References

Arntzenius, F. (1990) Physics and common causes. *Synthese*, 82

Beebee, H. and Papineau, D. (1997) Probability as a guide to life. *Journal of Philosophy*, 94, 217-243.

Eells, E. (1991) *Probabilistic Causality*. Cambridge: Cambridge University Press.

Hausman, D. (1998) *Causal Asymmetries*. Cambridge: Cambridge University Press.

Irzik, G. (1996) Can causes be reduced to correlations? *British Journal for the Philosophy of Science*, 47, 249-270.

Papineau, D. (1985a) Probabilities and causes. *Journal of Philosophy*, 82.

Papineau, D. (1985b) Causal asymmetry. *British Journal for the Philosophy of Science*, 36, 273-289.

Papineau, D. (1989) Pure, mixed and spurious probabilities and their significance for a reductionist theory of causation. In P. Kitcher and W. Salmon (Eds.) *Scientific Explanation: Minnesota Studies in the Philosophy of Science vol XIII*. Minneapolis: University of Minnesota Press, 307-348.

Papineau, D. (1993a) Can we reduce causal direction to probabilities? In D. Hull, M. Forbes and K. Okruhlik (Eds.), PSA 1992 vol 2. East Lansing: Philosophy of Science Association, 238-252.

Papineau, D. (1993b) The virtues of randomization. *British Journal for the Philosophy of Science*, 44, 437-450.

Papineau, D. (forthcoming) Causation as a guide to life.

Pearl, J. and Verma, T. (1994) A theory of inferred causation. In D. Prawitz, B. Skyrms and D. Westerståhl (Eds), *Logic, Methodology and Philosophy of Science IX* Amsterdam: Elsevier, 789-811.

Price, H. (1996) *Time's Arrow and Archimedes' Point*. Oxford: Oxford University Press

Reichenbach, H. (1956) *The Direction of Time*. Berkeley: University of California Press.

Spirtes, P, Glymour, C. and Scheines, R. (1993) *Causation, Prediction and Search*. New York: Springer-Verlag.

Suppes, P. (1970) *A Probabilistic Theory of Causality*. Amsterdam: North Holland.

3

Probabilistic Causality, Direct Causes and Counterfactual Dependence*

JAMES WOODWARD

California Institute of Technology

1 Introduction

A great deal of recent philosophical work on causation largely falls in two major traditions. On the one hand, there is the tradition of probabilistic theories of causality inaugurated by Suppes (1970). Suppes hoped to reduce causal claims to claims about probabilities. More recent work in this tradition eschews the goal of a complete reduction but still hopes to find systematic relationships between causal claims and claims about probabilities. For example, a standard suggestion (Cartwright 1983) about that connection is that causes must raise the probability of their effects across all causally homogenous background contexts or given all possible combinations of other factors that are causally relevant to the effect. Theories of this sort are generally intended as theories of so-called type causation: that is they are intended to capture causal claims that relate types of events or properties such as "impacts of rocks cause windows to break" and "smoking causes lung cancer". With a few exceptions (e. g., Eells 1991), they are generally not intended to be accounts of so-called token causation—that is, causal claims that relate

* Thanks to Chris Hitchcock for a number of helpful discussions.

Stochastic Causality.
Maria Carla Galavotti, Patrick Suppes and Domenico Costantini (eds.).

individual events such as "the impact of the ball thrown by Billy on January 12, 2000 at 3pm caused Smith's window to shatter".

One of the main alternatives to such theories in philosophy is the counterfactual approach to causation developed by David Lewis (Lewis, 1973, 1979) and his students. This approach is also reductionist in intent, but in contrast to probabilistic theories, the aim is to reduce causal claims to claims about counterfactual dependence, where the latter can be understood in a way that does not presuppose causal ideas. Moreover, in contrast to probabilistic theories, Lewis's theory is not intended as an account of type causation but rather as an account of token causation. Lewis's theory does succeed in capturing our common sensical judgments about token causal relationships in a range of (although by no means all) cases. However the theory requires a semantics for counterfactuals that is prima-facie quite mysterious—for example, we are often required to employ counterfactuals the antecedents of which are made true by miracles. Many writers have argued that such counterfactuals play no legitimate role in scientific practice.

What is the relationship between these philosophical theories and the treatments of causal claims one finds in discussions of experimental design and in disciplines like econometrics and epidemiology—treatments that employ, for example, the apparatus of structural equations and/or directed graphs? These disciplines differ along many different dimensions and there is no single generally accepted label for the work on causal inference and the representation of causal relationships one finds in them. To avoid cumbersome repetition, I will call them the causal modeling disciplines and will speak (in a way that obviously involves considerable idealization) of "the" causal modeling conception of causation. Causal inference in the causal modeling disciplines is usually based, at least in part, on statistical evidence and, in part for this reason, a great deal of work in these disciplines focuses on the relationship between causal claims and probabilistic relationships. This fact has led many philosophers to suppose that the causal modeling notion of causation must be something like the notion that probabilistic treatments of causation attempt to capture. Indeed, the assumption that this is the case has been one of the main motivations for the development of probabilistic theories.

In this essay, I will compare the treatment of causation assumed in the causal modeling disciplines with the probabilistic and counterfactual theories developed by philosophers. I will suggest that, contrary to what many philosophers have supposed, the causal modeling notion is in many respects closer to the counterfactual approach assumed in the Lewisian tradition than to probabilistic accounts of causality. In fact, the causal modeling treatment of causation clarifies and explains some of the puzzling features of Lewis's account of counterfactuals. Moreover, the causal modeling treatment yields a

notion of causation that, provided one is willing to make certain assumptions, does have systematic connections with facts about probabilities. However, these connections are considerably looser than those defended in standard formulations of probabilistic theories—so loose that the phrase "probabilistic theory of causation" seems a misnomer. In general, the causal modeling conception of causation brings together work in both the counterfactual and probabilistic causality traditions, capturing what is plausible in both.

As a point of departure, let me begin with an observation whose significance is insufficiently appreciated by many philosophers working in the probabilistic causality tradition. This is that the standard treatments of causal relationships one finds in the causal modeling disciplines employs two distinct kinds of resources or representations, both of which work together in problems of causal inference. First, as in probabilistic theories of causation it is assumed that we have a probability distribution **P** over some set of variables V whose causal relationships one is interested in investigating. Second, over and above this, one makes use of some additional device **G** to represent causal relationships among the variables in **V**. **G** is typically either system of equations or a directed graph. We may thus think of a causal model as an ordered triple of the form **<V, P, G>**. As we will see, if we are to adequately represent the connection between casual relationships and probabilities, *both* **P** and **G** are required—neither cannot be reduced to or replaced by the other. Roughly speaking the role of the directed graphs or structural equations is to represent information about patterns of counterfactual dependence among variables; more specifically, it is to tells us what would happen to the values of some variables under changes of a special sort involving what I will call *interventions*. (These are idealized experimental manipulations—see Section 4) in the values of other variables. The probability distribution **P**, by contrast, does not convey modal or counterfactual information of this sort. Instead, it conveys information about the actual distribution of values of variables. As we will see below, we may think of directed graphs and structural equations not as summarizing information about any particular probability distribution but rather as telling us various different distributions are connected to one another—in particular how such distributions (or certain features of them) will change under interventions or combinations of interventions. Seen from this perspective, a major problem with the probabilistic theories found in much of the philosophical literature is that they attempt to provide an account of causation by relying too heavily on just one kind of resource—the probability distribution **P**. In fact the full resources of the graphical or equational representation are required if one is to do justice to the notion of causation.

2 Directed Graphs and Equations

I begin with some brief remarks about the use of directed graphs and systems of equations to represent causal relationships. Let us assume that the causal claims that we are interested in modeling relate variables where variables represent properties or magnitudes that (as the name implies) are capable of taking more than one value[1]. Such claims will involve type-level rather than token level relationships. The familiar examples of so-called property or type causation discussed in the philosophical literature may be understood as relationships between two-valued or binary variables, with the variables in question taking one or another of two values, depending on whether the properties in question are instantiated or occur. Thus the claim that ingestion of aspirin causes recovery from headache may be understood as asserting a relationship of some kind between a variable A, representing whether or not aspirin is ingested, and a variable H representing whether or not relief from headache occurs. Of course variables need not be two-valued; they may also assume many values or be continuous.

A directed graph is an ordered pair <**V**, **E**> where **V** is a set of vertices which serve as the variables representing the *relata* of casual relationships and **E** a set of directed edges connecting these vertices. A directed edge from vertex or variable X to vertex or variable Y means that X *directly causes* Y. For now I will largely rely on the reader's intuitive understanding of this notion; it is discussed in more detail in Section 5 below. However, the basic idea is that X is a direct cause of Y if and only if there is a possible intervention (experimental manipulation) on X that would change the value of Y (or the probability distribution of Y) when all other variables in the system of interest are held fixed at some set of values in a way that is independent of the change in X. Put more simply, drawing an arrow from X to Y means that there is *some* change in the value of X that will change the value of Y (or the probability distribution of Y), given *some* set of values for the other variables. I assume that if X is a direct cause of Y, then X is a cause of Y, but that the converse of this claim is false. A sequence of variables $\{V_1, \ldots V_n\}$ is a *directed path* or *route* from V_1 to V_n if and only if for all i ($1 \leq i \leq n$) there is a directed edge from V_i to V_{i+1}. Y is a *descendant* of X if and only if there is a directed path from X to Y. If Y is a descendant of X, then X is an *ancestor* of Y. The direct causes of X are also said to be the *parents* of X[2]. As we will see below,

[1] To avoid cumbersome terminology, I will often use the word "variable" to refer both to the properties etc. that serve as *relata* in causal relationships and to the symbols that represent such properties. This conflation is, I believe, harmless.

[2] See Spirtes, Glymour and Scheines, 1993 for a more detailed discussion of the use of directed graphs to represent causal relationships.

a necessary but not a sufficient condition for X to be a cause of Y is that X be an ancestor of Y.

As an illustration of the use of directed graphs to represent causal relationships, suppose that A is a variable measuring atmospheric pressure, B a variable representing the reading of a particular thermometer and S a variable representing the occurrence or non-occurrence of a storm Then we may represent the claim that A is a direct common cause of B and S and that B does not cause S and S does not cause B by means of the following diagram.

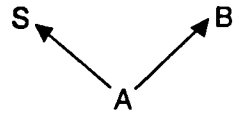


Figure 2.1

Causal relationships may also be represented by means of systems of equations. When underlying causal relationships are deterministic each endogenous variable Y (that is, each variable that represents an effect) is written as a function of all and only those variables that are its known or measured direct causes plus a so-called error term which represents the combined influence of all of the other direct causes of the endogenous variable. The presence of the error term makes possible conditional probabilities involving the measured variables that are strictly between zero and one. There is one equation for each endogenous variable. For example, if variables $X_1, \ldots, X_m$ are all of the known direct causes of Y, then Y may be written as (1.1) $Y = F_Y(X_1,\ldots,X_m) + U$. Analogous remarks apply to the indeterministic case, with the relevant equations specifying how the probability distribution of Y will change under manipulation of the right side variables representing direct causes in each equation.

What is the relationship between the representation of causal relationships by means of systems of equations and their representation by means of directed graphs? As we have seen, when we draw a directed graph with arrows from $X_1 \ldots, X_m$ into Y, we convey the information that Y is some function of $X_1 \ldots X_m$ and that all of the variables $X_1, \ldots, X_m$ are essential in the sense that for each such variable X_i , there is some combination of values of the others such that changing X_i will change Y. However, the graph does not further specify what this function is. It does not tell us *which* changes in X will change Y or by how much or for which values of other variables. By contrast, when we explicitly specify the function or equation relating Y to its direct causes (e. g. $Y = 3X_1 + 4X_2$), we convey more information than if we merely draw a graph with arrows from X_1 and X_2 directed into Y. Unlike the

directed graph, the explicit form of the equation specifies exactly how changing X_1 and X_2 will change Y. By contrast, the corresponding directed graph says simply that there is some change in X_1 (X_2) that, given some value of X_2 (X_1), will change Y.

3 Causation, Manipulation and Counterfactuals

I said above that directed graphs and systems of equations represent counterfactual claims about how changing the values of certain variables will change the values (or the probability distribution of the values) of others. Let me now try to be more precise about this idea and its underlying motivation. The notion of causation assumed in directed graphs and systems of equations is a *manipulability* conception of causation. The underlying idea is that causal relationships are just those relationships that are potentially usable for purposes of manipulation and control in the sense that if X is a cause of Y then if it were possible to change or wiggle the value of X in the right way and in the right circumstances, this would be a way of wiggling or changing the value of Y. Manipulability theories are thus a subspecies of counterfactual theories of causation; they are theories according to which the right counterfactuals for understanding causal claims are counterfactuals that have to do with what would happen under hypothetical manipulations.

Manipulability accounts of causation have been unpopular in contemporary philosophy; they are commonly criticized as both unilluminatingly circular and as leading to an unacceptably anthropomorphic or subjective notion of causation in the sense that they seem to restrict true or meaningful causal claims to those contexts in which manipulation by human being is possible (see, e.g., Hausman 1998). I have argued elsewhere (Woodward forthcoming) that while these criticisms are indeed apt when applied to the standard formulations of the manipulability theory one finds in the philosophical literature (such as von Wright 1971; Price 1991), there is natural way of developing an alternative version of manipulability theory that avoids such criticisms. As we will see (Section 4), the key to formulating an acceptable version of the manipuability theory is finding the right characterization of the notion of an intervention: an intervention on X with respect to Y can be characterized in a way that makes no reference to human beings or their activities (thus avoiding the anthropocentrism of traditional versions of the manipulability theory) and also in a way that makes no reference to the existence or non-existence of a causal relationship between X and Y (thus avoiding the vicious circularity that infects traditional versions).

But while it is possible in this way to formulate a version of the manipulability theory that avoids the standard criticisms, a natural (and deeper question) is why we should bother to do this. Why suppose that we can clarify or

explain anything about causal relationships by thinking about them within the framework of a manipulability theory? For reasons of space, I will confine myself to a few brief observations. First, researchers within the causal modeling disciplines tell us again and again that what they *mean* by causal relationships are those relationships that are exploitable for purposes of manipulation and control[3]. While it is of course possible that such pronouncements have no relation to the conception of causation assumed in the actual practice of these disciplines, this is a good prima-facie reason for taking the manipulability conception seriously.

Second, manipulability theories of causation provide a natural and attractive account of the underlying point or rationale of our practice of distinguishing between causal and non-causal relationships: if X and Y are correlated and if manipulation of X is possible, there are obvious practical advantages to knowing whether or not the relationship between X and Y is such that manipulating X can change Y. Moreover, at least in many contexts, it also seems clear that it is exactly this distinction that is at issue when we worry about whether a relationship is causal. Consider the well documented correlation between superior scholastic performance and attendance at private schools. Does this reflect a (3.1) causal connection between such attendance and performance or (3.2) is it rather the case that private school attendance per se has no effect on performance and that the correlation arises entirely from the fact that the very factors that lead to enrollment in private schools (e.g., affluent parents who are concerned about their children's education) also cause superior performance? Parents and educational researchers care about the answer to these questions exactly because they want to know whether they can manipulate performance by enrolling students in private schools—it is possible in principle to do this if (3.1) is correct but not if (3.2) is. More generally, human beings are often in the position of observing a correlation between X and Y and wondering whether this correlation reflects a relationship that will allow them to change Y by manipulating X or whether instead the observed correlation between X and Y will disappear under manipulation of X. According to a manipulability theory our notion of causation developed not as the result of an impulse to engage in dubious metaphysics or to project certain of our psychological states onto the world but rather to mark this practically important distinction. On this view, directed graphs and

[3] Illustrations are readily found in a variety of texts on experimental design and econometrics. A representative quotation from Cook and Campbell's highly influential (1979) is: *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 (1979, p. 36, emphasis in original).

systems of equations have a similar motivation; they reflect our concern to distinguish manipulation supporting relationships from mere correlations.

Like the notion of causation itself, counterfactuals have often been regarded with suspicion by empiricists. It is frequently suggested that they lack a clear meaning or that their truth conditions are so vague and context-dependent that they are not suitable for understanding or elucidating any notion (of causation or anything else) that might be of scientific interest. A famous example of Quine's illustrates the worry. Consider the counterfactual(s) (3.3) "If Julius Caesar had been in charge of U. N. forces during the Korean war, then he would have used (a) nuclear weapons or (b) catapults". It is hard to see on what basis one could decide whether the counterfactual (3.3) with (a) as consequent or the counterfactual (3.3) with (b) as consequent (or neither) is correct. A manipulability framework for understanding causation helps to address this worry. It suggests that the appropriate counterfactuals for elucidating causal claims are not just any counterfactuals but rather counterfactuals of a very special sort: those that have to do with the outcomes of hypothetical manipulations or experiments. It does seem plausible that counterfactuals that we do not know how to interpret as (or associate with) claims about the outcomes of well-defined manipulations will often be claims that lack a clear meaning or truth value. For example, (3.3a) and (3.3b) seem unclear for just this reason. It isn't just that we lack the technological means to carry out an experimental manipulation in which Caesar is placed in charge of the U. N. forces. The more fundamental problem is that we have no clear conception of what would be involved in carrying out such an experiment.

By contrast, a similar sort of skepticism about counterfactuals that are interpretable as claims about the outcomes of hypothetical (but otherwise well specified) experimental manipulations is much harder to sustain. Consider an experiment in which a large group of people suffering from a disease are randomly divided into a treatment and a control group with the former receiving some drug that is withheld from the latter. As it turns out, the incidence of recovery is much higher in the former than in the latter. Provided that the right sort of experimental controls have been followed, it is very natural to think of this experiment as providing good evidence for the truth of counterfactuals like the following: (3.4) "If those in the control group had received the drug, the incidence (or expected incidence) of recovery in that group would have been much higher." Indeed it is very plausible that it is precisely because the experimenters want to determine the truth value of counterfactuals like (3.4) that they conduct the experiment. Of course, the researchers may be mistaken in the conclusion they draw about the truth value of (3.4) but this does not distinguish (3.4) from any other empirical knowledge claim. The claims that (3.4) lacks a determinate meaning or truth value or is untestable in principle are, I suggest, much less plausible than the corresponding claims about (3.3)[4].

In contrast to counterfactuals like (3.3), counterfactuals like (3.4) play a central role both in practical deliberation and in experimental practice in science. We need to understand how such counterfactuals can be tested and evaluated but they should not be dismissed as meaningless or unscientific.

4 Interventions

I suggested above that one of the key elements in formulating a defensible version of a manipulability theory is the notion of an intervention. Heuristically (but only heuristically), one may think of an intervention on X with respect to Y as the sort of manipulation that might be carried out in an ideal experiment for the purpose of determining whether X causes Y. The basic idea may be illustrated by reference to the *ABS* system in Figure 2.1. It is clear that there are ways of changing B that will be associated with a corresponding change in S even though B does not cause S. For example, if we change B by changing A, or by means of some causal process that is perfectly correlated with changes in A, then S will also change, but this would not establish that B causes S. Plainly, an experiment in which B is manipulated in this way is a badly designed experiment for the purposes of determining

4 These remarks raise a natural question. Suppose that we grant that counterfactuals like (3.4) that can be tested experimentally have truth values and hence that the causal claims associated with them have truth values as well. What about causal claims and associated counterfactuals for which the relevant experimental manipulations are specifiable or well-defined but cannot actually be carried out, because of technological or other sorts of limitations? For example, consider the causal claim that (3.5) the position of the moon causally influences the tides and the associated counterfactual claim that (3.6) if the radius of the moon's orbit were to be changed as a result of an intervention, the motions of the tides would have been different. Assuming that the causal claim (3.5) is true, on a manipulability theory some associated counterfactual like (3.6) must be true as well. But why suppose that (3.6) has a definite truth value if, as is clearly the case, the associated manipulation cannot actually be carried out? While the matter deserves a more detailed discussion than I can give it here, the short answer is that once it is accepted that (a) counterfactuals have truth values when their antecedents refer to experiments that can be carried out, it is hard to avoid the view that (b) counterfactuals for which the associated experiments are well-defined but cannot be carried out also have truth values. The reason for this is that even for counterfactuals satisfying (a), it is not the actual carrying out of the associated experimental manipulations that endows them with definite truth values. Rather, such counterfactuals possess definite truth values independently of whether the relevant experimental manipulations are carried out. The experimental manipulations are a way of discovering what the truth values of the counterfactuals are; they do not somehow create those truth values. Similarly for counterfactuals satisfying (b)—if the manipulations specified in their antecedents cannot, as a practical matter, be carried out, this shows only that their truth or falsity cannot (at present) be directly determined by experimentation, not that they lack truth values.

whether B causes S. Similarly, an experiment in which the process that changes B also directly changes S would be badly designed for this purpose.

By contrast, consider the following experiment. We employ a random number generator which is causally independent of A and, depending just on the output of this device, repeatedly physically fix the barometer reading at different values by moving the dial to either a high or low reading and driving a nail through it. If it is really true that B does not cause S, then we expect that the changes in B produced by such interventions will no longer be associated with changes in S. If, on the contrary, S continues to be correlated with S under such interventions on B, this would be strong prima-facie evidence that B does cause S. In contrast to the previous experiments, an experiment of this sort would be a well designed experiment for the purposes of determining whether B causes S. The notion of an intervention is meant to capture the contrast between these two kinds of experiments: the second sort of experiment involves an intervention on B with respect to S while the first does not.

This reference to an "ideal experiment" naturally suggests an activity carried out by human beings. However, as I suggested above, the notion of an intervention can be given a completely nonanthropomorphic characterization, that makes no reference to human beings or their activities. I will not try to present the full details of this characterization here, but instead refer the reader to the characterization in Woodward (2000)[5]. Informally, however, we may think of an intervention I on X with respect to Y as an exogenous causal process that changes X in such a way and under conditions such that if any change occurs in Y, it occurs only in virtue of Ys relationship to X and not in any other way. Making such a characterization precise requires reference to the causal relationships between I and various other causes of Y and to the causal relationship between I and Y (for example, I must not be a direct cause of Y) but it does not require reference to the presence or absence of a causal relationship between X and Y. Thus, such a characterization will not be viciously circular in the sense that to know whether an intervention has been carried out on X with respect to Y, one must already know whether X causes Y.

The sense in which interventions involve *exogenous* changes in the variable intervened on is illustrated by the above example. When an intervention occurs on B, the value of B is determined entirely by the intervention, in a way that is (causally and probabilistically) independent of the value of A which was what previously determined the value of B. In this sense the intervention breaks or disrupts the previously existing endogenous causal relationship between A and B. If we represent such an intervention on B by

[5] For additional and broadly similar characterizations of the notion of an intervention see Spirtes, Glymour and Scheines (1993), Woodward (1997), Hausman (1998), and Pearl (2000).

drawing an arrow from an intervention variable I to B, then the result of the intervention will be to replace the structure in Figure 2.1 with the structure in Figure 4.1:

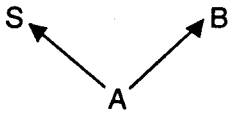


Figure 2.1

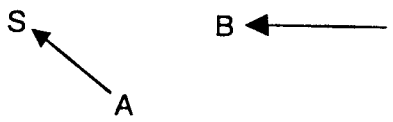


Figure 4.1

This illustrates the so-called "arrow-breaking" conception of interventions: an intervention on X breaks all arrows directed into X while preserving all other arrows in the graph, including those directed out of X. The breaking of arrows directed into X captures the idea that the value of X is now determined exogenously, entirely by the intervention variable I and that those variables that influenced the value of X prior to the intervention no longer do so[6].

This idea about how to represent interventions graphically is closely tied to an idea about the impact of interventions on systems of equations that Dan Hausman and I (Hausman and Woodward 1999) have elsewhere called "modularity". We may represent the structure in Figure 2.1 by means of the following two equations

(4.1) $B = aA$

(4.2) $S = bA$

An intervention on B will then correspond to replacing equation (4.1) with a different equation (4.3), $B = I$, specifying that the value of B is no longer determined by S but is instead set entirely by the value of the intervention variable I. Just as we assume, when we employ the arrow-breaking conception of interventions, that it is possible to carry out an intervention on B that leaves the arrow from A to S undisturbed, we also assume that when a system of equations like (4.1- 4.2) correctly represents some causal structure, it will be possible to carry out this operation of replacing one equation in the system (in this case, (4.1) with (4.3)) while leaving the other equations in the system (4.2) undisturbed. When a system of equations has this feature (that is, when one may disrupt or replace any one of the equations by means of an intervention on the dependent variable in that equation, without disrupting the other equations), I will say that the system is *modular* or equation-invariant. Within a probabilistic framework, modularity corresponds to the

[6] For additional discussion of the arrow-breaking interpretation, see Pearl (1995, 2000), Spirtes, Glymour and Scheines (1993) and Hausman and Woodward (1999).

following requirement: $Pr(X/Parents(X).Set\ Y) = Pr(X/Parents\ (X)$ for all Y distinct from X, where Set Y means that the value of Y has been set by an intervention.

One natural way of motivating the idea that systems of equations should be modular appeals to the idea that if a system of equations correctly represents causal structure, each equation in the system should represent the operation of a distinct causal mechanism. If we make the additional plausible assumption that a necessary condition for two mechanisms to be distinct is that it be possible (in principle) to interfere with the operation of one without interfering with the operation of the other and vice-versa , we have a justification for requiring that systems of equations that correctly represent casual structure should be modular. For example, a natural justification for supposing that we may replace (4.1) with (4.3) without altering (4.2) is that the mechanism by which A affects B is distinct from the mechanism by which A affects S. In what follows, I will assume that the systems of equations with which we are dealing are modular (and correlatively that graphical representations satisfy the arrow-breaking interpretation of interventions). For example, the definition of direct causation (**DC**) given below assumes modularity.

It also will be important to our subsequent discussion to understand that *intervening* to set the value of a variable to some value is conceptually quite different from *conditioning* on the value of that variable. (cf. Meek and Glymour 1994; Pearl 2000). Following a proposal due to Pearl (1995) let us suppose that the values of variable X that are set by interventions can be represented as the values of a new random variable, *set X*. (This will be a reasonable assumption when, as in the example above, the values of this variable are determined by a randomizing device). Then it will not in general be true that $Pr(Y/X) = Pr(Y/set\ X)$. In the above example, S and B are correlated and in fact $Pr(S/B) > Pr(S)$. However, assuming that B does not cause S, we would not expect S and ***set B*** to be correlated. Instead we would expect that $Pr(S/set\ B) = Pr(S)$.

The reason why intervening is different from conditioning is unmysterious. When we condition on a variable, we assume that whatever causal structure generates the values of that variable is left intact, so that the values in question continue to be generated by whatever endogenous causal factors are at work in that system. Thus when we condition on B in the above example, we assume that the values of B continue to be generated by A, in which case they will be correlated with the values of S, which are also generated by A. By contrast, as the above example illustrates, if a variable is endogenous, then intervening on it alters the causal structure of the system in which it figures—giving it a new exogenous causal history. Unless the variable intervened on is exogenous, the intervention will disrupt the previously existing pattern of correlations in the system, leading to a new set of probability relationships. We see this in the example under discussion, in which B and S are correlated when there is no intervention on B, but are uncorrelated in the new structure that results when there is an intervention on B. (The graph in Figure 2.1 thus tells us how the distribution of B and S will change under an intervention on B.) The difference between conditioning and intervening thus corresponds to the difference between two questions: What would it be reasonable for me to predict regarding the value of S when I observe the value of B, assuming that no intervention occurs and whatever system has been generating the values of B and S remains intact? What would it be reasonable for me to predict regarding the value of S, if I were to physically manipulate the value of B in the manner described above?

To say that there is an important difference between conditioning and intervening is not of course to say that there are no systematic connections between these two notions. Indeed to a very large extent, *the* problem of causal inference, at least in non-experimental contexts, is when (and how) it is possible to infer from information about conditional probability relationships to claims about what would happen under possible interventions—an issue to which I will turn in Section 6 below. However, from the perspective of a manipulablity theory, the structure of this problem is fundamentally obscured if we do not distinguish between conditioning and intervening.

It should also be clear from the above characterization that interventions function in broadly the same way as Lewisian miracles. When we consider, within Lewis' framework, a counterfactual like (4.4) "If the barometer reading had been low, a storm would have occurred" we imagine that the antecedent of this counterfactual is made true by the insertion of a small, localized miracle which decouples the value of the barometer reading from the value of the atmospheric pressure. This miracle makes B independent of S and this in turn prevents the sort of backtracking reasoning that might be used to argue for the truth of (4.4): (If the reading was low, that must be because A was low, in which case the storm would have occurred). Like an intervention on B, the insertion of such a miracle gives B an independent causal history and (at least in many cases) this is enough to insure that any change in the value of S is due to the value of B and hence that B is a cause of S.[7] Of course real-life interventions need not literally involve miracles, but we may think of this language as a picturesque way of expressing the idea that an intervention involves a change that comes into the system from the outside and disrupts

[7] As the phrase in parentheses suggests, while Lewisian miracles often function in the same way as interventions, they do not always do so. See Woodward (forthcoming) for discussion. One important difference between Lewis' theory and the interventionist approach is that the latter assigns an important role to counterfactuals concerning what will happen under combinations of interventions. These have no direct counterpart in Lewis' theory.

endogenous causal relationships. I thus suggest that Lewis' framework works as well as it does because it tracks the same sorts of relationships that are picked out by a manipulationist or intervention-based theory. In other words, Lewis' "unnatural" semantics for counterfactuals makes good scientific sense when motivated by ideas about the connection between causation and manipulation[8].

5 Direct Causes

I said above that directed graphs and systems of equations convey information about direct causal relationships. Given the notion of an intervention, the notion of a direct cause can be understood in manipulationist terms as follows:

> (**DC**) A necessary and sufficient condition for X to be a direct cause of Y with respect to some variable set Z is that there be a possible intervention on X that will change Y (or the probability distribution of Y) when all other variables in Z besides X and Y are also held fixed at some value by interventions.

Intervening to hold the variables in Z fixed at some values while changing X by means of an intervention means that the variables in Z are set to those values by a process that satisfies the conditions for an intervention while X is changed by some other process, also satisfying the conditions for an intervention, that is causally independent of and uncorrelated with the process that changes X.

As an illustration consider the following pairs of equations and corresponding graphical structures (error terms have been suppressed for expository convenience).

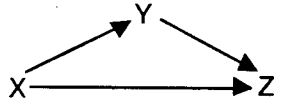


Figure 5.1

(5.1) $Y = aX$
(5.2) $Z = bX+cY$

Figure 5.2

(5.3) $Y= aX$
(5.4) $Z = dY$

[8] For similar observations, see Pearl (2000).

According to both (5.1-5.2) and (5.3-5.4), an intervention on X will change Y, an intervention on X will change Z and an intervention on Y will change Z. Nonetheless, (5.1-5.2) and (5.3-5.4) are associated with different graphical structures and make different claims about direct causal relationships. According to (5.1-5.2) X affects Z by two different routes, a direct route and an indirect route that goes through Y. By contrast, according to (5.3-5.4), X affects Z only by a single route that goes through Y. (5.1-5.2) claims that X is a direct cause of Z while (5.3-5.4) denies this. The definition **DC** captures this difference. According to (5.1-5.2), if we intervene to fix the value of Y and then intervene to change the value of X, the value of Z will change—hence X is a direct cause of Z[9]. By contrast, if (5.3-5.4) is correct, then, if we fix the value of Y (at any value), no intervention on X will change Y—hence X is not a direct cause of Y. In other words, while (5.1-5.2) and (5.3-5.4) agree about what will happen to Z under single interventions on either X or Y, they differ in what they predict about what will happen under *combinations* of interventions. In particular, we could determine whether (5.1-5.2) or (5.3-5.4) is the correct structure by doing an experiment in which the value of X is changed and the value of Z observed while the value of Y is held fixed.

DC requires that "all other variables in Z besides X and Y are held fixed at *some* value by interventions". This formulation is needed because some causal relationships are non-linear. Since the relationships in (5.1-5.2) and (5.3-5.4) are linear, the effect of a change in the value of X on Y when Z is fixed will be the same, regardless of the value at which Z is fixed. When relationships are non-linear, this will not be the case. Suppose that F is a variable that takes the values 0 or 1, depending on whether a fire occurs, S is a variable taking the values 0 or 1 depending on whether a short circuit occurs, and O is a variable that takes the values 0 or 1 depending on whether oxygen is present. Suppose that the causal relationship between these variables may be represented by means of the equation

(5.5) $F = S.O$

That is, a fire will occur when and only when both the short circuit and oxygen are present. If $O = 0$, an intervention that changes the value of S will not change the value of F. The formulation of **DC** nonetheless allows S to qualify as a direct cause of F; S is a direct cause of F because there is some value of O (namely $O = 1$) such that with O fixed at that value, a change in S will change F.

Plainly not all causes are direct causes. What is the connection between causation and direct causation? Consider the following candidates for neces-

[9] Obviously, this reasoning assumes that the system (5.1-5.2) is modular and in particular that fixing the value of Y in (5.1) does not change (5.2).

sary and sufficient conditions for X to be a cause of Y, which I label (**SC**) and (**NC**) respectively.

> (**SC**) If there is a possible intervention that changes the value of X such that carrying out this intervention will change the value of Y, or the probability distribution of Y, then X causes Y.
>
> (**NC**) If there are no possible interventions that can change the value of X, or if for all possible interventions that change the value of X, the value of Y (or the probability distribution of Y) does not change, then X does not cause Y.

I believe that (**SC**) is extremely plausible; indeed one may take it to be one of the core commitments of a manipulability theory of causation. By contrast, on the supposition that direct causes are causes, **NC** conflicts with **DC**. To see this consider, a structure like (5.1-5.2) with $b = -ac$. In this structure, the influence of X on Z along the direct and indirect routes will "cancel". In such a structure X is a direct cause of Y according to **DC** and hence (we are supposing) a cause. Nonetheless, there are no interventions on X that will change Y and hence according to **NC**, X fails to cause Y.

This example shows that we need to distinguish between two notions of "cause"[10]. Let us say that X is a *total cause* of Y if and only if it has a non-null total effect on Y—that is, if and only if there is some intervention on X alone (and no other variables) such that for some value of those other variables, this intervention on X will change Y. The *total effect* of a change dx in X on Y is then the change in the value of Y that would result from an intervention on X alone that changes it by amount dx (given the values of other variables that are not descendants of X.) For example, in (5.1-5.2) the total effect on Z of a change of dx in X is $(b+ac)\,dx$ and the total effect on Z of a change dy in Y is cdy. Let us say that X is a *contributing cause* of Y if and only if it makes a non-null contribution to Y along some directed path in the sense that there is some set of values of variables that are *not* on this path such that if these variables were fixed at those values, there is some intervention on X that will change the value of Y.[11] The *contribution* to a change in the value of Y due to

[10] For a similar distinction, see Hitchcock (forthcoming).

[11] Why not simply say that X is a contributing cause of Y if and only if X is an ancestor of Y where the ancestor relationship is defined in terms of **DC**? The difficulty with this suggestion is that even if X is an ancestor of Y, it is possible that there are no interventions on X that will change the value of Y, for any values of variables that are not on the directed path from X to Y. For example, suppose that Z is an intermediate variable on the path from X to Y, and that the functions linking X to Z and Z to Y compose in such a way that some intervention on X will change Z for some values of off path variables (so that X is a contributing cause of Z) and some intervention on Z will change Y (so that Z is a contributing cause of Y) but that none of the changes in Z that might be produced by changes in X are such that they will produce changes in Y. (Many of the counterexamples to the transitivity of causation in the philosophical literature have this sort of structure.) In this sort of case, X is not a means for changing Y and it follows from the general connection between causation and manipulation assumed in the manipulablity theory that X is not a contributing (or total) cause of Y.

a change dx in the value of X (or the effect on Y contributed by this change in X) along some directed path is the change in the value of Y that would result from this change in X, given that the values of off path variables are fixed by independent interventions. For example, if we were to add a third equation (5.6) to (5.1-5.2) relating an additional variable W to X ($(5.6)X = eW$)—that is, if we were to draw an additional arrow from W into X—then although W is not a direct cause of Z, W will be a contributing cause of Z since, freezing the value of Y, there are interventions on W that will change Z. In particular, the contribution (along the path $W \dashrightarrow X \dashrightarrow Z$) to the value of X due to changing the value of W by amount dW is just $ebdw$. When relationships are linear, as in the above examples, it does not matter, for the purposes of identifying either the total effect on Y of a change in the value of X or the contribution this change makes to Y along some route, what values "other" variables assume. When relationships are nonlinear, both total and contributed effect will be relative to the values of other variables. For example, with $O = 0$, a change in the value of S will have no total (or contributed) effect on F. With $O = 1$, the total (and contributed) effect of a change in S from 0 to 1 will be to change F from 0 to 1.

In the case of (5.1-5.2) with $b = -ac$, X is not a total cause of Y but it is a contributing cause. Total causes will satisfy both **SC** and **NC**; contributing causes will satisfy **SC** but need not satisfy **NC**. Direct causes are always contributing causes but contributing causes need not be direct. For example, when the third equation (5.6) is added to (5.1-5.2), W is a contributing although not a direct cause of Z. Both directed graphs and equations aim, in the first instance, at the representation of direct rather than total causal relationships. If we have full information about the functional relationships that represent direct causal relationships, we may recover total causal relationships from this, as illustrated in some of the examples above, but directed graphs, by themselves do not convey such information.

I have argued elsewhere (Woodward forthcoming) that there is an important sense in which the notion of a direct cause (and more generally the notion of a contributing cause) is more fundamental than the notion of a total cause but in what follows I will assume only that the notion of a direct cause is *a* (not necessarily the only) legitimate notion of cause. This assumption is, I believe, implicit in the use of directed graphs and systems of equations to represent causal relationships and also follows from the underlying logic of a manipulability approach to causation. Even when $b = -ac$ in (5.1-5.2), one may still use X to change or manipulate Z—all that one has to do is to fix Y at

some value and then wiggle X. Thus there is a perfectly good sense in which X remains a means to changing Z and this, I claim, is enough to establish that there is a legitimate sense in which X is a cause of Z. That sense is captured by the notion of a direct (or contributing) cause.

6 Direct Causes and Probabilities

As explained above, one reason why we need the notion of a direct cause is to capture or represent facts about what will happen under combinations of interventions. Such facts are not always captured by information about total causal relationships, when these are understood as satisfying **SC** and **NC**. For example, in both (5.1-5.2) and (5.3-5.4) X is a total cause of Y, Y is a total cause of Z and X is a total cause of Z. Nonetheless X, Y and Z differ in the direct causal relationships in which they stand in (5.1-5.2) and (5.3-5.4). In this section I will argue that we also need the notion of a direct cause for another reason: in order to formulate plausible conditions connecting causal claims to claims about conditional probability.

I will focus on one of the best known proposals about this connection, the condition (**CC**) formulated by Nancy Cartwright (1983, p. 26) (Broadly similar proposals are endorsed by a number of other writers including Eells (1991) and Eells and Sober (1983). According to (**CC**)

C causes E iff $\Pr(E/C.K_j) > Pr(E/K_j)$ for all state descriptions K_j over the set $\{C_i\}$ where $\{C_i\}$ satisfies

(i) If C_i is in $\{C_i\}$, then C_i causes either E or not E.
(ii) C is not in $\{C_i\}$
(iii) For all D, if D causes E or D causes not E, then either $D = C$ or D is in $\{C_i\}$
(iv) If C_i is in $\{C_i\}$, then C does not cause C_i.

CC can be interpreted in at least two ways—as a condition on total causes and as a condition on contributing causes. In what follows I will assume the latter interpretation (i. e., I will use "cause" to mean "contributing cause"), unless explicitly indicated otherwise. **CC** differs from the characterizations of causation considered above in a number of respects. First, **CC** requires that causes (or better, a change in the value of the cause variable from absent to present) raise the probabilities of their effect. By contrast, both the contributing and total notions of cause described above require only that a change in the value of the cause variable change the value or the probability distribution of the effect variable. This difference strikes me as largely (although perhaps not entirely) terminological. **CC** attempts to capture the notion of a positive causal factor or of a promoting cause as opposed to the notion of a negative causal factor or a preventive or inhibiting cause. By contrast, **DC**, as well as **NC** and **SC**, attempt to capture the broader notion of

one variable's being causally relevant (either positively or negatively) to another. A variety of considerations of convenience seem to me favor this broader usage[12].

Second, **CC** imposes what has come to be called a unanimity requirement: C causes E if and only if C raises the probability of E across *all* background contexts K (or all situations that are "otherwise causally homogenous with respect to E" (Cartwright 1983, p. 25)). Several commentators (e.g., Dupre 1984) have objected that this requirement is too strong on the grounds that it has unintuitive consequences. For example, it requires that we withdraw the claim that "smoking causes lung cancer among human beings" if we were to discover even a small subpopulation in which, perhaps as a result of some genetic quirk, smoking fails to raise the probability of lung cancer. I agree with this objection but think that we do not need to leave matters at the level of intuition. If I am correct in claiming that the underlying point or rationale of our classifying the relationship between X and Y as causal or noncausal has to do with whether or not X is a potential means for controlling or manipulating Y, then there is little motivation for the unanimity requirement, since even if we agree to restrict the notion of cause to mean "positive or promoting cause", it is clear that X can be used to manipulate Y in a way that is positive for Y even if the unanimity requirement is violated. Instead what the manipulability conception (and in particular, **SC**) suggests is something like the following: X will be means for manipulating Y in positive way (i.e., for promoting Y), if as there is at least one background context in which X raises the probability of Y.

If we look for a connection between causation and facts about probabilities that is in the spirit of **CC** but incorporates these two points, it will be a proposal of the following form: X causes Y if and only if X and Y are dependent conditional on certain other factors F. The problem of finding a connection between causation and probability then becomes one of specifying what these other factors F are. In other words, the question is this: what should be held fixed (that is, conditioned on) if the conditional dependence of C on E is to be used as a test for whether C causes E? **CC** says that the other factors F that should be conditioned on are all other causes of E with the exception of

[12] A minor terminological annoyance is that to assess whether X causes Y, other factors that are negatively as well as positively causally relevant to Y must be controlled for. Thus if "cause" is restricted to mean "positive cause", it is incorrect to say that the only factors that need to be controlled for are "other causes of Y". One needs some additional vocabulary to describe the other factors that need to be controlled for. A more fundamental difficulty is that once one moves away from cause variables that are binary valued, it often becomes unclear what value of the cause variable corresponds to the "absence" of the cause and hence what the state is in comparison with which the "presence" of the cause should raise the probability of the effect.

those causes of E that are on the causal chain from C to E. (As Cartwright explains (1983, p. 26) condition (iv) in **CC** is intended to exclude conditioning on such factors). The motivation for not holding fixed causal factors that are between C and E may seem obvious. If we are dealing with a causal structure like that represented in (5.3-5.4) and Figure 5.2 in which there is a single directed path from X to Z with Y as a causally intermediate variable, then one would expect that conditional on Y, X and Z will be independent. Hence, if Y is one of the background factors on which we condition when we test for whether X causes Z we will reach the mistaken conclusion that X does not cause Z. However, as Cartwright herself recognizes (1983, p.30, 1989, pp. 95ff), the claim that, as (iv) requires, we should *never* control for such intermediate variables is too strong[13]. Suppose that we are presented with a triangular structure like that in (5.1-5.2) and Figure 5.1 in which both X and Y are direct causes of Z and X is also a direct cause of Y. Clearly if the direct causal connection between X and Z is to reveal itself in the probabilistic dependence of Z on X conditional on some appropriately chosen set of other factors, these other factors must include Y which is causally intermediate between X and Y. That is, to capture the direct influence of X on Z, we must in some way control for or correct for the influence of Y on Z. Moreover, since as we have seen the total cause structure is the same in both (5.1-5.2) and (5.3-5.4), we need to know the direct causal relationships (and not just the total causal relationships) between X, Y and Z in order to know what to control for when we test for, e. g., whether X is a cause of Z.

What these examples bring out is that in determining what should be controlled for or conditioned on for the purposes of assessing whether X causes Z, we need more than information about the other causes (in either the contributing or total sense) of Z besides X. We also need to know how, as it were, those other causes are connected up—with one another, with X and with Z. It is just this information about direct causal relationships that is contained in the associated equational or directed graph structure and this in turn suggests that information about such structures is essential if the sort of project represented by **CC** (the project of formulating systematic relationships between causal claims and conditional probability relationships) is to have any hope of success.

This point of view is also reflected in the so-called causal Markov condition (**CM**). This is a generalization of familiar ideas about screening off,

[13] Cartwright (1989, p. 96) abandons requirement (iv); replacing it with a requirement involving information about singular causal processes. I fully agree with Cartwright that some additional information is needed to distinguish what needs to be controlled for in structures like (5.1-5.2) and (5.3-5.4). However, my suggestion is that what is needed is rather additional information about direct causal relationships at the type-level.

first formulated by Reichenbach (1956), and has figured heavily in recent work on causal inference by Judea Pearl (2000) and by Clark Glymour and his associates (Spirtes, Glymour and Scheines 1993). **CM** says that conditional on its parents or direct causes, every variable is independent of every other variable except its effects:

> **(CM)** For all Y distinct from X, if X does not cause Y, then $Pr(X/Parents(X)) = Pr(X/Parents(X).Y)$

As with **CC**, the word "cause" in **CM**, can be interpreted as either "contributing cause" or "total cause". ("Parents" must of course be understood as meaning "direct cause"). I will assume the former interpretation unless indicated otherwise, although I think that insofar as **CM** is plausible at all, it is equally plausible under either interpretation. Like **CC**, **CM** connects claims about causal relationships to facts about relationships between conditional probabilities. However, unlike **CC**, **CM** is formulated in terms of information about the direct causes of X. As we will see, this is a crucial difference.

It is well known that there are circumstances under which **CM** fails to hold. For example **CM** will be violated if purely accidental correlations that reflect no causal connections at all occur. **CM** can also break down in the presence of cyclic causal relationships or when variables are measured imperfectly or when their values are drawn from mixtures of distinct probability distributions[14]. Although the point is not widely appreciated, **CC** will also be violated in all of these circumstances. However, when such circumstances are excluded, there is a plausible case to be made that **CM** follows from a manipulability conception of causation—or so Dan Hausman and I have argued elsewhere (Hausman and Woodward 1999). **CM** thus has some claim to be regarded as a conceptual truth about causation. Moreover, although I will not attempt to argue for this claim here, I conjecture that insofar as there is *any* systematic connection between causation and conditional independence relationships in acyclic causal systems, it is captured by **CM**. That is, to the extent that **CM** fails to hold, *no* general test for causation—neither **CC** nor any competitor—formulated in terms of conditional independence relationships will work. This gives us a reason for focusing on **CM** and asking what if anything it implies about the connection between casual claims and conditional probabilities.

Contraposing **CM** gives a *sufficient* condition for causation in terms of conditional dependence relationships: assuming that Y is distinct from X, if $Pr(X/Parents(X)) \neq Pr(X/Parents(X).Y)$ then X causes (i. e., is a contributing cause of) Y. However, the converse of this claim is *not* correct, if "cause"

[14] For a discussion of the circumstances in which **CM** fails, see Hausman and Woodward (1999).

means "contributing cause". In (5.1-5.2) and Figure 5.1, with $b = -ac$, X and Z are uncorrelated and hence $Pr(X/Parents(X)) = Pr(X/Parents\ (X).\ Z)$. Nonetheless X is a contributing cause of Z. We would have a necessary condition for contributing causation if we were willing to assume the converse of **CM**, which Spirtes, Glymour and Scheines (1993) call Faithfulness (**F**).

(**F**): If X causes (is a contributing cause of) Y, then $Pr\ (X/Parents\ (X) \neq Pr\ (X/Parents\ (X).Y)$

Taken together, **CM** and **F** provide necessary and sufficient conditions, formulated in terms of facts about relationships between conditional probabilities for X to be a contributing cause of Y. Interestingly, the conjunction **CM.F** is a condition that is rather different in form from **CC**, even allowing for the fact that the notion of direct causation plays no role in **CC**. While **CC** looks (roughly) at whether X and Y are dependent conditional on all *other* causes of Y (besides X) that meet certain additional conditions, **CM.F** asks whether X and Y are dependent conditional on *all direct* causes of X.

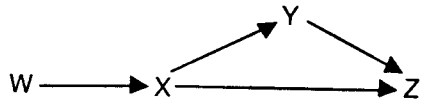
What justification is there for assuming (**F**)? Spirtes, Glymour and Scheines (1993) advance a measure-theoretic argument: given certain assumptions, violations of Faithfulness will be rare in the measure-theoretic sense. However, this is at best a reason for assuming faithfulness in causal inference problems—assuming faithfulness will only rarely mislead us. As Spirtes, Glymour and Scheines readily concede, "rare" does not mean impossible. To the extent that our interest is in giving a condition that is strictly necessary for it to be true that X causes Y (and not just a fallible test for whether X causes Y), assuming **F** is problematic. In contrast to **CM**, even those who advocate **F** do not suppose that it has any claim to be regarded as a conceptual truth about causation.

Is there some other candidate for a necessary condition that is more plausible than **F**? Although I lack the space for a detailed exploration of the possibilities, I think that there are reasons for skepticism. For example, the suggestion that a necessary condition for X to be a contributing cause of Y is that Y be dependent on X conditional on all direct causes of Y that are not identical with or descendants of X fails for several reasons, including the possibility of failures of faithfulness. The fundamental problem is that **CM** says merely that if certain causal relationships hold, then certain conditional independence relationships follow. **CM** doesn't say that if these causal relationships hold, then *only* these conditional independence relationships and no others hold[15]. However, something like this latter assumption seems required

[15] As (5.1-5.2) show, it is perfectly possible for independence relationships that do not follow from **CM** to hold because of, e. g., cancellations among coefficients in equations. It is cases of this sort that constitute violations of faithfulness.

if we are to have a necessary condition connecting causation and conditional probabilities. For such a necessary condition to hold, it would need to be true not just that (a) some inequality between conditional probabilities always indicates the presence of a causal relationship but that (b) causal relationships always reveal themselves in some inequality between conditional probabilities. While there is some reason to think that (a) is built into our concept of causation, it is hard to see how such an argument could be made on behalf of (b).

Does this assessment change if instead we look for necessary and sufficient conditions for X to be a total cause (rather than a contributing cause) of Y that are formulated in terms of conditional dependence relationships? As suggested above, **CM** yields a plausible sufficient condition for X to be a total cause of Y: (**TSC**) If Y is distinct from X and $Pr(X/Parents\ (X)) \neq Pr\ (X/Parents\ (X)).\ Y)$ then X is a total cause of Y. However, we cannot replace the reference to $Parents(X)$ in **TSC** with some condition formulated in terms of total causes—that is, the condition will fail to be sufficient if we fail to control for direct causes of X that are not total causes. As an illustration, consider the following structure



(5.1) $Y = aX$
(5.2) $Z = bX+cY$
(6.1) $W = dX$

Suppose as before that $ac = -b$ Then X is not a total cause of Z. Instead, Y is the only total cause of Z. Assuming **CM**, W is independent of Z conditional on both X and Y, but W is *not* independent of Z conditional just on Y. Nonetheless W is not a total cause of Z.[16] The inference from the fact that Z and W are dependent conditional on all of the total causes of Z to the conclusion that W causes Z is mistaken.

Formulating a sufficient condition for total causation in terms of conditional dependence relations thus requires information about direct causal relationships and this is an additional reason for thinking that the notion of a direct cause is an indispensable one. Moreover, again because of the possibility of violations of faithfulness, the converse of **TSC** does not hold. Consider a structure in which X and Y are the only direct causes of Z, Y is exogenous, and X is not, and Y is not a cause of X Then with right values of the coefficients linking X to Z and Y to Z, it is possible for X to be independent of Z

[16] Thanks to Chris Hitchcock for supplying this example.

and hence independent of Z conditional on *Parents (X)*. Nonetheless, in this structure X is a total cause of Z.

In summary, we may draw two more general conclusions. First, we require the notion of direct causation if we are to formulate any plausible connection between causation (whether contributing or total) and probability. Second any defensible connection between causation and probability is likely to involve only a sufficient rather than a necessary and sufficient condition. In this sense, the connection will be far weaker than the necessary and sufficient conditions sought in the philosophical literature on probabilistic causation. If we want necessary and sufficient conditions for causation that apply even in circumstances in which **CM** is violated, a counterfactual approach is more promising.

References

Cartwright, N. (1983) *How the Laws of Physics Lie.* Oxford: Oxford University Press.

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

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

Eells, E. (1991) *Probabilistic Causality.* Cambridge: Cambridge University Press.

Eells, E. and Sober, E. (1983) Probabilistic causality and the question of transitivity. *Philosophy of Science,* 50, 35-57.

Hausman, D. (1998) *Causal Asymmetries.* Cambridge: Cambridge University Press.

Hausman, D. and Woodward, J. (1999) Independence, invariance, and the causal Markov condition. *British Journal for the Philosophy of Science,* 50, 521-583.

Hitchcock, C. (forthcoming) *The Intransitivity of Causation Revealed in Equations and Graphs.*

Lewis, D. (1973) Causation. *Journal of Philosophy* 70, 556-567. Page references in text to reprinting in Lewis (1986).

Lewis, D. (1979) Counterfactuals dependence and time's arrow. *Nous* 13, 455- 76. Page references in text to reprinting in Lewis (1986).

Lewis, D. (1986) *Philosophical Papers,* Vol. 2, New York: Oxford University Press. Vol. 63, 113-37.

Meek, C. and Glymour, C. (1994) Conditioning and intervening. *British Journal for the Philosophy of Science* 45, 1001-1021.

Pearl, J. (1995) Causal diagrams for empirical research. *Biometrika* 82, 669-688.

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

Price, H. (1991) Agency and probabalistic causality. *British Journal for the Philosophy of Science* 42 157 -76.

Reichenbach, H. (1956) *The Direction of Time.* Berkeley: University of California Press.

Spirtes, P., Glymour, C. and Scheines, R. (1993) *Causation, Prediction and Search.* New York: Springer-Verlag.

Woodward, J. (1997) Explanation, invariance and intervention. *PSA 1996,* vol 2. S26- 41.

Woodward, J. (1999) Causal interpretation in systems of equations. *Synthese,* 121, 199-247.

Woodward, J. (2000) Explanation and invariance in the special sciences. *British Journal for the Philosophy of Science,* 51, 197-254.

Woodward, J. (forthcoming) Causality and manipulation.

Von Wright, G. (1971) *Explanation and Understanding.* Ithaca, New York: Cornell University Press.