Economics and Philosophy, **19** (2003) 321–340
DOI: 10.1017/S0266267103001184

CRITICAL NOTICE: *CAUSALITY* BY JUDEA PEARL

James Woodward
California Institute of Technology

1

This is a wonderful book; indeed, it is easily one of the most important and creative books I have ever read on the subject of causation and causal inference. *Causality* is impressive on many levels and should be of great interest to many different audiences. Philosophers will find of particular interest Pearl's defense of what might be described as a broadly manipulationist or interventionist treatment of causation: Causal claims have to do with what would happen under ideal, suitably surgical experimental manipulations ('interventions'). But Pearl moves far beyond philosophical generalities on this theme: He makes the case for the approach that he favors by developing a formal apparatus ('a calculus of interventions', as he calls it) for talking about interventions and how they relate to causation and information about probabilities that is at once both intuitively compelling and genuinely useful for purposes of calculation and estimation. He also actively engages the philosophical literature on causation at many points – for example, there are illuminating discussions of evidential decision theory, of probabilistic theories of causation of the sort favored by Eells and Cartwright, of Lewisian counterfactual theories of causation, and of the relationship between so-called type and token causal claims, as well as a quite novel treatment of the latter.

For those interested in foundational questions at the intersection of econometrics, causal inference and statistics, Pearl offers an account, in terms of claims about the outcomes of hypothetical experiments, of what systems of structural equations mean (or should be understood to mean) when interpreted causally. This account draws on ideas that go back to such econometricians as Haavelmo, Frisch, Strotz and Wald, among others, but Pearl expresses these ideas in a way that seems to me to be clearer and analytically sharper and relates them successfully to much more

general ideas about causal inference. Pearl also has illuminating things to say in this connection about such matters as exogeneity, confounding, Simpson's paradox, and the relationship between his own framework and the counterfactual approach to causation associated with writers like Rubin, Holland, and Robbins. In doing so, he provides a powerful defense of the idea that causation is a legitimate (indeed, necessary and unavoidable) subject of inquiry in statistics and econometrics.

In a book that is this rich, a detailed discussion of everything of interest is obviously impossible. I will confine myself to sketching in a non-technical way some of the main ideas of Pearl's approach (section 2) and will then raise some questions about the way in which he understands the notion of an intervention (section 3) and about the details of his characterizations of several causal notions (section 4).

2

Pearl uses systems of equations and directed graphs to represent causal relationships. In particular, a set of causal relationships may be represented by a system of equations (1) $x_i = f_i(pa_i, u_i)$, $i = 1, 2, \ldots, n$, where pa_i represents the *parents* or direct causes of x_i that are explicitly included in the model and u_i represents a so-called error term that summarizes the impact of causally relevant variables that are excluded from that equation. The error terms are assumed to be governed by a joint distribution function $P(u_i)$ and this in turn has the consequence that other variables in the model are also governed by a non-degenerate probability distribution – that is, because of the presence of the error terms, probabilities like $P(X_j \mid X_i)$, $i \neq j$ will have values other than 0 or 1.

Equations like (1) do not just describe the *de facto* relationships that happen to obtain between the values of the variables they relate. Instead they are to be understood as representing 'counterfactual' (p. 310) relationships – that is, as telling us what the response of x_i would be if the variables on the right hand side of (1) were set to various values in an appropriate way by means of interventions or hypothetical experimental manipulations. I will explore in more detail below just what this means. Pearl thinks of each equation as representing 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.

What then is an intervention? Pearl characterizes the notion as follows:

> (**PI**) The simplest type of external intervention is one in which a single variable, say X_i, is forced to take on some fixed value x_i. Such an intervention, which we call 'atomic', amounts 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

https://doi.org/10.1017/S0266267103001184 Published online by Cambridge University Press

> new mechanism that sets the value x_i while leaving all other mechanisms unperturbed. Formally, this atomic intervention, which we denote by do $(X_i = x_i)$ or do (x_i) for short, amounts to removing the equation $x_i = f_i(pa_i, u_i)$ from the model and substituting $X_i = x_i$ in the remaining equations. (p. 70)

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. For example, in the following system of equations,

(2) $Y = aX + U$

(3) $Z = bX + cY$

an intervention that sets $Y = y$ will replace equation (2) with the equation

(2′) $Y = y$,

indicating that the value of Y is now determined entirely by the intervention and is no longer influenced by X and U. The equation (3) will be unaltered by this intervention, so that the value of Z will be determined by the contribution made by the value to which Y has been set – that is, cy – and the contribution made by whatever value X assumes.

Pearl's talk of 'lifting' the variable intervened on from the influence of its (previous) parents may seem puzzling to philosophers who are accustomed to associate causal relationships with the instantiation of (presumably inviolable) 'laws of nature', but in fact the underlying idea is quite intuitive. An intervention replaces a situation in which the variable intervened on, X, is sensitive to changes in the values of certain variables with a new situation in which the value of X is no longer sensitive to such changes but instead depends only on the value assigned to it by the intervention.[1] Many real-life experiments aim at (and succeed) in accomplishing this. For example, in an experiment to test the impact of a drug on recovery, in which subjects are randomly assigned to a treatment group which receives the drug and a control group from which the drug is withheld, the idea is that who receives the drug should be

[1] As Woodward and Hitchcock (forthcoming) and Hausman and Woodward (forthcoming) observe, one natural way of representing this is to think of the intervention variable as a 'switch' variable. For some range of values of the intervention variable I (values for which I takes an 'off' value) the variable intervened on, X, is a function of its parents and the value of I. For other values of I (values for which I is 'on'), the value of X is a function of the value of I alone.

https://doi.org/10.1017/S0266267103001184 Published online by Cambridge University Press

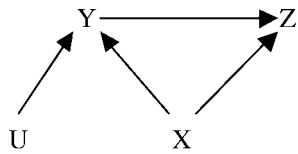
determined entirely by this random assignment (the intervention), and not, as it perhaps was previously, by such other factors as the subjects' own decisions.

As the above quotation indicates, Pearl represents the proposition that the value of X has been set by an intervention to some particular value x by means of a 'do' operator *do* $(X = x)$ or $do(x)$. This allows for simple definitions of various causal notions. For example, the 'causal effect' of X on Y associated with the 'realization' of a particular value x of X is defined as $P(y \mid do(x))$ – this represents the 'total effect' of do $(X = x)$ on Y through all different paths from X to Y. By contrast, the 'direct effect' of $X = x$ on Y is $P(y \mid do(x), do\ (S_{xy}))$ where S_{xy} is the set of all endogenous variables except X and Y in the system. That is, the direct effect is the distribution that y takes under an intervention that sets $X = x$ and fixes by interventions the values of all other variables in the system. According to Pearl, this represents the sensitivity of Y to changes in X alone (p. 126).

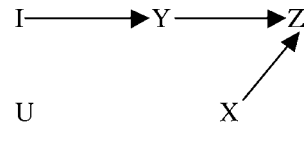
I will return to these characterizations below, but to understand them it is important to recognize that conditioning on the information that the value of X has been set to x will in general be quite different from conditioning on the information that the value of X has been observed to be x, but not as the result of an intervention. For example, in the case in which X and Y are joint effects of the common cause Z, X and Y will not be independent. However, we expect that 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. Unlike ordinary conditioning, interventions involve alterations in causal structure. An account of causal relationships based on interventions so conceived differs in important respects from the sorts of attempts to connect causation and manipulation found in the so-called manipulability accounts of causation advocated by some philosophers (e.g. von Wright, 1971, Menzies and Price, 1993). These writers attempt to characterize causal relationships in terms of some primitive notion of human action, which is claimed to be non-causal in character or at least graspable independently of other ordinary causal notions. Such theories have considerable difficulty explaining how causal notions can ever be legitimately extended to cases in which the appropriate manipulations by human beings are impossible, such as causal relationships involving large scale cosmological processes in the early universe. By contrast, **PI** characterizes the notion of an intervention in purely causal terms that make no essential reference to human action – a purely natural process, not involving human activity at any point, will qualify as an intervention as long as it has the right causal characteristics. Moreover, although this point is not discussed in detail by Pearl, his characterizations of causal notions must be understood

counterfactually or hypothetically. For the causal effect of $X = x$ on Y to be well-defined, it is not required that an intervention that sets $X = x$ actually occur or that it be technologically possible for humans to carry it out. Instead, all that matters is what the distribution of Y would be *were* such an intervention to occur.[2]

Pearl also uses directed graphs to represent systems of causal relationships. Here the basic representational convention is quite simple: A directed edge or arrow from the variable X_1 to the variable X_2 means that X_1 is a direct cause or parent of X_2. Thus the causal structure represented by the equations (2) - (3) and (2′) - (3) may also be represented by the following directed graphs.

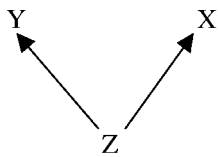


2-3

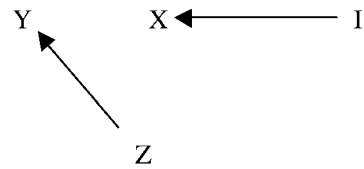


2′-3

As this suggests, interventions have a particularly straightforward representation in terms of directed graphs: An intervention I on a variable Y breaks all (other) arrows directed into Y, indicating that the value of Y is set entirely by I, while preserving all other arrows in the graph, including those directed out of Y. Thus in the case of the common cause structure described above, the effect of an intervention I on one of the effects X is to replace the structure 4 with the structure



4



4′

[2] It would desirable, however, to have more guidance from Pearl about when talk of this distribution and the intervention associated with it is well-defined. For example, what about interventions on variables (e.g., gender, race, membership in a particular species) that for logical or conceptual reasons may be unmanipulable? If R is a variable that represents subject's race, does it make sense to talk of 'lifting' R from the influence of its parents while all other causal relationships are unchanged? Can we meaningfully talk about the educational attainment that a subject would have if he were assigned a different race by an intervention?

https://doi.org/10.1017/S0266267103001184 Published online by Cambridge University Press

Graphs give us a non-parametric representation of causal relationships. A structure in which there are arrows from X and Y into Z tells us that Z is some (non-trivial) function of X and Y but does not tell us what the precise functional form of the relationship between these variables is – e.g., whether it is linear. Much of the interest and power of Pearl's use of directed graphs comes from the fact that he is able to derive many interesting results that depend only on the graphical structures used to represent causal relationships, rather than on more specific assumptions about the functional form of those relationships. In representing causal relationships by means of graphs and equations, Pearl departs in several important respects from common philosophical frameworks for thinking about the relationship between causal and probabilistic claims. Note first that while the variables in equations like (1) are assumed to be governed by a probability distribution, in Pearl's framework causal relationships are *not* directly represented by probabilistic relationships among variables. Instead they are represented by a distinct mathematical object – a system of equations or a directed graph. This contrasts with the common practice among philosophers (especially those who have attempted to develop probabilistic theories of causation) of using just one formal object – a probability distribution, perhaps supplemented by various informal notions such as that of a 'causally homogenous background context' – in constructing theories of causation.[3] This difference is in turn closely related to the tendency within philosophical discussion to attempt to understand causal relationships in terms of the notion of conditioning (i.e., in terms of relationships among conditional probabilities among variables that do not involve interventions) and Pearl's contrasting insistence that causal relationships are to be analyzed in terms of the notion of an intervention, where an understanding of the effect of the latter is provided in terms of graphs and equations. Roughly speaking, within Pearl's framework, equations and directed graphs express the modal and counterfactual commitments carried by causal claims – they record what would happen (what would change and what would remain undisturbed) if certain changes or interventions were to occur. By contrast, the probability distribution records facts about the actual distribution of values of the variables in the model and does not by itself tell us what would happen if causal relationships in the model were changed in various ways, as by

[3] One central problem with such theories is that the import of the probabilistic relationships among X, Y and these background factors B_i for whether X causes Y will depend on the direct causal relationships between X and B_i, between Y and B_i and among the B_i themselves – in other words, it will depend on the rest of the causal route structure as captured by the associated set of equations or directed graph. Just having a list of the B_i or all the other causes of Y besides X is not enough – the information supplied by the associated equations or directed graph is essential too. This is an implication of the Causal Markov condition, discussed below. See (Woodward, 2001) for additional discussion.

an intervention. In my view, this framework results in substantial gains in clarity over standard philosophical treatments of causation in probabilistic contexts.

While causal and statistical relationships are thus sharply separated in Pearl's treatment, it is nonetheless assumed that they will often bear certain systematic relationships to one another. One such relationship which is systematically exploited by Pearl is the Causal Markov (**CM**) condition which connects claims about causal structure to claims about conditional independence relations. A directed graph and an associated probability distribution satisfy the Causal Markov condition if and only if, conditional on its parents, every variable is (probabilistically) independent of every other variable except its descendants, where X_2 is a descendant of X_1 if and only if there is a directed path from X_1 to X_2 – i.e., if and only if X_2 is an effect of X_1. **CM** is a generalization of the ideas about screening off that are a familiar part of philosophical theorizing – **CM** implies, for example, that a single common cause Z will screen off its joint effects X and Y from one another ($P(X \mid Y.Z) = P(X \mid Z)$); that if two variables X and Y are probabilistically dependent, then either X causes Y, Y causes X or their dependence is due to some common cause or causes; and so on. It is easy to prove the following result (**D**): If causal relationships are deterministic, as with equations of form (1) and the error terms u_i are distributed independently of one another and of the right hand side variables in the equations in which they occur, then **CM** (which, it should be noted, is formulated just in terms of variables and their parents and makes no reference to the error terms u_i) must hold.

Pearl's justification for the use of **CM** appeals mainly to **D**. This again reflects another point of contrast between Pearl's approach and those that are common among philosophers: unlike many philosophers, Pearl does not take very seriously the possibility that causal relationships of the sort he is interested in (causal relationships among macroscopic variables of the sort found in engineering, biomedical, or social scientific – but not quantum mechanical – contexts) are indeterministic in a way that precludes their being adequately modeled by deterministic equations with an error term. In other words, he thinks of the apparent indeterminism infecting causal relationships in these domains as really reflecting the operation of deterministic relationships that involve some variables that we are unable to measure. He thus avoids one of the issues that has recently been a source of considerable interest to philosophers, whether there is any reason to expect **CM** to hold in irreducibly indeterministic contexts.[4] His discussion

[4] Nancy Cartwright has been a persistent critic of the idea that there is any general reason to expect **CM** to hold in indeterministic contexts – see Cartwright (1999 and 2002). For additional discussion and defense of **CM**, see Glymour (1999), Spirtes, Glymour and Scheines (2000), and Hausman and Woodward (1999 and forthcoming).

https://doi.org/10.1017/S0266267103001184 Published online by Cambridge University Press

raises the general question of what, if anything, is lost when causal relationships are modeled deterministically with omitted variables, rather than by the indeterministic frameworks favored by many philosophers.

Given the Markov condition **CM** and, on occasion, various other conditions connecting causal claims and probabilities, Pearl is able to investigate a variety of questions about what one can learn about causal relationships from probabilistic information. For example, he provides a very simple characterization of when two different graphs will be 'observationally equivalent' in the sense that, assuming that both conform to **CM**, they imply exactly the same facts about independence relations in the associated probability distribution. In the absence of other constraints or sources of information, the answer to this question places limits on our ability to infer causal claims from statistical information and also on the reducibility of causal relationships to probabilistic relationships.

Questions about identification of causal effects also receive a thorough exploration. In general, this is the problem of inferring the value of expressions involving the 'do' operator from probabilistic information that does not involve the 'do' operator and hence is observable, and other kinds of information, such as information about graphical structure. Suppose that we are given the structure of a graph G describing the causal relationships among a set of variables V. Suppose that we wish to know whether the causal effect of X on Y, $P(y \mid do(x))$, where X and Y are variables in V, is identifiable from measurements on X, Y and various other variables in V. Given **CM**, Pearl provides simple graphical criteria (his 'back door' and 'front door' criteria – pp. 79ff) for identifiability and a simple formula for calculating $P(y \mid do(x))$ from such measurements and our knowledge of graphical structure. As an illustration, consider a structure in which it is known that X causes Y only via an intermediate variable Z, but in which there is also an unmeasured common cause U that also contributes to the correlation between X and Y, although it does not directly affect Z. Despite the confounding influence of U, the causal effect of X on Y turns out to be identifiable from measurements on X, Y and Z.

Pearl also gives conditions for the identifiability of direct effects and formulas for calculating these. Again these are stated only in terms of graphical structure and measurements on some appropriate subset of variables in V; it is not required, for example, that the investigator make assumptions about the precise functional forms of the equations linking the variables in V.

3

We have seen that the notions of an intervention and of a causal mechanism are central and closely interconnected ideas within Pearl's framework. In this section I want to explore the way in which Pearl characterizes these notions and their interconnections and to compare these with an alternate

characterization.[5] My remarks will be those of a sympathetic critic – I fully agree with Pearl's basic premise that the most promising approach for understanding the sorts of causal claims that figure in the social and biomedical sciences will appeal to claims about what will happen under interventions. I will suggest, however, that while Pearl's notion of an intervention is ideally suited to the particular use to which he puts it – that of predicting what will happen to the probability distribution of one variable under a certain kind of change in the value of another (or others) when the causal graph or some portion of it connecting those variables and all others in the system of interest is known, there are other sorts of purposes for which one might wish to use the notion of an intervention for which **PI** is less useful.

One such purpose is *semantic* or *interpretive*: One might hope to use the notion of an intervention to provide a non-trivial account or interpretation of what it is for there to be a causal relationship or causal mechanism connecting two variables. For example, it might be suggested that X causes Y if suitable manipulations or interventions that change the value of X are associated with changes in the value of Y for some values of other variables in the system of interest. Similarly, one might hope to use the notion of an intervention to explain what it means to (or when it is justifiable to) draw an arrow from one variable to another in a directed graph when the graph is given a causal interpretation. Thus it might be suggested that it is appropriate to draw an arrow from X to Y (X is a direct cause of Y) if and only if under some interventions that change the value of X, the value of Y changes when all other variables in the graph are fixed at some value by interventions. Providing accounts of this sort that capture the content of various causal claims in terms of claims about what would happen under possible manipulations is certainly one of the aspirations of traditional manipulability theories of causation. Moreover, it seems clear that Pearl shares this aspiration; for example, he explicitly claims that claims like 'X causes Y' as well as other sorts of causal claims can be 'defined' in terms of his *do* operator.[6]

[5] In thinking about the contrast between the two conceptions of intervention – **PI** and **EI** – that are described in this section, I have been helped a great deal by two papers of Nancy Cartwright's (2002 and forthcoming) that explore in some detail the consequences of different ways of thinking about interventions. In particular, the contrast between **PI** and **EI** is, I believe, related to but not identical with the contrast that Cartwright draws in the first paper between *causal law variation* and *value variation*. As far as I know, however, Cartwright does not share my grounds for preferring **EI** to **PI**. Conversations with Dan Hausman and Christopher Hitchcock have also been very helpful in clarifying my thinking on this topic.

[6] This claim is particularly clear in Pearl's response to a reader's question, 'Has causality been defined?' on his web page (http://www.cs.ucla.edu/~judea/) under 'Discussion with readers'.

https://doi.org/10.1017/S0266267103001184 Published online by Cambridge University Press

A second, closely related use to which one might hope to put the notion of an intervention has to do with the *epistemology* of experimentation: One would like to model cases in which we have limited information about causal relationships in the system of interest and use experimental interventions to learn about previously unknown causal relationships or to correct mistaken views about such relationships.

For a variety of reasons, Pearl's definition of 'intervention', quoted above under **PI**, does not appear to be well-suited for either of these projects. One simple way of seeing this in connection with the interpretive project is to note that **PI** defines the notion of an intervention with respect to the *correct* causal graph for the system in which the intervention occurs – hence **PI** does not give us a notion of intervention that can be used to provide an interpretation for what it is for such a graph to be correct. Put slightly differently, the difficulty is that Pearl takes the notion of a causal mechanism (as represented by an arrow or a functional relationship) as primitive and defines the notion of an intervention in terms of this primitive, thus losing any possibility of using the notion of an intervention to characterize the notion of a causal mechanism or relationship. For it to be informative to say such things as, e.g., "there is a causal mechanism connecting X to Y if the value of Y would change under some intervention on X", we need a notion of 'intervention' that is different from **PI** and that can be used to provide some independent purchase on what a mechanism is and what it is for it be left intact.

This point is obscured, I believe, because it is clear that Pearl has defined or characterized something of interest in terms of $P(y \mid do(x))$. But what he has given us is not a definition of claims like 'X causes Y'. Instead, he has shown us how to define or characterize a quantitative causal notion $P(y \mid do(x))$ in terms of a qualitative causal primitive plus probabilistic information. The quantitative causal notion $P(y \mid do(x))$ corresponds to something like 'the distribution that y would assume if the value of X were to be set to $X = x$ by a certain kind of surgical change' and the causal primitive in terms of which it is defined corresponds to some qualitative or non-parametric notion of causal dependence between the variables X and Y and others in the system of interest, as expressed by the associated directed graph. The definition of this quantitative notion is valuable and important, but contrary to what Pearl seems to suggest, it cannot serve as a definition of what it is for X to cause Y or for there to be a causal mechanism connecting X to Y. Providing a definition of $P(y \mid do(x))$ in terms of a probability distribution and an associated causal graph, which is what Pearl has done, is not the same thing as providing a semantics or interpretation for the graph itself. Because $do(x)$ is defined with respect to a true or correct graph, we cannot appeal to it to provide such a semantics.

We may also put this point in terms of the epistemology of experimentation. One consequence of Pearl's requirement (**PI**) that an intervention

https://doi.org/10.1017/S0266267103001184 Published online by Cambridge University Press

on X leave intact *all* other mechanisms besides those connecting X to its parents is that it builds into the definition of an intervention on X that interventions must leave intact the mechanism, if any, that connects X to its possible effect Y. This in turn has the apparent consequence that in order to know whether an experimental manipulation of X counts as an intervention in the sense of **PI** we must know already whether any mechanism connecting X to Y has been altered by the change in X, which seems to preclude our finding out about the nature of the relationship between X and Y and which changes will disrupt it by intervening on X.

The following examples may help to illustrate these points. First, consider an investigator who is presented with two variables P and Q, who knows nothing about whether there is any causal mechanism connecting them or about the nature of that mechanism, and who wishes to learn by experimentation whether P causes Q. Suppose that the investigator experimentally manipulates P and (i) does or (ii) does not observe a change in Q. What can she conclude within Pearl's framework? If she knows that she has performed a series of interventions in the sense of **PI** that set P at various levels, then she may conclude that the causal effect of P on Q for these various levels of P is just the distribution of Q under each setting of P, but how does she determine whether or not she has performed such an intervention? The investigator knows that *if* she has changed the value of P in a way that has disrupted the connection between P and its parents but (iii) has not disrupted the mechanism connecting P to Q and (iv) has not disrupted any other mechanism, then she has performed a Pearl-type intervention, but since she knows little or nothing about the mechanism, if any, connecting P to Q (that's what she wants to learn about by doing her experiment), it is unclear how she is to assess (iii) and (iv). Examples like this suggest that if we can learn about causal relationships by conducting experiments at all, there must be some alternative way of characterizing the notion of an intervention that requires less or different prior causal information than that presupposed by **PI**.[7]

[7] It is true enough that we may *sometimes* have good reason to think that a certain kind of change in X would disrupt the causal relationship between X and Y, should any such relationship exist, without knowing whether in fact there is such a relationship. For example, if I see a wire running from a switch to a light bulb and wish to know whether manipulating the position of the switch will cause the light to go on or off, my general background knowledge gives me good reason to believe that a manipulation of the switch that breaks the wire will disrupt any causal connection that might exist between the switch and the light, even if I don't know whether in fact such a connection exists in the unmanipulated system. However, in this case I rely on a very substantial amount of background knowledge about what sorts of mechanisms connect switches and lights and what sorts of changes will disrupt these. The issue, as I see it, is whether (and what) we can learn from experiments when we lack this sort of background knowledge. My view is that we can sometimes learn whether there is a mechanism connecting X to $Y$and about

Consider next a researcher who believes that the activity of some drug causes recovery *R* from an illness – i.e., she believes that there is a causal mechanism connecting *D*, a variable measuring who receives the drug, to *R*. As it happens, the drug itself is ineffective, but the act *A* of administering (or seeming to administer) the drug has a substantial placebo effect on recovery. This is shown by the fact that recovery rates are the same for patients who are given the drug by process *A* as they would be for patients who (unbeknownst to them) are given an inactive substitute for the drug by process *A*. However, what the researcher does is to conduct an experiment *EX. 1* in which she compares the recovery rate in a treatment group for which the drug is administered by process *A* with a control group which receives no drug or substitute by any process. Finding a higher recovery rate in the treatment group, she regards her belief that the drug is efficacious as confirmed. From the researcher's perspective, the experimental manipulation she performs seems to break the causal connection between *D* and its previous causes *C* (who gets the drug is now determined by *A* rather than by *C*) and to leave intact the causal mechanism which she supposes connects *D* to *R*, as well as all other mechanisms (or at least all relevant mechanisms – see below) and thus to count as an intervention in the sense of **PI**. Of course, since we know the correct model, we know that the researcher has not performed an intervention on *D* alone in Pearl's sense **PI** – instead a second variable *A* has also been manipulated in a way that, depending on how the example is modeled, may involve the creation of a new causal link from *A* to *R*.

My concern about **PI** is not that it classifies the experimenter's manipulation as an intervention when it is not, but rather that it provides no basis for (or at least misidentifies the real basis for) distinguishing between those experimental manipulations that, like the one just described, are defective from the point of view of assessing the causal relationships between *D* and *R* and those that are acceptable.[8] One way of seeing this is to compare the experimenter just described with another experimenter who performs a second, ideal experiment *EX. 2* (randomized, double blind, with adequate controls for placebo effects and other confounders etc.) to test the effects of a second drug on recovery and who finds a substantial difference in recovery between the treatment and control group. However ideal, this experimental manipulation will have *some* additional effects besides its effects on *D'* (a variable measuring who receives the second

the sorts of changes that will disrupt it even if we have little prior information about either of these – we learn this precisely by carrying out experimental interventions on *X*. It is not obvious how this is possible if to recognize whether a given change in *X* counts as an intervention we must also know whether the change disrupts any mechanism connecting *X* to its effects.

[8] Thanks to Chris Hitchcock for some very helpful comments on this point.

drug) and will introduce some new causal relationships and/or disrupt some old ones besides those connecting D' to its (previous) parents. For example, this experimental manipulation will also alter the value of the variable A' which describes the manner of administering the drug in this experiment, so that it will not be true that only the value of D' is altered. Administration of this second drug will change the position of air molecules, will result in written records that would not otherwise take the same form, and will involve movement of equipment that disrupts or alters previously existing causal relationships, all in apparent violation of **PI**. In other words, **PI** is violated in both experiments and it thus looks as though **PI** does not capture why the first intervention is unacceptable but the second is acceptable.[9]

Let us call an experimental manipulation of X that is 'ideal' in the sense that it provides good evidence that X causes some second variable Y a *well designed manipulation* and those manipulations that fail to provide such evidence *badly designed*. What the above discussion brings out is that if our interest is in distinguishing between those experimental manipulations of X that are well designed and those that are badly designed, this distinction cannot be based on the idea that the former involve processes that change only the mechanism connecting X to its parents and no other mechanisms, while the latter do not meet this condition. In other words, **PI** does not capture the distinction between well and badly designed manipulations

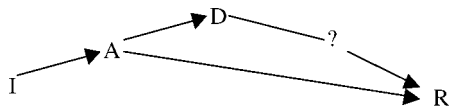
[9] In correspondence, Pearl has drawn my attention to his home page where he discusses a very similar example in response to a question raised by a reader (at http://bayes.cs.ucla.edu/BOOK-2K/do-x.html). The example involves a study in which incubated children at risk for retrolental fibroplasia were injected with vitamin E. It turned out that 'the actual effective treatment was opening the pressurized oxygen-saturated incubators several times a day to give the injections, thus lowering the barometric pressure and oxygen levels in the blood of the infants'. In response Pearl writes,

> In your example of the vitamin E injection (above), there is another variable being manipulated together with X, namely the incubator cover, Z, which turns the experiment into a $do(x, z)$, condition instead of $do(x)$. Thus, the experiment was far from ideal, and far even from the standard experimental protocol, which requires the use of placebo.

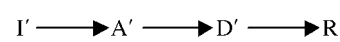
The problem with this response from my point of view is that it does not really give us any insight into what is wrong with the experiment. As argued above, any real life experiment, no matter how ideal, will involve the manipulation of more than one variable. Thus one cannot capture what was defective in the experiment by observing that it involved the manipulation of two variables X and Z rather than just X. What goes wrong in the experiment has to do with the *relationship* between the additional variable Z that is manipulated and the outcome of interest – incidence of retrolental fibroplasia (R). In particular, the problem is not that Z as well as X is changed, but rather that Z affects R independently of X. If the manipulation affected Z as well as X but the only path from Z to R went through X, there would be (at least in this respect) nothing wrong with the experiment.

that we have been describing. For similar reasons, we cannot argue that the manipulation in *EX.* 1 is badly designed on the grounds that it alters some mechanism besides the mechanism connecting *D*, the variable intervened on, to its parents or that it manipulates other variables in addition to *D*. Again, virtually any real-life experimental intervention, however well-designed, will do these things. What makes the manipulation in *EX. 1* badly designed and the manipulation in *EX. 2* well designed has to do not with the fact that additional mechanisms or variables are changed but rather with the *relationships* between the additional changes these manipulations introduce and *R* and *D* (or *D′*).

Intuitively, the problem in *EX. 1* is that *A* affects *R* via a causal route that does not go through *D*. In *EX. 2*, the change in the position of the air molecules does not play a similar role – that is, it does not exert an influence on recovery that is independent of *D′* and it is presumably also not related to *D′* and *R* in some other way that would mislead us about the causal relationship, if there is one, between *D′* and *R*. Similarly, in *EX. 2*, while the experimental manipulation affects *A′* and not just *D*, the structure of the experiment insures that these changes in *A′* affect *R*, if at all, only through the changes *A′* produces in *D′* and not in some other way. That is why, despite the fact that *A′* produces these additional changes, it qualifies as a well designed manipulation. This difference is captured in the following two diagrams.



EX. 1



EX. 2

Can we find a notion of intervention (call it intervention*) that corresponds to the difference between well and badly designed manipulations of the sort described in *EX.* 1 and *EX.* 2, that is more useful for the interpretive and epistemological projects described above, and which, unlike **PI**, does not have the result that few if any real-life experimental manipulations will qualify as interventions? One alternative proposal along these lines which I and others have discussed elsewhere,[10] is that an intervention* on *X* must always be defined with respect to a second variable *Y* – an intervention *I* might be an intervention* on *X* with respect to *Y* but not an intervention* on *X* with respect to some third variable *Z*. This approach

[10] The basic idea of this proposal can be found in Spirtes, Glymour and Scheines (1993 and 2000). They are, however, not responsible for the details of the elaboration (**EI**) offered below. **EI** is discussed in more detail in Woodward (forthcoming) and Woodward and Hitchcock (forthcoming).

also makes it explicit that an intervention must change the value of the variable intervened on for some specific individual or system – i.e., it must make sense to think of the individual as capable of possessing at least two different values of the variable in question. Somewhat more precisely, the idea is

> (**EI**) A necessary and sufficient condition for a process I to count as an intervention* on the value of some variable X possessed by some individual i with respect to some second variable Y is that I changes the value of X possessed by i in such a way that if any change in the value of Y occurs, it only occurs through the change in the value of X and not via some other route. Graphically, this amounts to the requirements that (i) I is the only cause of X – all other arrows into X are broken, (ii) any directed path from the intervention variable I to Y must go through X, (iii) any directed path from any cause Z of I that goes through Y must also go through X, and (iv) 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.

Note that requirements (ii) and (iii) do not specify that there *is* a directed path from X to Y, but only that *if* I or any cause of I affects Y, it must do so through X. (ii) and (iii) thus may be satisfied both in circumstances in which there is and in which there is not a directed path from X to Y. Moreover, it is arguable that one can at least sometimes have good grounds for belief about whether these requirements are or are not satisfied without knowing whether in fact X causes Y or anything very precise about the causal mechanism, if any, connecting the two. Thus, in the second experiment above, the experimenter may not know whether D' affects R (that is why she bothers to do the experiment) but may know on the basis of background knowledge, derived from other experiments, that placebo effects are possible – that is, that administration of the drug is the sort of thing that may affect R independently of D' and hence that she needs to design the experiment in such a way as to exclude this possibility. It is this sort of consideration that the characterization (**EI**) attempts to capture. Unlike (**PI**), (**EI**) gives us some insight into why an investigator should not be concerned if the administration of a drug changes the position of air molecules (assuming these do not affect recovery independently of the activity of the drug) but should be concerned about other causal relationships (such as a possible link between some cause of A and R that is independent of D) that may be created or altered when the drug is administered.[11]

[11] Obviously, it does *not* follow from the fact that some process I is an intervention* on X in the sense of **EI** that it leaves all other mechanisms besides the mechanism connecting X to its parents undisturbed, as **PI** requires. This point is also emphasized in Cartwright (2002). As I explain below, **EI** has the consequence that an intervention* on X may disrupt the relationship between X and its effect Y and still count as an intervention*. In such a

Once we have the notion of an intervention*, characterized along the lines of **EI**, we may use it to provide a characterization of various causal locutions (the causal effect of X on Y, etc.) along lines broadly similar to those recommended by Pearl (only 'broadly similar' because I will suggest characterizations below that differ somewhat in detail from Pearl's).

Since the characterization **EI** obviously makes use of causal information of various sorts, any characterization of causal relationships in terms of (**EI**) will not be reductive in the sense that it explains causal notions in terms of notions that are not themselves causal in character. On the other hand, the causal information that is presupposed when **EI** is used to characterize, for example, the causal effect of X on Y, is not causal information about the very relationship we are trying to characterize – the connection between X and Y – but is instead causal information about the relationship between the intervention* and X or between the intervention* and other causes of Y besides X. In this sense a characterization of what it is for X to cause Y in terms of (**EI**) seems less overtly circular than a characterization in terms of (**PI**).[12]

There is yet another issue concerning what it means to leave a mechanism intact and how this relates to the notion of an intervention that illustrates a difference between **PI** and **EI**. Suppose that within a certain

case, the relationship between X and Y will not be invariant (will not continue to hold) under this intervention*, although, if it is causal at all, the relationship must be invariant under some other intervention*.

[12] I suggested above that it is a point in favor of **EI** and against **PI** that on the latter characterization even ideal real-life experiments will rarely, if ever, qualify as interventions while at least some real life experiments will conform to **EI**. One possible response is that this difference is immaterial. After all, we have already acknowledged, within the general framework of a manipulability or interventionist approach to causation, that we are going to end up talking hypothetically or counterfactually about what would happen if interventions that are technologically impossible for humans to carry out were to occur and that this point will hold regardless of what notion of intervention (whether **PI**, **EI** or some third alternative) we adopt. If so, who cares whether, so to speak, we end up going hypothetical sooner rather than later? That is, given that we will need to talk about what would happen under various interventions that cannot in fact be carried out, what's wrong with talking about what would happen under interventions in the sense of **PI**, even if these are rarely if ever actually realized? I confess to some uncertainty about what the rules of the game are here. Nonetheless, it does seem that if the notion of an intervention is to play a role in the interpretive project – if the point of introducing the notion is to begin with contexts in which suitably controlled experiments can actually be carried out, to use claims about the outcomes of such experiments to get a purchase on what causal claims mean in such contexts, and to then extend this idea to yield an understanding of causal claims in contexts in which the notion of an intervention or experimental manipulation is well-defined but impossible to carry out, then this project is in some tension with the use of a notion of intervention that, like **PI**, is rarely realized even in the most well designed experiment. In other words, the worry is that on a conception of intervention like **PI**, we lose some of the motivation for introducing the notion and its intuitive connection with real-life experiments.

range, $0 < X < r_1$, the restoring force Y exerted by a spring S is linearly related to its extension X: (4) $Y = kX$; that within a range of larger extensions $r_1 < X < r_2$, Y is related to X by some non-linear function (5) $Y = g(X)$; and that (6) for values of $X > r_2$, the spring will break. Suppose that S is initially extended to some length $< r_1$ and then to an extension between r_1 and r_2 via some process that is (otherwise) appropriately exogenous – e.g., it satisfies the conditions in (**EI**) above. If we wish to apply **PI**, should we think of this as a change that 'perturbs' the mechanism connecting Y to X, so that this extension is not an intervention? If the mechanism linking Y to X is simply what is represented by (4), the answer is obviously 'yes'. However, why not think of the mechanism linking Y and X as whatever is described by the conjunction of (4), (5) and a third equation specifying that if the extension ever exceeds r_2, $Y = 0$, regardless of the subsequent value of the extension? This 'mechanism' is not perturbed by extending the spring to lengths greater than r_1 or indeed even by breaking the spring. So relative to this understanding of what the mechanism governing the spring is, even an extension that breaks the spring can count as an intervention, according to **PI**. Indeed, on this conception of what the mechanism governing the spring is, it appears that no manipulation that takes the form of the extension could possibly disrupt the mechanism of the spring. This brings out the extent to which the notion of an intervention is relativized by **PI** to a prior specification of the functional forms governing the mechanisms in the system of interest. The same manipulation may or may not count as an intervention in the sense of **PI**, depending on the level of detail with which the functions governing the mechanisms in that system are specified.

Again it seems to me desirable to have available a notion of intervention that, like (**EI**), is not so relativized and which also does not automatically yield the conclusion that if the equation governing the spring is conceived as (4) then any extension of the spring, no matter how exogenous in other respects, that is out of the range of extensions in which (4) holds does not qualify as an intervention. Unlike **PI**, **EI** allows us to say that an appropriately exogenous extension of the spring beyond the point at which it exerts a linear restoring force may nonetheless qualify as an intervention and hence that (4) holds or is invariant under some interventions and not others. (**EI**) thus allows us to capture the idea that we can learn about the mechanism(s) of the spring and the range of extensions for which it (they) are invariant by conducting experiments. With **EI**, we can still define the total causal effect on Y of $X = x$, as Pearl does, in terms of the response of Y to an intervention* on X, although the details of the characterization will be different (see below). Of course, if we stretch the spring to an extension that breaks it, the causal effect of any subsequent extension on the restoring force will, according to this definition, be null, but this is as it should be.

https://doi.org/10.1017/S0266267103001184 Published online by Cambridge University Press

4

With these remarks as background, let me now turn to Pearl's characterization of causal effect. As explained above, Pearl identifies the total effect of a particular realization x of X on Y with $P(y \mid do(x))$ or as he explains elsewhere, 'the distribution of y while X is held constant at x [presumably by an intervention – J.W.] and all other variables are permitted to run their natural course' (p. 164). While I may not understand this characterization correctly, I find the claim that it gives us the effect of $X = x$ on Y a bit puzzling. Although this is not entirely clear, I assume that Pearl intends his graphs and equations to relate variables, the values or realizations of which apply to individual systems or units in the population of interest. Moreover, as explained above, Pearl's framework is deterministic. Consider in this light a simple linear regression equation $Y = aX + U$. In this case, I take it that Pearl's characterization is meant to suggest that we should identify the total effect of $X = x$ on Y with the distribution of Y when X is set to x and U is allowed to follow whatever distribution it would 'naturally' take. The variation in the value of U results in a spread of values (and a non-degenerate probability distribution for) Y, given by $P(y \mid do(x))$, even for a fixed value of $X = x$. But on the face of things, this is *not* the effect of $X = x$ on Y, since, among other things, it includes, in addition to the contribution made to X by Y, the contribution that U makes to Y. Indeed, what $P(y \mid do(x))$ seems to give us is not even $ax + u$, where u is the particular value taken by U on a particular occasion on which X is set to x, but rather the distribution of Y that results from setting $X = x$ and then adding to this result the contribution to Y made by individual values of U that are generated on other occasions in accord with whatever distribution U takes. If the relationship between X and Y was genuinely indeterministic, then a spread in the value of Y would result for a fixed value of X even if U was held fixed at single value $U = u$ rather than being allowed to vary. But if the relationship is deterministic, it is hard to see why there should be any variation at all in the 'effect' on Y of setting X to a particular value. And whether the relationship is deterministic or indeterministic, it is also hard to see why this effect should include any contribution from U.[13]

[13] There is another closely related issue here that deserves some comment. Recall Pearl's characterization of an intervention (**PI**) – interventions break the relationship between the variable intervened on and its parents while leaving all other mechanisms intact. If Pearl's intention is to identify $P(y/do(x))$ in the above example with the distribution of Y that would result if X were set to x and U allowed to follow its natural or previous distribution, then it seems to me that another additional assumption is being made about the effect of an intervention on X – namely, that (a) it leaves the *probability distribution* of U unchanged or invariant – that is, under the intervention, U continues to follow whatever distribution it would have in the absence of the intervention. This assumption (a) should be distinguished from the even stronger assumption (b) made in clause (iv) of **EI** that an

It seems to me to that the best way to understand the notion of total effect of X on Y is not just in terms of setting X to a particular value but rather in terms of the *difference* made to the value of Y by a specified *change* in the value to which X is set under an intervention, assuming that other causes of Y do not change in value under this intervention.[14] 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 such as U. For example, in the case in which the relationship between X and Y is governed by the deterministic linear relationship $Y = aX + U$, it is natural to think of the total effect on Y of a change in the value of X from $X = x_1$ to $X = x_2$ as $ax_1 - ax_2$. In the more general non-linear case, the change in the value of Y caused by a given change in the value of X may depend on the values of the other causes of Y. Thus the characterization of the notion of causal effect must be relativized to a background context B_i which incorporates information about these other values. In particular, in deterministic contexts, we might 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**) $Y(B_i)_{do(x)} - Y(B_i)_{do(x')}$ – 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 involved in $do(x)$ is now understood in terms of **EI** rather than **PI** – i.e., as an intervention*. In non-deterministic contexts, the causal effect of X on Y may be defined, analogously, as $P(Y(B_i)_{do(x)}) - P(Y(B_i)_{do(x')})$. Note that this definition exploits the fact that clause (iv) in **EI** characterizes the notion of an intervention* in such a way that the values of the other causes of Y in context B_i are left unchanged under an intervention* on X. My claim is that we need some assumption like (iv) about the invariance of the values of other causes of Y under an intervention on X if we are to capture the idea of the causal effect of X on Y.

None of this is to say is to say that the quantity $P(y \mid do(x))$, when $do(x)$ is understood along the lines Pearl favors, is uninteresting or unimportant. Indeed from the perspective of someone who can intervene to set the value

intervention on X leaves the *individual values* of U and of any other exogenous variables unchanged. The importance of these invariance assumptions and the contrast between (a) and (b) is emphasized in (Freedman, forthcoming). It may be that Pearl is treating the invariance assumption (a) as automatically satisfied whenever an intervention in the sense of **PI** occurs but if so, it seems to me that that this should be made explicit. On the face of things, all that **PI** guarantees is that the mechanism governing Y – namely the mechanism described by $Y = aX + U$ – is not disturbed by an intervention on X and a change in the distribution of U need not amount to a change in the mechanism connecting X and U to Y (a change in the distribution of U need not break the arrows from X and U to Y).

[14] The basic idea of defining the effect of X on Y in terms of the difference that changes in the value of X make to the value of Y derives from the framework developed by Donald Rubin and Paul Holland, among others. See Rubin (1974) and Holland (1986).

https://doi.org/10.1017/S0266267103001184 Published online by Cambridge University Press

of X but who cannot affect individual values of U, which can be assumed to follow some fixed distribution regardless of the value to which X is set, and who wants to know what the distribution of Y would be under such interventions on X, it is clearly $P(y \mid do(x))$ rather than the effect of X on Y, which is of most immediate interest. Many problems involving interventions and control have this sort of structure and hence Pearl's investigation of the conditions under which $P(y \mid do(x))$ can be identified is potentially of great practical importance.

REFERENCES

Cartwright, N. 1999. *The dappled world: A study in the boundaries of science.* Cambridge University Press

Cartwright, N. 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* 53:411–453

Cartwight, N. 2003. Two theorems on invariance and causality. *Philosophy of Science* 70:203–24

Freedman, D. Forthcoming. On specifying graphical models for causation, and the identification problem. Technical Report No. 601, Department of Statistics, University of California at Berkeley

Glymour, C. 1999. Rabbit hunting. *Synthese* 121 (1–2):55–78

Hausman, D. and J. Woodward. 1999. Independence, invariance, and the causal Markov condition. *British Journal for the Philosophy of Science* 50:1–63

Hausman, D. and J. Woodward. Forthcoming. Causality and the causal Markov condition: A restatement. *British Journal for the Philosophy of Science*

Holland, P. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81:945–970

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

Pearl, J. 2000. *Causality.* Cambridge University Press

Rubin, D. 1974. Estimating causal effects of treatment in randomized and non-randomized studies. *Journal of Educational Psychology* 66:688–701

Spirtes, P., C. Glymour, and R. Scheines. 2000. *Causation, prediction, and search.* MIT Press

Woodward, J. and C. Hitchcock. Forthcoming. Explanatory generalizations: A counterfactual account. *Nous*

Woodward, J. 2001. Probabilistic causality, direct causes, and counterfactual dependence. In *Stochastic dependence and causality*, ed. D. Constantini, M. C. Galavotti, and P. Suppes, 39–63. CSLI Publications

Woodward, J. Forthcoming. *Making things happen: A theory of causal explanation.* Oxford University Press

von Wright, G. 1971. *Explanation and understanding.* Cornell University Press