Stanford Encyclopedia of Philosophy

Causation and Manipulability

First published Fri Aug 17, 2001; substantive revision Sat May 6, 2023

Manipulability theories of causation, according to which causes are to be regarded as handles or devices for manipulating effects, have considerable intuitive appeal and are popular among many scientists and statisticians. This article surveys several prominent versions of such theories advocated by philosophers, and the many difficulties they face. Until recently, philosophical statements of the manipulationist approach have generally been reductionist in aspiration and have assigned a central role to human action. These contrast with more recent discussions employing a broadly manipulationist framework for understanding causation, such as those due to the computer scientist Judea Pearl and others, which are non-reductionist and rely instead on the notion of an intervention. This is simply an appropriately exogenous causal process; it has no essential connection with human action. This interventionist framework manages to avoid at least some of these difficulties faced by traditional philosophical versions of the manipulability theory and helps to clarify the content of causal claims.

1. Introduction

A commonsensical idea about causation is that causal relationships are relationships that are potentially exploitable for purposes of manipulation and control: very roughly, if C is genuinely a cause of E, then if I can manipulate C in the right way, this should be a way of manipulating or changing E. This idea is the cornerstone of manipulability theories of causation developed by philosophers such as Gasking (1955), Collingwood (1940), von Wright (1971), Menzies and Price (1993), and Woodward (2003). It is also an idea that is advocated by many non-philosophers. For example, in their extremely influential text on experimental design (1979) Cook and Campbell write:

> *The paradigmatic assertion in causal relationships is that manipulation of a cause will result in the manipulation of an effect*. … Causation implies that by varying one factor I can make another vary. (Cook & Campbell 1979: 36, emphasis in original)

Similar ideas are commonplace in econometrics and in the so-called structural equations or causal modeling literature, and more recently have been forcefully reiterated by the computer scientist Judea Pearl in very influential book length treatment of causality (Pearl 2009).

At least until relatively recently philosophical discussion has been unsympathetic to manipulability theories: two standard complaints have been that manipulability theories are unilluminatingly circular and that they lead to a conception of causation that is unacceptably anthropocentric or at least insufficiently general in the sense that it is linked much too closely to the practical possibility of human manipulation (see, e.g., Hausman 1986, 1998). Both objections seem *prima facie* plausible. Suppose that X is a variable that takes one of two different values, 0 and 1, depending on whether some event of interest occurs. Then for an event or process M to qualify as a manipulation of X, it would appear that there must be a causal connection between M and X: to manipulate X, one must *cause* it to change in value. How then can we use the notion of manipulation to provide a non-circular account of causation? Moreover, it is uncontroversial that causal relationships can obtain in circumstances in which manipulation of the cause by human beings is not practically possible—think of the causal relationship between the gravitational attraction of the moon and the tides or causal relationships in the very early universe. How can a manipulability theory avoid generating a notion of causation that is so closely tied to what humans can do that it is inapplicable to such cases?

These philosophical criticisms of manipulability theories contrasts with the widespread view among statisticians, theorists of experimental design, and many social and natural scientists that an appreciation of the connection between causation and manipulation can play an important role in clarifying the meaning of causal claims and understanding their distinctive features. This in turn generates a puzzle. Are non-philosophers simply mistaken in thinking that focusing on the connection between causation and manipulation can tell us something valuable about causation? Does the widespread invocation of something like a manipulability conception among practicing scientists show that the usual philosophical criticisms of manipulability theories of causation are misguided?

The ensuing discussion is organized as follows. Section 2 describes a well-known version of an manipulability theory due to Menzies and Price (1993) which assigns a central role to the notion of agency or free action. Section 3 describes reasons why the notion of a free action seems an inadequate basis for the formulation of a manipulability theory. Section 4 introduces the notion of an intervention which allows for a more adequate statement of the manipulability approach to causation and which has figured prominently in recent discussion. Section 5 considers Pearl's "interventionist" formulation of a manipulability theory and an alternative to it, due to Woodward (2003). Section 6 takes up the charge that manipulability theories are circular. Section 7 explores how interventionist ideas can be used to explicate a variety of different causal concepts. Section 8 returns to the relationship between interventions and human actions, while §9 discusses the role of counterfactuals in interventionist theories. Sections 10, 11 and 12 consider the scope of manipulability accounts, while §13 considers some objections to such accounts and §14 some recent positive developments.

As we shall see, the somewhat different assessments of manipulability accounts of causation within and outside of philosophy derive in part from the different goals or aspirations that underlie the versions of the theory developed by these two groups. Early philosophical defenders of the manipulability conception such as von Wright and Menzies and Price attempted to turn the connection between causation and manipulability into a reductive analysis: their strategy was to take as primitive the notion of manipulation (or some related notion like agency or bringing about an outcome as a result of a free action), to argue that this notion is not itself causal (or at least does not presuppose all of the features of causality the investigator is trying to analyze), and to then attempt to use this notion to construct a non-circular reductive definition of what it is for a relationship to be causal. Philosophical critics have (quite reasonably) assessed such approaches in terms of this aspiration (i.e., they have tended to think that manipulability accounts are of interest only insofar as they lead to a non-circular analysis of causal claims) and have found the claim of a successful reduction unconvincing. By contrast, statisticians and other non-philosophers who have explored the link between causation and manipulation generally have not had reductionist aspirations—instead their interest has been in unpacking what causal claims mean and in showing how they figure in inference by tracing their interconnections with other related concepts (including manipulation but also probability) but without suggesting that the notion of manipulation is itself a causally innocent notion.

The impulse toward reduction contributes to the other features that critics have found objectionable in standard formulations of the manipulability theory. To carry through the reduction, one needs to show that the notion of agency is independent of or prior to the notion of causality and this in turn apparently requires that human actions or manipulations be given a special status—they can't be ordinary causal transactions, but must instead be an independent fundamental feature of the world in their own right. This both seems problematic on its own terms (it is *prima facie* inconsistent with various naturalizing programs) and leads directly to the problem of anthropocentricity: if the only way in which we understand causation is by means of our prior grasp of an independent notion of agency, then it is hard to see what could justify us in extending the notion of causation to circumstances in which manipulation by human beings is not possible and the relevant experience of agency unavailable. Both von Wright and Menzies and Price struggle with this difficulty.

One way out of these problems is to follow Pearl and others in reformulating the manipulability approach in terms of the notion of an intervention, where this is characterized in purely causal terms that make no essential reference to human action. Some human actions will qualify as interventions but they will do so in virtue of their causal characteristics, not because they are free or carried out by humans. This "interventionist" reformulation allows the manipulability theory to avoid a number of counterexamples to more traditional versions of the theory. Moreover, when so reformulated, it is arguable that the theory may be extended readily to capture causal claims in contexts in which human manipulation is impossible. However, the price of such a reformulation is that we lose the possibility of a reduction of causal claims to claims that are non-causal. Fortunately (or so §§7 and 8 argue) an interventionist formulation of a manipulability theory may be non-trivial and illuminating even if non-reductive.

2. Agency Theories.

A comparatively early and influential statement of a manipulability theory which assigns a central role to human agency is due to von Wright (1971; see An Early Version of an Agency Theory for further discussion). However, this entry will focus on the more recent version developed by Peter Menzies and Huw Price (1993) (also discussed in a series of papers written by Price alone [1991, 1992, and more recently, 2017 which argues that the contrast between agency versions of a manipulability theory and non-agency versions is not as sharp as is suggested below) . Menzies' and Price's basic thesis is that:

> … an event A is a cause of a distinct event B just in case bringing about the occurrence of A would be an effective means by which a free agent could bring about the occurrence of B. (1993: 187)

They take this connection between free agency and causation to support a probabilistic analysis of causation (according to which "A causes B" can be plausibly identified with "A raises the probability of B") provided that the probabilities appealed to are what they call "agent probabilities," where

> [a]gent probabilities are to be thought of as conditional probabilities, assessed from the agent's perspective under the supposition that antecedent condition is realized *ab initio*, as a free act of the agent concerned. Thus the agent probability that one should ascribe to B conditional on A is the probability that B would hold were one to choose to realize A. (1993: 190)

The idea is thus that the agent probability of B conditional on A is the probability that B would have conditional on the assumption that A has a special sort of status or history—in particular, on the assumption that A is realized by a free act. A will be a cause of B just in case the probability of B conditional on the assumption that A is realized by a free act is greater than the unconditional probability of B; A will be a spurious cause of B just in case these two probabilities are equal. As an illustration, consider a structure in which atmospheric pressure, represented by a variable Z, is a common cause of the reading X of a barometer and the occurrence of a storm Y, with no causal relationship between X and Y. X and Y will be correlated, but Price's and Menzies' intuitive idea is that conditional on the realization of X by a free act, this correlation will disappear, indicating that the correlation between X and Y is spurious and does not reflect a causal connection from X to Y. If, by contrast, this correlation were to persist, this would be an indication that X was after all a cause of Y. (What "free act" might mean in this context will be explored below, but a charitable interpretation, although not one that Price and Menzies explicitly adopt, is that the manipulation of X should satisfy the conditions we would associate with an ideal experiment designed to determine whether X causes Y—thus, for example, the experimenter should manipulate the position of the barometer dial in a way that is independent of the atmospheric pressure Z, perhaps by setting its value after consulting the output of some randomizing device.)

Menzies and Price claim that they can appeal to this notion of agency to provide a non-circular, reductive analysis of causation. They claim that circularity is avoided because we have a grasp of the *experience* of agency that is independent of our grasp of the general notion of causation.

> The basic premise is that from an early age, we all have direct experience of acting as agents. That is, we have direct experience not merely of the Humean succession of events in the external world, but of a very special class of such successions: those in which the earlier event is an action of our own, performed in circumstances in which we both desire the later event, and believe that it is more probable given the act in question than it would be otherwise. To put it more simply, we all have direct personal experience of doing one thing and thence achieving another. … It is this common and commonplace experience that licenses what amounts to an ostensive definition of the notion of 'bringing about'. In other words, these cases provide direct non-linguistic acquaintance with the concept of bringing about an event; acquaintance which does not depend on prior acquisition of any causal notion. An agency theory thus escapes the threat of circularity. (1993: 194–5)

Menzies and Price recognize that, once the notion of causation has been tied in this way to our "personal experience of doing one thing and hence achieving another" (1993: 194), a problem arises concerning unmanipulable causes. To use their own example, what can it mean to say that "the 1989 San Francisco earthquake was caused by friction between continental plates" (1993: 195) if no one has (or given the present state of human capabilities could have) the direct personal experience of bringing about an earthquake by manipulating these plates? Their response to this difficulty is complex, but the central idea is captured in the following passages

> … we would argue that when an agent can bring about one event as a means to bringing about another, this is true in virtue of certain basic intrinsic features of the situation involved, these features being essentially non-causal though not necessarily physical in character. Accordingly, when we are presented with another situation involving a pair of events which resembles the given situation with respect to its intrinsic features, we infer that the pair of events are causally related even though they may not be manipulable. (1993: 197)
>
> Clearly, the agency account, so weakened, allows us to make causal claims about unmanipulable events such as the claim that the 1989 San Francisco earthquake was caused by friction between continental plates. We can make such causal claims because we believe that there is another situation that models the circumstances surrounding the earthquake in the essential respects and does support a means-end relation between an appropriate pair of events. The paradigm example of such a situation would be that created by seismologists in their artificial simulations of the movement of continental plates. (1993: 197)

One problem with this suggestion has to do with how we are to understand the "intrinsic" but (allegedly) "non-causal" features in virtue of which the movements of the continental plates "resemble" the artificial models which the seismologists are able to manipulate. It is well-known that small scale models and simulations of naturally occurring phenomena that superficially resemble or mimic those phenomena may nonetheless fail to capture their causally relevant features because, for example, the models fail to “scale up”—because causal processes that are not represented in the model become quite important at the length scales that characterize the naturally occurring phenomena. Thus, when we ask what it is for a model or simulation which contains manipulable causes to “resemble” phenomena involving unmanipulable causes, the relevant notion of resemblance seems to require that the same *causal* processes are operative in both. If the extension of their account to unmanipulable causes requires a notion of resemblance that is already causal in character and which, *ex hypothesi* cannot be explained in terms of our experience of agency, then their reduction fails.

It might be thought that this difficulty can be avoided by the simple expedient of adopting a counterfactual formulation of the manipulability theory. Indeed, it is clear that *some* counterfactual formulation is required if the theory is to be even remotely plausible: after all, no one supposes that A can only be a cause of B if A is in fact manipulated. One thus might consider a formulation along the lines of:

(CF) A causes B if and only if B would change if an appropriate manipulation [by humans] on A *were* to be carried out.

The suggestion under consideration attempts to avoid the difficulties posed by causes that are not manipulable by human beings by contending that for (CF) to be true, it is not required that the manipulation in question be practically possible for human beings to carry out or even that human beings exist. Instead all that is required is that if human beings were to carry out the requisite manipulation of A (e.g., the continental plates), B (whether or not an earthquake occurs) would change. (The possibility of adopting such a counterfactual formulation is sympathetically explored, but not fully endorsed by Ernest Sosa and Michael Tooley in the introduction to their 1993.)

One problem with this suggestion is that, independently of whether a counterfactual formulation is adopted, the notion of a free action or human manipulation cannot by itself, for reasons to be described in §3, do the work (that of distinguishing between genuine and spurious causal relationships) that Menzies and Price wish it to do. But in addition to this, a counterfactual formulation along the lines of (CF) seems completely unilluminating unless accompanied by some account of how we are to understand and assess such counterfactuals and, more specifically, what sort of situation or possibility we are supposed to envision when we imagine that the antecedent of (CF) is true. Consider, for example, a causal claim about the very early universe during which temperatures are so high that atoms and molecules and presumably anything we can recognize as an agent cannot exist. What counterfactual scenario are we supposed to envision when we ask, along the lines of (CF), what would happen if human beings were to exist and were able to carry out certain manipulations in this situation? A satisfying version of an agency theory should give us an account of how our experience of agency in ordinary contexts gives us a purchase on how to understand and evaluate such counterfactuals. To their credit, Menzies and Price attempt to do this, but it is not clear that they are successful.

3. Causation and Free Action

As we have seen, Menzies and Price assign a central role to “free action” in the elucidation of causation. They do not further explain what they mean by this phrase, preferring instead, as the passage quoted above indicates, to point to a characteristic experience we have as agents. It seems clear, however, that whether (as soft determinists would have it) a free action is understood as an action that is uncoerced or unconstrained or due to voluntary choices of the agent, or whether, as libertarians would have it, a free action is an action that is uncaused or not deterministically caused, the persistence of a correlation between A and B when A is realized as a “free act” is *not* sufficient for A to cause B. Suppose , in the example described above, the position of the barometer dial X is set by a free act (in either of the above senses) of the experimenter but that this free act (and hence X) is correlated with Z , the variable measuring atmospheric pressure, perhaps because the experimenter observes the atmospheric pressure and freely chooses to set X in a way that is correlated with Z . (This possibility is compatible with the experimenter’s act of setting X being free in either of the above two senses.) In this case, X will remain correlated with Y when produced by a free act, even though X does not cause Y . Suppose, then, that we respond to this difficulty by adding to our characterization of A’s being realized by a free act the idea that this act must not itself be correlated with any other cause of A. (Passages in Price 1991 suggest such an additional proviso, although the condition in question seems to have nothing to do with the usual understanding of free action.) Even with this proviso, it need not be the case that A causes B if A remains correlated with B when A is produced by an act that is free in this sense, since it still remains possible that the free act that produces A also causes B via a route that does not go through A. As an illustration, consider a case in which an experimenter's administration of a drug to a treatment group (by inducing patients to ingest it) has a placebo effect that enhances recovery, even though the drug itself has no effect on recovery. There is a correlation between ingestion of the drug and recovery that persists under the experimenter's free act of administering the drug even though ingestion of the drug does not cause recovery.

4. Interventions

Examples like those just described show that if we wish to follow Menzies and Price in defending the claim that if an association between A and B persists when A is given the right sort of "independent causal history" or is "manipulated" in the right way, then A causes B, we need to be more precise by what we mean by the quoted phases. There have been a number of attempts to do this in the recent literature on causation. The basic idea that all of these discussions attempt to capture is that of a "surgical" change in A which is of such a character that if any change occurs in B, it occurs only as a result of its causal connection, if any, to A and not in any other way. In other words, the change in B, if any, that is produced by the manipulation of A should be produced only via a causal route that goes through A. Manipulations or changes in the value of a variable that have the right sort of surgical features have come to be called *interventions* in the recent literature (e.g., Spirtes, Glymour, and Scheines 2000; Meek and Glymour 1994; Hausman 1998; Pearl 2009; Woodward 1997, 2000, 2003; Woodward and Hitchcock 2003; Cartwright 2003) and this entry will follow this practice. The characterization of the notion of an intervention is rightly seen by many writers as central to the development of a plausible version of a manipulability theory. One of the most detailed attempts to think systematically about interventions and their significance for understanding causation is due to Pearl 2009 and is discussed in the following section.

5. Structural Equations, Directed Graphs, and Manipulationist Theories of Causation

A great deal of recent work on causation has used systems of equations and directed graphs to represent causal relationships. Judea Pearl (e.g., Pearl 2009) is an influential example of this approach. His work provides a striking illustration of the heuristic usefulness of a manipulationist framework in specifying what it is to give such systems a causal interpretation.[1] Pearl characterizes the notion of an intervention by reference to a primitive notion of a causal mechanism. A functional causal model is a system of equations $X_i = F(Pa_i, U_i)$ where Pa_i represents the parents or direct causes of X_i that are explicitly included in the model and U_i represents an error variable that summarizes the impact of all excluded variables. Each equation represents a distinct causal mechanism which is understood to be "autonomous" in the sense in which that notion is used in econometrics; this means roughly that it is possible to interfere with or disrupt each mechanism (and the corresponding equation) without disrupting any of the others. The simplest sort of intervention in which some variable X_i is set to some particular value x_i amounts, in Pearl's words, to

> lifting X_i from the influence of the old functional mechanism $X_i = F_i(Pa_i, U_i)$ and placing it under the influence of a new mechanism that sets the value x_i while keeping all other mechanisms undisturbed. (Pearl 2009: 70; I have altered the notation slightly)

In other words, the intervention disrupts completely the relationship between X_i and its parents so that the value of X_i is determined entirely by the intervention. Furthermore, the intervention is surgical in the sense that no other causal relationships in the system are changed. Formally, this amounts to replacing the equation governing X_i with a new equation $X_i = x_i$, substituting for this new value of X_i in all the equations in which X_i occurs but leaving the other equations themselves unaltered. In a graphical representation of causal relationships (see below), an intervention of this sort on a variable X_i breaks or removes all other arrows directed into X_i, so that the value of X_i is now completely fixed by the intervention. Pearl's assumption is that the other variables that change in value under this intervention will do so only if they are effects of X_i.

Again, if we want to use this notion of an intervention to elucidate what it is for X to cause Y it is natural to move to a counterfactual formulation in the sense that what matters for whether X causes Y is what would happen to Y if an intervention on X of the sort described above were to occur. Following what has become an established usage I will call such counterfactuals, the antecedents of which correspond to claims about interventions (If X were set to value x under an intervention, then…) *interventionist counterfactuals*. These are the counterfactuals that (under some interpretation, perhaps not necessarily involving Pearl's particular notion of an intervention) seem most suitable for formulating a manipulability theory of causation.

The need for such a counterfactual formulation raises several questions that will be explored in more detail below. First, how should one understand (what is the appropriate interpretation of or semantics for) the counterfactuals in question? Without attempting to answer this question in detail, it seems plausible that if interventionist counterfactuals are to be useful in elucidating causal claims, their semantics must be different from the familiar Lewis/Stalnaker possible world semantics in some respects, as is argued by Woodward (2003), Briggs (2012), and Fine (2012). For example, on the Lewis/Stalnaker semantics, counterfactuals with logically or metaphysically impossible antecedents are always vacuously true, but if there are causal claims that might be associated with such counterfactuals, we don't want them to be automatically true (cf. §12).

A second difference is that an interventionist counterfactual of form "If an intervention were to set $X = x$, then $Y = y$" is most naturally understood as requiring for its truth that *all* such interventions (or at least all such interventions within the background circumstances in which the causal model of interest is taken to hold) would be followed by $Y = y$. This has the consequence that, for example, "strong centering" which holds for the Lewis/Stalnaker semantics, does *not* hold for interventionist counterfactuals. According to strong centering the actual world is more similar to itself than any other possible world. Thus if both p and q hold in the actual world, then the "counterfactual" (that is, subjunctive conditional) "if p were the case, q would be the case", is automatically true, As an illustration of the difference this makes, suppose that X and Y obey the following intervention–supporting functional relation: If and only if $X = 1.5$, then $Y = 3$. Suppose that in the actual world, $X = 1.5$, $Y = 3$. Now consider the counterfactual C : If $1 < X < 3$, then $Y = 3$. Assuming strong centering, the closest world to the actual world in which the antecedent of C is true is the actual world in which $X = 1.5$. In this world, $Y = 3$, so C is true. By contrast, C is false under an interventionist interpretation, since values of X between 1 and 3 other than 1.5 are not followed by 3. Arguably the interventionist verdict that C (and the associated causal claim that "X being between 1 and 3 causes $Y = 3$)" are false is the correct view. Several other differences between interventionist counterfactuals and the Lewis/Stalnaker semantics will be noted below.

A second general issue, related to the one just described, concerns the sense, if any, in which interventions must be "possible" and the bearing of this on the truth of the associated causal claims. Returning to the notion of intervention associated with Pearl above, note that this notion says nothing about whether there is an actual or even possible causal factor that might accomplish the surgical modification Pearl describes. We may if we wish represent such an intervention I by means of arrow directed into the variable X_i that is intervened on which breaks all other arrows directed into X_i (and Pearl sometimes uses this representation) but both the I variable and this arrow seem dispensable. We could instead just think of X_i as set to some new value in the arrow-breaking or equation replacement manner described above, with no further restrictions on when such a setting operation is possible (or when it is permissible or legitimate to invoke it). I will call this a *setting intervention*. This contrasts with an alternative conception of interventions and their connection to causal claims according to which the truth of a claim like "X causes Y" requires that interventions on X must be "possible" in some non-trivial sense of this notion, which then must be specified. (In other words, the truth of "X causes Y" requires both that Y changes under an intervention on X *and* that this intervention be possible.) When a possibility condition of this sort is imposed, I will say we are making use of a *possibility constrained* notion of intervention. Use of this notion raises the difficult question of how the relevant notion of possibility should be understood. I will suggest below that the best way of making sense of this notion is in terms of some notion of conceptual or mathematical (or if you like, "metaphysical") coherence—roughly speaking, the issue is whether there is an appropriately empirically grounded theoretical/mathematical apparatus that allows for a coherent description of the possible intervention in question and allows us to determine what would happen if the intervention were realized. In some cases (see below) such a description may be available even though the intervention in question may not be physically possible.[2] Recognizing obvious worries about the clarity of this notion of possibility (which in my view should be acknowledged by defenders of this notion), one might think that it is preferable to always employ the setting notion in formulating an interventionist account. However, as we shall see, formulations in terms of the "possibility constrained" notion have appealing features (they seem to do a better job of capturing the truth conditions for some causal claims) and a number of writers seem to rely on such a conception.

Returning to Pearl, and following his framework, let us represent the proposition that the value of X has been set by an intervention to some particular value, x_0, by means of a "do" operator ($do(X = x_0)$, or more simply, $do\, x_0$). It is important to understand that conditioning on the information that the value of X has been *set* to x_0 will in general be quite different from conditioning on the information that the value of X has been *observed* to be x_0 (see Meek and Glymour 1994; Pearl 2009). For example, in the case in which X and Y are joint effects of the common cause Z, and X does not cause Y, $P(Y/X = x_0) \neq P(Y)$; that is, Y and X are not independent. However, $P(Y/ do(X = x_0) = P(Y)$; that is, Y will be independent of X, if the value of X is set by an intervention. This is because the intervention on X will break the causal connection from Z to X, so that the probabilistic dependence between Y and X that is produced by Z in the undisturbed system will no longer hold once the intervention occurs. In this way, we may capture Menzies' and Price's idea that X causes Y if and only if the correlation between X and Y would persist under the right sort of manipulation of X.

This framework allows for a simple definitions of various causal notions. For example, Pearl defines the "causal effect" of X on Y associated with the "realization" of a particular value x of X as:

(C) $P(Y \mid do\, x)$

that is, as the distribution that Y would assume under an intervention that sets the value of X to the value x. Again, it is obvious that this is a version of a counterfactual account of causation.

One of the many attractions of this approach is that it yields a very natural account of what it is to give a causal interpretation to a system of equations of the sort employed in the so-called causal modeling literature. For example, if a linear regression equation $Y = aX + U$ makes a causal claim, it is to be understood as claiming that if an intervention were to occur that sets the value of $X = x_0$ in circumstances $U = u_0$, the value of Y would be $y = ax_0 + u_0$, or alternatively that an intervention that changes X by amount dx will change Y by amount a dx. As another illustration consider the system of equations

(1) $Y = aX + U$
(2) $Z = bX + cY + V$

We may rewrite these as follows:

(1) $Y = aX + U$
(3) $Z = dX + W$

where $d = b + ac$ and $W = cU + V$. Since (3) has been obtained by substituting (1) into (2), the system (1)–(2) has exactly the same solutions in X, Y, and Z as the system (1)–(3). Since X, Y and Z are the only measured variables, (1)–(2) and (1)–(3) are "observationally equivalent" in the sense that they imply or represent exactly the same facts about the patterns of correlations that obtain among the measured variables. Nonetheless, the two systems correspond to different causal structures. (1)–(2) says that X is a direct cause of Y and that X and Y are direct causes of Z. By contrast, (1)–(3) says that X is a direct cause of Y and that X is a direct cause of Z but says nothing about a causal relation between Y and Z. We can cash this difference out within the interventionist/manipulationist framework described above—(2) claims that an intervention on Y will change Z while (1)–(3) denies this. (Recall that an intervention on Y with respect to Z must not be correlated with any other cause of Z such as X, and will break any causal connection between X and Y.) Thus while the two systems of equations agree about the correlations so far observed, they disagree about what would happen under an intervention on Y. According to an interventionist/manipulationist account of causation, it is the system that gets such counterfactuals right that correctly represents the causal facts.

One possible limitation of the notion of a setting intervention (or at least Pearl's characterization of it) concerns the scope of the requirement that an intervention on X_i leave intact *all* other mechanisms besides the mechanism that previously determined the value of X_i. If, as Pearl apparently intends, we understand this to include the requirement that an intervention on X_i must leave intact the causal mechanism if any, that connects X_i to its possible effects Y, then an obvious worry about circularity arises, at least if we want to use the notion of an intervention to characterize what it is for X_i to cause Y. A closely related problem is that given the way Pearl characterizes the notion of an intervention, his definition (C) of the causal effect of X on Y, seems to give us not the causal contribution made by $X = x$ alone to Y but rather the combined impact on Y of this contribution and whatever contribution is made to the value of Y by other causes of Y besides X. For example, in the case of the regression equation $Y = aX + U$, the causal effect in Pearl's sense of $X = x$ on Y is apparently $P(Y) = ax + U$, rather than, as one might expect, just ax. In part for these reasons (and for other reasons, described below), Woodward (2003) and Woodward and Hitchcock (2003) explore a different way of characterizing the notion of an intervention which does not make reference to the relationship between the variable intervened on and its effects. For Woodward and Hitchcock, in contrast to Pearl, an intervention I on a variable X is always defined with respect to a second variable Y (the intent being to use the notion of an intervention on X with respect to Y to characterize what it is for X to cause Y). Such an intervention I must meet the following requirements (M1–M4):

(M1) I must be the only cause of X; i.e., as with Pearl, the intervention must completely disrupt the causal relationship between X and its previous causes so that the value of X is set entirely by I,
(M2) I must not directly cause Y via a route that does not go through X as in the placebo example,
(M3) I should not itself be caused by any cause that affects Y via a route that does not go through X, and
(M4) I leaves the values taken by any causes of Y except those that are on the directed path from I to X to Y (should this exist) unchanged.

This characterization makes explicit reference to conditions that must be satisfied by the intervention variable I. Although perhaps not mandatory, questions about what it means for such an I to be possible and how we are to understand the antecedents of the associated interventionist counterfactuals ("If an intervention satisfying (M1)–(M4) on X were to occur,…") thus arise in a natural way on this characterization—or at so I will assume in what follows.

Putting aside these issues about possibility for the present, the most natural way of defining the notion of causal effect in the framework associated with (M1)–(M4) is in terms of the *difference* made to the value of Y by a change or difference in the value of X. (This is also effectively the definition of causal effect adopted in Rubin 1974. Focusing on differences in this way allows us to isolate the contribution made to Y by X alone from the contribution made to Y by its other causes. Moreover, since in the non-linear case, the change in the value of Y caused by a given change in the value of X will depend on the values of the other causes of Y, it seems to follow that the notion of causal effect must be relativized to a background context B_i which incorporates information about these other values. In deterministic contexts, we might thus define the causal effect on Y of a change in the value of X from $X = x$ to $X = x'$ in circumstances B_i as:

(CD) $Y_{\text{do } x, B_i} - Y_{\text{do } x', B_i}$

that is, as the difference between the value that Y would take under an intervention that sets $X = x$ in circumstances B_i and the value that Y would take under an intervention that sets $X = x'$ in B_i, where the notion of an intervention is now understood in terms of (M1)–(M4) rather than in the way recommended by Pearl. In non-deterministic contexts, the characterization of causal effect is less straightforward, but one natural proposal is to define this notion in terms of expectations: If we let $EP_{\text{do } x, B_i}(Y)$ be the expectation of Y with respect to the probability distribution P if X is set to $X = x$ by means of an intervention, then the causal effect on Y of a change in X from $X = x''$ to $X = x$ might be defined as: $EP_{\text{do } x, B_i}(Y) - EP_{\text{do } x', B_i}(Y)$.

This Section will not attempt to adjudicate among these and various other proposals concerning the best way to characterize the notions of intervention and causal effect. Instead, we make the following comment on the general strategy they embody. Note first that the notion of an intervention, when understood along either of the lines described above, is an unambiguously causal notion in the sense that causal notions are required for its characterization—thus the proposals variously speak of an intervention on X as breaking the causal connection between X and its causes while leaving other causal mechanisms intact or as not affecting Y via a causal route that does not go through X. This has the immediate consequence that one cannot use the notion of an intervention to provide a reduction of causal claims to non-causal claims. Moreover, to the extent that reliance on some notion like that of an intervention is unavoidable in any satisfactory version of a manipulability theory, any such theory must be non-reductionist. Indeed, we can now see that critics who have charged manipulability theories with circularity have in one important sense understated their case: manipulability theories turn out to be "circular" not just in the obvious sense that for an action or event I to constitute an intervention on a variable X, there must be a causal relationship between I and X, but in the sense that I must meet a number of other causal conditions as well.

6. Is Circularity a Problem?

Suppose that we agree that any plausible version of a manipulability theory must make use of the notion of an intervention and that this must be characterized in causal terms. Does this sort of "circularity" make any such theory trivial and unilluminating? It is arguable that it does not, for at least two reasons. First, it may be, as writers like Woodward (2003) contend, that in characterizing what it is for a process I to qualify as an intervention on X for the purposes of characterizing what it is for X to cause Y, we need not make use of information about the causal relationship, if any, between X and Y. Instead, it may be that we need only to make use of *other* sorts of causal information, e.g., about the causal relationship between I and Y or about whether I is caused by causes that cause Y without causing X, as in (M1)–(M4) above. To the extent that this is so, we may use one set of claims about causal relationships (e.g., that X has been changed in a way that meets the conditions for an intervention) together with correlational information (that X and Y remain correlated under this change) to characterize what it is for a different relationship (the relationship between X and Y) to be causal. This does not yield a reduction of causal talk to non-causal talk, but it is also not viciously circular in the sense that it presupposes that we already have causal information about the very relationship that we are trying to characterize. One reason for thinking that there must be *some* way of characterizing the notion of an intervention along the lines just described is that we do sometimes learn about causal relationships by performing experiments—and it is not easy to see how this is possible if to characterize the notion of an intervention on X we had to make reference to the causal relationship between X and its effects.

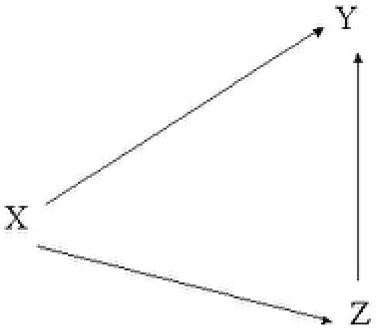
A related point is that even if manipulability accounts of causation are non-reductive, they can *conflict* with other accounts of causation, leading to different causal judgments in particular cases. As an illustration consider a simple version of manipulability account along the lines of (CD), according to which a sufficient condition for X to cause (have a causal effect on Y) is that some change in the value of X produced by an intervention is associated with a change in the value of Y (in the background circumstances of interest). Such an account implies that omissions (e.g., the failure of a gardener to water a plant) can be causes (e.g., of the plant's death) since a change under an intervention in whether the gardener waters is associated with a change in the value of the variable measuring whether the plant dies. For a similar reason relationships involving "double prevention" (Hall 2000) or "causation by disconnection" (Schaffer 2000) count as genuine causal relationships on interventionist accounts. Consider, by contrast, the verdicts about these cases reached by a simple version of a *causal process* theory (in the sense of Salmon 1984, Dowe 2000) according to which a necessary condition for a particular instantiation x of a value X to cause a particular instantiation y of a value Y is that there be a spatio-temporally continuous process connecting x to y involving the transfer of energy, momentum or perhaps some other conserved quantity. According to such a theory, "causation" by omission or by double prevention does not qualify as genuine causation. Similarly, if an "action at a distance" version of Newtonian gravitational theory had turned out to be correct, this would be a theory that described genuine causal relationships according to interventionist accounts of causation, but not according to causal process accounts. Whether one regards the verdicts about these cases reached by causal process accounts or by interventionist accounts as more defensible, the very fact that the accounts lead to inconsistent judgments shows that interventionist approaches are not trivial or vacuous, despite their “circular”, non-reductive character.

7. The Plurality of Causal Concepts

A second respect in which reliance on the notion of an intervention need not be thought of as introducing a vicious circularity is this: So far, I have been following Menzies and Price in assuming that there is just one causal notion or locution (A causes B, where A and B are types of events) that we are trying to analyze. But in fact there are many such notions. For example, among causal notions belonging to the family of so-called type causal notions (i.e., causal claims that relate types of events or variables) there is a distinction to be drawn between what we might call claims about total or net causes and claims about direct causes. Even if the notion of an intervention presupposes some causal notion such as some notion of type causation, it may be that we can use it to characterize other causal notions.

As an illustration consider the causal structure represented by the following equations and associated directed graph

$$Y = aX + cZ$$
$$Z = bX$$



In this structure, there are two different causal routes from X to Y—a direct causal relationship and an indirect relationship with Z as an intermediate variable. If a = –bc, there is cancellation along these two routes. This means that no intervention on X will change the value of Y. In one natural sense, this seems to mean that X does not cause Y, assuming that (C*) a necessary condition for X to cause Y is that some interventions on X are associated with changes in the value of Y, as an obvious extension of CD seems to suggest. In another natural sense, however, X does seem to be a cause—indeed a direct cause—of Y. We can resolve this apparent inconsistency by distinguishing between two kinds of causal claims (for a related distinction, see Hitchcock 2001b)—the claim X is a total or net cause of Y, where this is captured by (C*) and the claim that X is a direct cause of Y, where this is understood along the following lines: X is a direct cause of Y if and only if under some intervention that changes the value of X, the value of Y changes when all other variables in the system of interest distinct from X and Y including those that are on some causal route from X to Y, are held fixed at some value, also by interventions. (For related, but different, characterizations of direct causation along these lines, see Pearl 2009 and Woodward 2003.) Fixing the other values of other variables means that each of these values is fixed by separate interventions that are independent of each other and of the intervention that changes the value of X. The effect of intervening to fix the values of these variables is thus that each variable intervened on is disconnected from its causes, including X. In the example under discussion, X qualifies as a direct cause of Y because if we were to fix the value of Z in a way that disconnects it from the value of X, and then intervene to change the value of X, the value of Y would change. This idea can then be generalized to provide a characterization of “contributing” causation along a causal route, i.e., to capture the sense in which X is an indirect cause of Y along the route that goes through Z (Woodward 2003).

So far our focus has been on type causal claims of various kinds. There are also a number of proposals in the literature that provide interventionist treatments of token or actual cause claims (these have to do with the event of X’s taking on a particular value being an actual cause of Y’s taking on a particular value, as when it is claimed that Jones’ smoking caused his lung cancer), including those that involve various forms of pre-emption and over-determination (e.g., Halpern and Pearl 2005 a, b; ; Hitchcock 2001a; Woodward 2003; Hitchcock 2007a; Hall 2007; Glymour and Wimberly 2007; Halpern and Hitchcock 2015; Halpern 2016; Weslake 2013

[Other Internet Resources]). Considerations of space preclude detailed description, but one strategy that has been explored is to appeal to what will happen to the effect under *combinations* of interventions that both affect the cause and that fix certain other variables to specific values. As an illustration, consider a standard case of causal pre-emption: Gunman one shoots (s_1) victim, causing his death d, while gunman two does not shoot but would have shot (s_2) also causing d, if s_1 had not occurred. If we fix (via an intervention) the behavior of the gunman two at its actual value (he does not shoot), then an independent intervention that alters whether gunman one shoots will alter whether victim dies, thus identifying s_1 as the actual cause of d, despite the absence of counterfactual dependence (of the usual sort) between d and s_1. Accounts along these lines are able to deal with a number (although admittedly not all; see Hitchcock 2007a for details) of the standard counterexamples to other counterfactual treatments of token causation.

It is worth adding that although this appeal to combinations of interventions may seem artificial, it maps on to standard experimental procedures in an intuitive way. Consider a case of genetic redundancy—gene complex G_1 is involved in causing phenotypic trait P but if G_1 is inactivated another gene complex G_2 (which is inactive when G_1 is active) will become active and will cause P. The geneticist may test for this possibility by, first, intervening on G_2 so that it is fixed at the value = inactive, then intervening to vary G_1 and observing whether there is a corresponding change in P. Second, the investigator may intervene to render G_1 inactive and then, independently of this intervening to change G_2 and observing whether there is a change in P. As this example illustrates, we may think of different complex causal structures in which there are multiple pathways, redundancy, cancellation and so on, as encoding different sets of claims about what will happen under various possible combinations of interventions.

Thus even if a “manipulationist” or “interventionist” framework does not yield a reduction of causal talk to non-causal talk, it provides a natural way of marking the distinctions among a number of different causal notions and exhibiting their interrelations. More generally, even if a manipulationist account of causation does not yield a reduction but instead simply connects “causation” (or better, various more specific causal concepts) with other concepts within the same circle, we still face many non-trivial choices about how the concepts on this circle are to be elucidated and connected up with one another. For example, it is far from obvious how to characterize the notion of an intervention so as to avoid the various counterexamples to a version of the manipulability theory like that of Menzies and Price. It is in part because the notion of manipulation/intervention has an interesting and complex fine structure—a structure that is left largely unexplored in traditional manipulability theories-- that working out the connection between causation and manipulation turns out to be interesting and non-trivial rather than banal and obvious.

8. Interventions That Do Not Involve Human Action

We noted above that a free action need not meet the conditions for an intervention, on any of the conceptions of intervention described in §5. It is also true that a process or event can qualify as an intervention even if it does not involve human action or intention at any point. This should be apparent from the way in which the notion of an intervention has been characterized, for this is entirely in terms of causal and correlational concepts and makes no reference to human beings or their activities. In other words, a purely “natural” process involving no animate beings at all can qualify as an intervention as long as it has the right sort of causal history—indeed, this sort of possibility is often described by scientists as a “natural experiment”. Moreover, even when manipulations are carried out by human beings, it is the causal features of those manipulations and not the fact that they are carried out by human beings or are free or are attended by a special experience of agency that matters for recognizing and characterizing causal relationships. Thus, by giving up any attempt at reduction and characterizing the notion of an intervention in causal terms, an “interventionist” approach of the sort described under §§4–5 and 7 avoids the second classical problem besetting manipulability theories—that of anthropocentrism and commitment to a privileged status for human action. For example, under this approach X will qualify as a (total) cause of Y as long as it is true that for some value of X that if X were to be changed to that value by a process having the right sort of causal characteristics, the value of Y would change. Obviously, this claim can be true even if human beings lack the power to manipulate X or even in a world in which human beings do not or could not exist. There is nothing in the interventionist version of a manipulability theory that commits us to the view that all causal claims are in some way dependent for their truth on the existence of human beings or involve a “projection” on to the world of our experience of agency.

9. Interventions and Counterfactuals

We noted above that interventionist versions of manipulability theories are counterfactual theories. What is the relationship between such theories and more familiar versions of counterfactual theories such as the theory of David Lewis? Lewis’ theory is an account of what it is for one individual token event to cause another while, as explained above, versions of interventionist treatments are available for different sorts of type causal claims as well as token causal claims. But if we abstract away from this, there are both important similarities and important differences between the two approaches. As readers of Lewis will be aware, any counterfactual theory must explain what we should envision as changed and what should be held fixed when we evaluate a counterfactual the antecedent of which is not true of the actual world—within Lewis’ framework, this is the issue of which worlds in which the antecedent of the counterfactual holds are “closest” or “most similar” to the actual world. Lewis’ answer to this question invokes a “similarity” ordering that ranks the importance of various respects of resemblance between worlds in assessing overall similarity. (Lewis 1979). For example, avoiding diverse, widespread violations of law is said to be the most important consideration, preserving perfect match of particular fact over the largest possible spatio-temporal region is next in importance and more important than avoiding small localized violations of law, and so on. As is well-known the effect of this similarity ordering is, at least in most situations, to rule out so-called “back-tracking” counterfactuals (e.g., the sort of counterfactual that is involved in reasoning that if the effect of some cause had not occurred, then the cause would not have occurred). When the antecedent of a counterfactual is not true of the actual world, Lewis’ similarity metric commonly leads us (at least in deterministic contexts) to think of that antecedent as made true by a “small” miracle.

The notion of an intervention plays a somewhat (but only somewhat) similar role within manipulability theories of causation to Lewis’ similarity ordering. Like Lewis’ ordering, the characterization of an intervention tells us what should be envisioned as changed and what should be held fixed when we evaluate the sorts of counterfactuals that are relevant to elucidating causal claims. For example, on Pearl’s understanding of an intervention, in evaluating an interventionist counterfactual like “If X were to be set by an intervention to such and such a value, the value of Y would be so and so”, we are to consider a situation in which the previously existing causal relationship between X and its causes is disrupted, but all other causal relationships in the system of interest are left unchanged. A moment’s thought will also show that, as in Lewis’ account, both Pearl’s (in its setting version) and the characterization of interventions in terms of M1–M4 rule out backtracking counterfactuals—for example, in evaluating a counterfactual of the form “if an intervention were to occur that changes E, (where E is an effect of C), then C would change”, Pearl holds that we should consider a situation in which the relationship between E and its causes (in this case, C) is disrupted, but all other causal relationships are left unchanged, so that C still occurs, and the above counterfactual is false, as it should be. Moreover, there is a clear similarity between Lewis’ idea that the appropriate counterfactuals for analyzing causation are often counterfactuals the antecedents of which are made true by “miracles”, and the idea of an intervention as an exogenous change that disrupts the mechanism that was previously responsible for the cause event *C*—both of these notions function so as to provide C with the kind of “independent causal history” (recall Menzies and Price) that allows us to distinguish the effects (if any) of C on E from the effects of other “confounding” variables on E. From this perspective, one might think of an interventionist treatment of causation as explaining why Lewis’ account, with its somewhat *ad hoc* looking similarity ordering, works as well as it does—Lewis’ account works because his similarity ordering picks out roughly those relationships that are stable under interventions and hence exploitable for purposes of manipulation and control and, as a manipulability theory claims, it is just these relationships that are causal.

As noted above, however, this is *not* to say that the two approaches are identical or always yield identical assessments of particular causal and counterfactual claims. One central difference is that Lewis’ account is reductionist in aspiration—the elements that go into his similarity metric (avoidance of big miracles, perfect match of particular facts etc.) are (at least officially) characterized in non-causal, non-modal terms. By contrast, as explained above, the notion of an intervention and the standards for evaluating counterfactuals to which it gives rise are characterized in causal terms, so that the resulting account is non-reductionist.

There are other differences as well, a number of which are explored in an important paper by Briggs (2012). We have already noted that strong centering holds in Lewis' semantics but not for counterfactuals with an interventionist interpretation. In addition, the inference from (i) "if p or q were the case, r would be the case" to (ii) "if p were the case, r would be the case" is invalid within Lewis' semantics but valid if these counterfactuals are given an interventionist interpretation (Briggs 2012; Fine 2012). Very roughly this is because within an interventionist framework, (i) is interpreted as claiming that for any realization of its antedent-- either *p* or *q*-- *r* will follow. It is arguable that in each of these cases, the assessments provided by the interventionist interpretation are correct, assuming that what we want to capture are those counterfactuals that behave in a way that is appropriate for causal interpretation. In addition, Woodward 2003 describes several specific examples in which the two approaches diverge in their judgments about which causal relations are present and in which the interventionist approach seems more satisfactory.[3]

10. Possible and Impossible Interventions

In the versions of a manipulability theory considered under §5ff above, causal claims are elucidated in terms of counterfactuals about what would happen under interventions. As we have seen, the notion of an intervention should be understood without reference to human action, and this permits formulation of a manipulability theory that applies to causal claims in situations in which manipulation by human beings is not a practical possibility.

However, as already intimated, interesting questions arise about how far this framework may be extended to other sorts of cases in which interventions are not "possible". These also illustrate some additional differences between thinking of interventions as setting or, alternatively, in terms of M1–M4 and as possibility constrained. Consider the (presumably true) causal claim (G):

(G) The gravitational attraction of the moon causes the motion of the tides.

Within Pearl's framework and using the notion of a setting intervention, it might be argued that there is no problem with capturing claims like (G), at least if we assume (as we did above) that the relevant setting operation is always "possible" or legitimate: we just imagine the gravitational attraction of the moon set to some different value via a setting intervention (we don't need to specify how this comes about—whether the mass of the moon or its distance from the earth etc. is different) and then note (by applying Newtonian gravitational theory) that the motion of the tides would be different.[4]

Suppose, by contrast, we require that interventions be "possible" in some more demanding sense (that is, we adopt a notion of possibility constrained intervention) and we consider counterfactuals of the form: "if an intervention meeting M1–M4 were to occur that sets the gravitational attraction of the moon to a different value, then…". It may well be that there is no physically possible process that will meet the conditions M1–M4 for intervention on the moon's position with respect to the tides—all possible processes that would alter the gravitational force exerted by the moon may be insufficiently "surgical". For example, it may very well be that any possible process that alters the position of the moon by altering the position of some other massive object will have an independent impact on the tides in violation of condition (M2) for an intervention. Woodward (2003) argues that nonetheless we have a principled basis in Newtonian mechanics and gravitational theory themselves for answering questions about what would happen if such a surgical intervention were to occur and that this is enough to vindicate the causal claim (G). On this view of the matter, what is crucial is not whether the antecedent of the relevant counterfactual is nomologically or physically possible but rather whether we are in possession of well-grounded scientific theories and accompanying mathematics that allow us to reliably answer questions about what would happen under the supposition of such antecedents. We count interventions as "possible" as long as this is the case. Such interventions should be distinguished from interventions that are logically, conceptually or mathematically inconsistent or incoherent (see below for additional illustrations).

11. The Scope of Interventionist Accounts

One context in which issues about the range of cases in which interventionist account may be legitimately or fruitfully applied concerns "cosmological" claims in which fundamental physical theories are understood as applying to the whole universe. Consider the following claim

(4) The state S_t of the entire universe at time t causes the state S_{t+d} of the entire universe at time t + d, where S_t and S_{t+d} are specifications in terms of some fundamental physical theory.

On an interventionist construal, (4) is unpacked as a claim to the effect that under some possible intervention that changes S_t, there would be an associated change in S_{t+d}. This raises the worry that it is unclear what would be involved in such an intervention (given that there is nothing in addition to S_t that might realize the intervention) and unclear how to assess what would happen if it were to occur, given the stipulation that S_t is a specification of the entire state of the universe.

Commenting on an example like this, Pearl writes:

> If you wish to include the whole universe in the model, causality disappears because interventions disappear—the manipulator and the manipulated lose their distinction. (2009: 419-20)

Note that here Pearl seems to invoke a notion of intervention that is different from (and stronger than) a pure setting conception. After all, as Reutlinger (2012) notes, it is arguable that there is no problem about imagining the state of the universe at S_t set to some different value and then determining by reference to the laws governing its evolution what its state will be at S_{t+1}.[5] Pearl's remark seems to assume that the imagined intervention has to meet some additional constraint beyond this (having to do in some way with the possibility of realizing the intervention). Pearl's claim is controversial—it is discussed sympathetically by Hitchcock 2007b and Woodward 2007 and criticized by other writers such as Reutlinger 2012.

We will not try to resolve the issues surrounding this particular claim of Pearl's here, but there is a related and more general issue concerning the implications of interventionism for the status of causal claims in physics, even outside of cosmological contexts, that deserves discussion. Return to the contrast between explicating causal claims by appealing to a pure setting notion of intervention and, alternatively, explicating them by reference to interventions that meet some further non-trivial constraints regarding possibility, as discussed above. Consider cases in which there is a physical law according to which there is counterfactual dependence between Y and X but interventions on X are in some appropriately relevant sense impossible. A pure setting treatment may conclude that such relationships are causal while an account relying on a possibility constrained notion of intervention will not.

Two possible illustrations are discussed in Woodward (2016). The field equations of General Relativity describe a lawful or nomological relationship between the stress energy tensor and the spacetime metric. "Setting" the former to different values (by specifying initial and boundary conditions) one may calculate the associated different values of the latter. One may doubt, however, that it is appropriate to think of the field equations as describing a *causal* relationship between the stress-energy tensor and the metric. It is arguable that a possibility constrained interventionist account supports this judgment: specification of the stress energy tensor requires reference to the metric in such a way that interventions on the former with respect to the latter will violate the conditions M1–M4 for an intervention. One might conclude on these grounds that although there is a relation of nomic dependence between the state of the stress energy tensor and the metric, this relation is not causal. Employment of a setting conception of intervention seems to realize the opposite conclusion.

As a second example, the spins of the entangled particles in an EPR–type experiment are lawfully related by a conservation law. It is arguable (cf. Skyrms 1984; Butterfield 1992) that many standard philosophical theories, including regularity and Lewis-style counterfactual theories treat this relationship as causal, and a setting version of an interventionist theory seems to suggest a similar conclusion. By contrast, various no signaling theorems are commonly interpreted as implying that it is impossible both to intervene on one of the separated spin settings and to use the relationship between the two settings to manipulate the other setting. In this case a possibility constrained version of interventionism can judge that no causal relationship is present. Although the matter is controversial among philosophers, most physicists agree with this judgment of non-causality.

Both of these examples illustrate different implications of setting and possibility constrained versions of interventionism in physics contexts and how the latter framework requires more than just the presence of a nomically sufficient condition or law-based counterfactual dependence for causation. More generally, if one thinks, as many philosophers of physics and some physicists do, that causal concepts do not apply, at least in any straightforward way, to some or many fundamental physics contexts, then it is arguably a consideration in favor of a version of interventionism that imposes a non-trivial possibility constraint that it might be used to support this judgment. By contrast, a setting version of interventionism will tend to find causation in physics whenever there is nomic dependence.

There has been considerable discussion recently both about the extent to which fundamental physics is causal and about what interventionist frameworks imply about the status of causal claims in physics. A number of the essays in Price and Corry 2007 (Price 2007; Hitchcock 2007b; Woodward 2007) express varying degrees of skepticism about the applicability of causal notions in portions of physics, in part on the basis of interventionist considerations. By contrast, Frisch (2014) argues vigorously that many physical theories, at least in classical physics, such as classical electromagnetism, make extensive use of causal concepts and that the relevant notion of cause is captured by the interventionist framework and associated technical ideas (such as structural equations and directed graphs). He suggests that writers like Price, Hitchcock, and Woodward (in his 2007 but see his 2016 for a more nuanced view) underestimate the degree to which interventionist ideas of causation are applicable to such contexts. Of course it is also possible, consistently, with the views of both Frisch and these other writers, that causal notions, understood along possibility constrained interventionist lines, are important in many areas of physics but that there are other physical theories that are not fruitfully interpreted as making causal claims, whether understood along interventionist or other lines. In any case, the question of the scope of interventionist theories and their implications for causal claims in fundamental physics is an important and at present unresolved issue.[6]

12. (Alleged) Causes That Are Unmanipulable for Logical, Conceptual, or Metaphysical Reasons

Several statisticians (e.g., Holland 1986; Rubin 1986), as well as similarly minded epidemiologists (e.g., Hernan and Taubman 2008) who advocate treatments of causation in terms of manipulation-based ideas (in this case in terms of “potential outcome” theory) have held that causal claims involving causes that are unmanipulable in principle are defective or lack a clear meaning—they think of this conclusion as following directly from a manipulationist approach to causation. What is meant by an unmanipulable cause is not made very clear, but what these authors have in mind are not candidate causes that cannot be manipulated as a practical matter, but rather candidates that are such that we lack any clear conception of what would be involved in manipulating them or any basis for assessing what would happen under such manipulations—cases in which manipulation seems “impossible” for conceptual or (if you like) “metaphysical” reasons. Proposed examples include such candidate causes as race, membership in a particular species, and gender. Other examples discussed in this connection involve cases in which there are many different things one might have in mind by “manipulation” of the candidate causes with different results flowing from alternative understandings of “manipulation” so that the claims in question are taken, from a manipulationist or interventionist standpoint, to be unclear or ambiguous. All such cases contrast with the case involving (G) above, where the notion of manipulating the moon’s orbit seems perfectly clear and well-defined, and the problem is simply that the world happens to be arranged in such a way that an intervention that produces such a change is not physically possible.

A sympathetic reconstruction of the position under discussion might go as follows. On an interventionist account of causation, causes (whether we think of them as events, types of events, properties, facts, or what have you) must be representable by means of *variables*—where this means, at a minimum, that it must be possible for the cause to change or to assume different values, for whatever object, unit or system those values are assigned to, as when it is possible for the same particle to be either at position p_1 specified by a position variable P or at some alternative position p_2. This is required if we are to have a well-defined notion of manipulating a candidate cause and well-defined answers to counterfactual queries about what would happen if the cause were to be manipulated in some way—matters which are central to what causal claims mean on any version of a manipulability theory worthy of the name. However, for some putative causes, there may be no well-defined notion of change or variation in value and if so, a manipulability theory will not count these as genuine causes. Suppose, for example, we lack any coherent conception of what it is for something to exist but to be non-physical. Then there will be no well-defined notion of intervening to change whether something is or is not a physical object and being a physical object will not be a factor or property that can serve as a cause. (Of course it is consistent with this that there are true and perhaps even lawful generalizations about all physical objects.) For example, although to the best of our knowledge, it is a law of nature that

(L) no physical object can be accelerated from a velocity less than that of light to a velocity greater than light,

(L) is not, according to this version of a manipulability theory, a *causal* generalization: being a physical object is not a cause of the incapacity in question.

Moreover, even with respect to variables that can take more than one value, the notion of an intervention or manipulation will not be well-defined if there is no well-defined notion of *changing* the values of that variable. Suppose that we introduce a variable "animal" which takes the values {lizard, kitten, raven}. By construction, this variable has more than one value, but if, as seems plausible, we have no coherent idea of what it is to change a raven into lizard or kitten, there will be no well-defined notion of an intervention for this variable and being an animal (or being a raven) will not be the sort of thing that can count as a *bona fide* cause on a manipulability theory. The notion of changing the value of a variable seems to involve the idea of an alteration from one value of the variable to another in circumstances in which the very same system or entity can possess both values and this notion seems inapplicable to the case under discussion.

Note that, just as with some of the examples considered in §12, this conclusion does *not* seem to follow on the pure setting interpretation of the interventionist account. One can, after all, set up an equation $Y = X$ with a candidate variable X, taking the values 0 and 1, according to whether some object is a kitten or lizard and a candidate effect variable Y, taking the values 0 and 1 according to whether that object is warm-blooded or cold-blooded. Then setting X to 0 rather than 1 changes whether $Y = 0$ or 1, and if this is sufficient for causation, being a kitten rather than a lizard causes warmbloodiness rather than coldbloodiness. If one thinks, that there is something defective or problematic about these causal claims, this requires, within an interventionist framework, a richer conception of what is required for causation than what is suggested by the setting conception of intervention. A similar point applies to the other examples described in this section.

Some readers will take it to be intuitively obvious that, e.g., being a raven can be a cause of some particular organism's being black, that being a kitten can be a cause of warm-bloodiness and so on. If causal claims like

(R) "Raveness causes blackness"

are true, it will be an important advantage of a setting version of interventionism over a formulation in terms of a possibility constrained notion of intervention that the former but not the latter is able to capture claims like (R). By contrast, others will think that claims like (R) are, if not false, at least unclear and unperspicuous, and that it is a point in favor of a possibility constrained version of the interventionist account that it can capture this. Those who take this second view will think that claims like (R) should be replaced by claims that involve causes that are straightforwardly manipulable. For example, (R) might be replaced by a claim that identified the genetic factors and biochemical pathways that are responsible for raven pigmentation—factors and pathways for which there is a well-defined notion of manipulation and which are such that if they were appropriately manipulated, this would lead to changes in pigmentation. Theorists like Rubin and Holland will think that such a replacement would be clearer and more perspicuous than the original claim (R). Another illustration of this general idea that replacing claims involving non-manipulable candidate causes with claims involving candidate manipulable causes clarifies their meaning is discussed in The Role of the Manipulability Theory in Clarifying Causal Claims.

13. Some Criticisms of Interventionist Accounts

A number of other criticisms besides the classic charges of anthropomorphism and circularity have been advanced against interventionist accounts. One complaint is that interventionist accounts (at least as I have formulated them) appeal to counterfactuals and that counterfactuals cannot be (as it is often put) “barely true”: if a counterfactual is true, this must be so in virtue of some “truth maker” which is not itself modal or counterfactual. Standard candidates for such truth makers are fundamental laws of nature or perhaps fundamental physical/chemical processes or mechanisms. Often the further suggestion is made that we can then explain the notion of causation in terms of such truth makers rather than along interventionist lines—for example, the notion of causation (as well as the truth conditions for counterfactuals) might be explained in terms of laws (Hiddleston 2005). Thus appealing to interventionist counterfactuals is not necessary, once we take account of the truth conditions of such counterfactuals.

These claims raise a number of issues that can be explored only briefly. First, let us distinguish between providing an ordinary scientific explanation for why some counterfactual claim is true and providing truth conditions (or identifying a truth maker) in the sense described above, where these truth conditions are specified in non-modal, non-counterfactual terms. The expectation that (i) whenever some macro-level interventionist counterfactual is true, there will be some more fundamental scientific explanation of why it is true seems plausible and well grounded in scientific practice. By contrast, the expectation that (ii) for every true counterfactual there must be a truth maker that can be characterized in non-modal, non-counterfactual terms is a metaphysical doctrine that requires some independent argument; it does not follow just from (i). Suppose that it is true that

(5) if subjects with disease D were to be assigned treatment via an intervention with drug G, they would be more likely to recover.

Then it is very plausible that there will be some explanation, which may or may not be known at present, that explains why (5) is true in terms of more fundamental biochemical mechanisms or physical/chemical laws and various initial and boundary conditions. What is less obviously correct is the further idea that we can elucidate these underlying mechanisms/laws without appealing to counterfactuals. It is this further idea that is appealed to when it is claimed that it must be possible to describe a truth maker for a counterfactual like (5) that does not itself appeal to counterfactual or modal claims. The correctness of this idea is not guaranteed merely by the existence of an explanation in the ordinary sense for why (5) is true; instead it seems to depend on whether a reductivist account of laws, mechanisms, etc. in terms of non-modal primitives can be given—a matter on which the jury is still out.[7]

A different line of criticism has been advanced against interventionist accounts in several recent papers by Nancy Cartwright (e.g., 2001, 2002). According to Cartwright such accounts are “operationalist”. Classical operationalism is often criticized as singling out just one possible procedure for testing some claim of interest and contending that the claim only makes sense or only has a truth value when that procedure can actually be carried out. Similarly, Cartwright complains that the interventionist account “overlooks the possibility of devising other methods for measuring” causal relationships and also suggests that the account leads us to

> withhold the concept [of cause] from situations that seem the same in all other aspects relevant to its application just because our test cannot be applied in those situations. (2002: 422)

If interventionism is formulated as above, this criticism seems misplaced. The interventionist account does not hold that causal concepts apply or make sense only when the appropriate interventions can actually be carried out. Nor does it deny that there are other ways of testing causal claims besides carrying out interventions. Instead, interventionism holds that causal claims apply or have truth values whenever the appropriate counterfactuals concerning what would happen if interventions were to be performed have truth values. As explained above, interventionists think that sometimes such counterfactuals are true even if the interventions in question cannot actually be performed. Similarly, interventionists can readily agree that causal claims may be tested and confirmed by, for example, purely observational data, not involving interventions or manipulations—

their view, though, is that what is confirmed in this way is a claim about what would happen if certain interventions were to be performed. In fact, thinking of causal claims in this way helps to clarify why certain strategies for causal inference with observational data, such as the use of instrumental variables, are more likely to lead to reliable conclusions than alternatives (Woodward 2015).

In a related criticism, Cartwright contends that the interventionist account is “monolithic”: it takes just one of the criteria commonly thought to be relevant to whether a relationship is causal—whether it is potentially exploitable for purposes of manipulation—and gives it a privileged or pre-eminent place, allowing it to trump other criteria (like spatio-temporal contiguity or transmission of energy-momentum), when it comes into conflict with them. By contrast, Cartwright favors a “pluralistic” account, according to which a variety of diverse criteria are relevant to whether a relationship is causal and which of these are most appropriate or important will depend on the causal claim at issue.

The interventionist account is indeed mono-criterial. Whether this feature is objectionable depends on whether there are realistic cases in which (i) intervention-based criteria and criteria based on other considerations come into conflict *and* (ii) it is clear that the causal judgments supported by these other criteria are more defensible than those supported by interventionist criteria. Cartwright does not present any uncontroversial cases of this kind. We have seen that interventionist accounts and accounts that take, e.g., spatio-temporal continuity to be crucial for causation do yield conflicting judgments in some realistic cases (e.g., those involving double prevention), but it is far from clear that the interventionist account is mistaken in the judgments that it recommends about such cases.

Two still more recent criticisms directed against M1–M4 and possibility constrained notions of interventionism are Reutlinger 2012 and Glynn 2013. These are discussed in the supplementary document Additional Recent Criticisms of the Interventionist Account.

14. Some Recent Positive Developments.

The material above has largely focused on the use of interventionist or manipulability based ideas to provide an interpretation of causal claims, with little attention paid to the use of these ideas in causal inference—that is, inference to causal relationships from experimental and non-experimental data. The latter is an important subject in its own right. Roughly speaking, if one thinks of causal claims as claims about the outcomes of possible manipulations or experiments, then this suggests distinctive ways of conceptualizing problems of causal inference from non-experimental data: these may be conceptualized as problems of inferring from such data (and other assumptions) what the outcome of a possible experiment would be without doing the experiment in question. This point of view can used to motivate or rationalize the use of such procedures as instrumental variables or regression discontinuity designs—see, e.g., Angrist and Pischke 2009 for econometric applications of these ideas.

Another important extension of interventionist ideas, also with a focus on inference but containing conceptual innovations as well is due to Eberhardt (Eberhardt 2007, Eberhardt and Scheines 2007). These authors generalize the notion of intervention in two ways. First, they consider interventions that do not deterministically fix the value of variable(s) intervened on but rather merely impose a probability distribution on those variables. Second, they explore the use of what have come to be called “soft” interventions. These are interventions that unlike the fully surgical (“hard”) interventions considered above (both Pearl’s setting interventions and the notion associated with M1–M4), do not completely break the previously existing relationships between the variable X intervened on and its causes C but rather supply an exogenous source I of variation to X that leaves its relations to C intact but where I is uncorrelated with C. Certain experiments are naturally modeled in this way. For example, in an experiment in which subjects are randomly given various amounts of additional income (besides whatever income they have from other sources) this additional income functions as a soft, rather than a hard intervention. Soft interventions may be possible in practice or in principle in certain situations in which hard interventions are not. Eberhardt 2007 and Eberhardt and Scheines 2007 explore what can be learned from various combinations of soft and hard, indeterministic and deterministic interventions together with non-experimental data in various contexts. Unsurprisingly each kind of intervention and associated data have both advantages and limitations from the point of view of inference.

Bibliography

Angrist, Joshua D. and Jörn-Steffen Pischke, 2009, “Mostly Harmless Econometrics”, Princeton: Princeton University Press.

Butterfield, Jeremy, 1992, “David Lewis Meets John Bell”, *Philosophy of Science*, 59(1): 26–43. doi:10.1086/289652

Briggs, Rachel, 2012, “Interventionist Counterfactuals”, *Philosophical Studies*, 160(1): 139–166. doi:10.1007/s11098-012-9908-5

Cartwright, Nancy, 2001, “Modularity: It Can—and Generally Does—Fail”, in Maria Carla Galavotti, Patrick Suppes, and Domenico Constantini (eds.) *Stochastic Causality*, Stanford: CSLI Publications.

–––, 2002, “Against Modularity, the Causal Markov Condition, and Any Link Between the Two: Comments on Hausman and Woodward”, *British Journal for the Philosophy of Science*, 53(3): 411–53. doi:10.1093/bjps/53.3.411

–––, 2003, “Two Theorems on Invariance and Causality”, *Philosophy of Science*, 70(1): 203–24. doi:10.1086/367876

Collingwood, R.G., 1940, *An Essay on Metaphysics*, Oxford: Clarendon Press.

Cook, Thomas D. and Donald T. Campbell, 1979, *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Boston: Houghton Miflin Company.

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

Eberhardt, Frederick, 2007, *Causation and Intervention*, (Ph.D. Thesis), Carnegie Mellon University. [Eberhardt 2007 available online]

Eberhardt, Frederick and Richard Scheines, 2007, “Interventions and Causal Inference”, *Philosophy of Science*, 74(5): 981–995. doi:10.1086/525638

Fine, Kit, 2012, “Counterfactuals Without Possible Worlds”, *Journal of Philosophy*, 109(3): 221–246. doi:10.5840/jphil201210938

Frisch, Mathias, 2014, *Causal Reasoning in Physics*, Cambridge: Cambridge University Press.

Gasking, Douglas, 1955, “Causation and Recipes”, *Mind*, 64(256): 479–487.

Glymour, Clark and Frank Wimberly, 2007, “Actual Causation and Thought Experiments”, in Joseph Keim Campbell, Michael O’Rourke, and Harry S. Silverstein (eds.) *Causation and Explanation*, Cambridge, MA: MIT Press, pp 43–67.

Glynn, Luke, 2013, “Of Miracles and Interventions”, *Erkenntnis*, 78 (Supplement 1): 43–64. doi:10.1007/s10670-013-9436-5

Haavelmo, Trygve, 1944, “The Probability Approach in Econometrics”, *Econometrica*, 12 (Supplement), pp. iii–vi, 1–115. doi:10.2307/1906935

Hall, Ned, 2000, “Causation and the Price of Transitivity”, *The Journal of Philosophy*, 97(4): 198–222. doi:10.2307/2678390

Hall, Ned, 2007, “Structural Equations and Causation”, *Philosophical Studies*, 132(1): 109–136.

Halpern, Joseph Y., 2016, *Actual Causality*, Cambridge, MA: MIT Press.

Halpern, Joseph Y. and Christopher Hitchcock, 2015, “Graded Causation and Defaults”, *British Journal for the Philosophy of Science*, 66(2): pp. 413–457. doi:10.1093/bjps/axt050

Halpern, Joseph Y. and Judea Pearl, 2005a, “Causes and Explanations: A Structural Model Approach; Part I: Causes”, *British Journal for the Philosophy of Science*, 56(4): 843–87. doi:10.1093/bjps/axi147

–––, 2005b, “Causes and Explanations: A Structural Model Approach; Part II: Explanations”, *British Journal for the Philosophy of Science*, 56(4): 889–911. doi:10.1093/bjps/axi148

Hausman, Daniel M., 1986, “Causation and Experimentation”, *American Philosophical Quarterly*, 23(2): 143–54

–––, 1998, *Causal Asymmetries*, Cambridge: Cambridge University Press.

Hernan, Miguel and Taubman, Sarah, 2008, “Does Obesity Shorten Life? The Importance of Well-defined Interventions to Answer Causal Questions ”, *International Journal of Obesity*, 32 (Supplement 3): S8–14.

Hiddleston, Eric, 2005, Review of *Making Things Happen* (Woodward 2003), *Philosophical Review*, 114(4): 545–47. doi:10.1215/00318108-114-4-545

Hitchcock, Christopher, 2001a, "The Intransitivity of Causation Revealed in Equations and Graphs", *The Journal of Philosophy*, 98(6): 273–99. doi:10.2307/2678432

–––, 2001b, "A Tale of Two Effects", *Philosophical Review*, 110(3): 361–96. doi:10.1215/00318108-110-3-361

–––, 2007a, "Prevention, Preemption, and the Principle of Sufficient Reason", *Philosophical Review*, 116(4): 495–532. doi:10.1215/00318108-2007-012

–––, 2007b, "What Russell Got Right", in Price and Corry 2007: 45–65.

Hitchcock, Christopher and James Woodward, 2003, "Explanatory Generalizations, Part II: Plumbing Explanatory Depth", *Nôus*, 37(2): 181–99. doi:10.1111/1468-0068.00435

Holland, Paul W., 1986, "Statistics and Causal Inference", *Journal of the American Statistical Association*, 81(396): 945–960.

Lewis, David, 1973, "Causation", *Journal of Philosophy*, 70(17): 556–567.

–––, 1979, "Counterfactual Dependence and Time's Arrow", *Nôus*, 13(4): 455–76. doi:10.2307/2215339

Maudlin, Tim, 2007, *The Metaphysics Within Physics*, Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780199218219.001.0001

Meek, Christopher and Clark Glymour, 1994, "Conditioning and Intervening", *British Journal for the Philosophy of Science*, 45(4): 1001–1021. doi:10.1093/bjps/45.4.1001

Menzies, Peter and Huw Price, 1993, "Causation as a Secondary Quality", *British Journal for the Philosophy of Science*, 44(2): 187–203. doi:10.1093/bjps/44.2.187

Norton, John D., 2007, "Causation as Folk Science", in Price and Corry 2007: 11–44.

Pearl, Judea, 2009, *Causality*, New York: Cambridge University Press.

Price, Huw, 1991, "Agency and Probabilistic Causality", *British Journal for the Philosophy of Science*, 42(2): 157–76. doi:10.1093/bjps/42.2.157

–––, 1992, "Agency and Causal Asymmetry", *Mind*, 101(403): 501–520.

–––, 2007 , "Causal Perspectivalism", in Price and Corry 2007: 250–292.

–––, 2017, "Causation, Intervention and Agency: Woodward on Menzies and Price", in H. Beebee, C. Hitcock, and H. Price (eds.), *Making a Difference: Essays on the Philosophy of Causation*, Oxford University Press, pp. 73–98.

Price, Huw and Richard Corry (eds), 2007, *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, Oxford: Oxford University Press.

Reutlinger, Alexander, 2012, "Getting Rid of Interventions", *Studies in the History and Philosophy of Biological and Biomedical Sciences*, 43(4): 787–95. doi:10.1016/j.shpsc.2012.05.006

Rubin, Donald B., 1974, "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology*, 66(5): 688–701.

–––, 1986, "Comment: Which Ifs Have Causal Answers?", *Journal of the American Statistical Association*, 81(396): 961–962.

Salmon, Wesley C., 1984, *Scientific Explanation and the Causal Structure of the World*, Princeton: Princeton University Press.

Schaffer, Jonathan, 2000, "Causation by Disconnection", *Philosophy of Science*, 67(2): 285–300. doi:10.1086/392776

Skyrms, Brian, 1984, "EPR:Lessons from Metaphysics", *Midwest Studies in Philosophy*, 9(1): 245–55. doi:10.1111/j.1475-4975.1984.tb00062.x

Sosa, Ernest and Michael Tooley (eds.), 1993, *Causation*, Oxford: Oxford University Press.

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

von Wright, Georg Henrik, 1971, *Explanation and Understanding*, Ithaca, NY: Cornell University Press.

Woodward, James/Jim, 1997, "Explanation, Invariance, and Intervention", *Philosophy of Science*, 64(supplement): S26–S41.

–––, 2000, "Explanation and Invariance in the Special Sciences", *British Journal for the Philosophy of Science*, 51(2): 197–254. doi:10.1093/bjps/51.2.197

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

–––, 2007, "Causation with a Human Face", in Price and Corry 2007: 66–105.

–––, 2014, “Simplicity in the Best Systems Account of Laws of Nature”, *British Journal for the Philosophy of Science*, 65: 91–123.

–––, 2015, “Methodology, Ontology and Interventionism”, *Synthese*, 192(11): 3577–3599. doi:10.1007/s11229-014-0479-1

–––, 2016, “Causation in Science”, in *Oxford Handbook of the Philosophy of Science*, edited by Paul Humphreys, New York: Oxford University Press, 163–184; longer online version at doi:10.1093/oxfordhb/9780199368815.013.8

Woodward, James and Christopher Hitchcock, 2003, “Explanatory Generalizations, Part I: A Counterfactual Account”, *Noûs*, 37(1): 1–24. doi:10.1111/1468-0068.00426

Academic Tools

- How to cite this entry.
- Preview the PDF version of this entry at the Friends of the SEP Society.
- Look up topics and thinkers related to this entry at the Internet Philosophy Ontology Project (InPhO).
- Enhanced bibliography for this entry at PhilPapers, with links to its database.

Other Internet Resources

- Weslake, Brad, 2013, “A Partial Theory of Actual Causation”, unpublished manuscript.
- Links to Judea Pearl’s work on causality

Related Entries

causal models | causation: counterfactual theories of | causation: probabilistic | causation: the metaphysics of | mechanism in science | Salmon, Wesley | scientific explanation: causal approaches to

Library of Congress Catalog Data: ISSN 1095-5054