this poverty seems of no concern within the tradition of the counterfactual analysis. How many theorems do we find there? What a contrast to the rich development in Spirtes et al. (1993), Pearl (2000), and Woodward (2003)! Woodward completely agrees with this criticism. The reason is clear: as soon as we want to get to details, if only in a theoretical way, we have to describe the possible worlds, with predicate and individual terms or with variables and their realizations, etc.; and thereby we capture only small worlds, i.e., models, and nothing more. This is an unavoidable trade-off.

My conclusion from all this is this: We grasp causation within models, we intend to grasp absolute causation, and we have no clear idea about the relation between the two things and only very partial ways of grasping all the don't-know-what's not included in our models. My further conclusion is that we should reverse the priorities. We should not stare at absolute causation, deplore the insufficiencies of model-relative causation, and try to directly repair them. We should rather be happy to start with model-relative causation and then try to thereby approach absolute causation. This is the reversed strategy I recommend.

So, I propose to start with the above model-relative explication of direct and indirect causation. This allows rigorous and instructive theorizing about the model-relative notion. For instance, as indicated above, the model-relative versions of the causal Markov and the minimality condition simply turn out to be provable. Thus we catch up with Spirtes et al. 1993). As Woodward notes, this is not absolute causation. Can we somehow approach it from this starting point?

I think so. Not directly, though. However, we can study how model-relative causation in smaller and larger models relates. That is, it may be hard to say how model-relative causation in a small model unfolds in a larger model (though not everything goes). We can precisely study, however, how model-relative causation in a larger model appears in a smaller model. I have only theorems about causal dependence between variables and none about causation

between facts. Roughly, one of my theorems (see Spohn 2012, 395ff.) is that, given faithfulness, causal dependence of Y on X in a smaller model means that in the larger model Y causally depends on X or shares a common causal predecessor with X. (Faithfulness is not a necessary condition, but I do not know of any weaker sufficient condition.) In a way, this is not surprising. It is very similar to Reichenbach's common cause principle, which is intuitively very convincing (and turns out provable as well, taken in a model-relative way; see Spohn 2012, 384ff.).

The idea now is that causation in the larger model is a substitute for model-transcendent causation in the smaller model. Of course, it is still model-relative causation. But we may conceive of absolute causation as causation relative to the universal model containing all variables whatsoever. Admittedly, this universal model is absolutely fictitious and ill defined, just as Lewis' grand worlds, which we may postulate, but can describe only very partially. Still, we may take the way in which causation relates within smaller and larger models as paradigmatic or indicative of the relation between model-relative and absolute causation and hence state, e.g., that model-relative causal dependence is *either* causal dependence *or* having a common causal predecessor in the absolute sense (see Assertion 14.42 in Spohn 2012, 397).

This move makes none of our substantial epistemological problems with (the intended) absolute causation vanish; it does not even mitigate them. We always move within limited models, we always try to extend them to larger, but still limited models, and we always face the issue what this might tell us about full reality or absolute causation. It is this issue that has been tackled in the work mentioned above. And my hunch is that it can be stated in an equivalent way within my proposal. (However, this should be checked. Maybe my hunch is too optimistic; this would throw doubt on my proposal. Or maybe my framework allows improving on that issue in some ways.)

Still, my proposal contains a determinate conceptual strategy: first to give a rigorous account of model-relative causation and then to approach absolute causation in those model-

relative terms. This is clearer and more consequent than trying to account for absolute causation through amending model-inherent theorizing by inexplicable model-transcendent conditions. In other words: it reestablishes priority of the Socratic over the Euclidean method.

I said in the beginning that both, the epistemological and the conceptual order, are legitimate and important. They are indeed. Implicitly, though, my discussion note was a pleading for the priority of the conceptual over the epistemological order. We should strictly pursue the conceptual order first and not compromise it with the epistemological order, as I interpret Woodward as doing. Only then we can know what our epistemological problems are and attempt to solve them.

References

Cartwright, Nancy. 1979. "Causal Laws and Effective Strategies." *Noûs* 13:419-437.

Cartwright, Nancy. 1989. *Nature's Capacities and Their Measurement*. Oxford: Clarendon Press.

Glymour, Clark. 2004. "Critical Notice on: James Woodward, Making Things Happen." *British Journal for the Philosophy of Science* 55:779-790.

Halpern, Joseph Y. 2016. *Actual Causation*. Cambridge, Mass.: MIT Press.

Lewis, David. 1973. "Causation." *Journal of Philosophy* 70:556-567.

Pearl, Judea.1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann.

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

Spirtes, Peter, Clark Glymour, and Richard Scheines. 1993. *Causation, Prediction, and Search*. Berlin: Springer. 2nd ed. 2000, Cambridge, Mass.: MIT Press.

Spohn, Wolfgang. 1978. *Grundlagen der Entscheidungstheorie*, Kronberg/Ts.: Scriptor. Out of print, pdf-version at: https://www.philosophie.uni-konstanz.de/ag-spohn/personen/prof-dr-wolfgang-spohn/books-lecture-notes/

Spohn, Wolfgang. 1980. "Stochastic Independence, Causal Independence, and Shieldability." *Journal of Philosophical Logic* 9:73-99.

Spohn, Wolfgang. 1983. *Eine Theorie der Kausalität*, unpublished Habilitationsschrift, Universität München, pdf-version at: https://www.philosophie.uni-konstanz.de/ag-spohn/personen/prof-dr-wolfgang-spohn/books-lecture-notes/

Spohn, Wolfgang. 2010. "The Structural Model and the Ranking Theoretic Approach to Causation: A Comparison." In *Heuristics, Probability and Causality. A Tribute to Judea Pearl*, ed. Rina Dechter, Hector Geffner, and Joseph Y. Halpern, 493-508. San Mateo, CA: Kauffmann.

Spohn, Wolfgang.2012. *The Laws of Belief. Ranking Theory and Its Philosophical Applications*. Oxford: Oxford University Press.

Spohn, Wolfgang. 2015. "Conditionals: A Unified Ranking-Theoretic Perspective." *Philosophers' Imprints* 15, No. 1:1-30.

Woodward, James. 1979. "Scientific Explanation." *British Journal for the Philosophy of Science* 30:41-67.

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