Suppes, P. and J. A. de Barros: 1996, 'Photons, Billiards and Chaos', in P. Weingartner and G. Schurz (eds.), *Law and Prediction in the Light of Chaos Research, Lecture Notes in Physics*, Springer-Verlag, Berlin, pp. 189–201.

Suppes, P., J. A. de Barros, and G. Oas: 'A Collection of Probabilistic Hidden-Variable Theorems and Counterexamples', in R. Pratesi and L. Ronchi (eds.), *Conference Proceedings Vol. 60, Waves, Information and Foundations of Physics*, Società Italiana Di Fisica, Bologna, pp. 267–291.

Suppes, P. and M. Zanotti: 1981, 'When are Probabilistic Explanations Possible?', *Synthese* **48**, 191–199.

Department of Philosophy
Stanford University
Stanford CA 94305-2155
USA
E-mail: suppes@csil.stanford.edu

JIM WOODWARD

CAUSAL INTERPRETATION IN SYSTEMS OF EQUATIONS*

Individual equations or systems of equations are commonly used in the econometrics and causal modeling literature to make causal claims. Consider, for example, a linear regression equation with n independent variables $X_1, \ldots, X_n$ and an error term U.

$$(1) \qquad Y = a_1 X_1 + \cdots + a_n X_n + U$$

For example, with $n = 2$, Y in (1) might be interpreted as measuring the height of individual plants in some population, and X_1 and X_2 as the amount of water and fertilizer that these plants receive. Discussions of regression tell us that an equation like (1) can be used for several different purposes. It can be used simply to describe or represent the pattern of correlations among $Y, X_1, \ldots, X_n$ but it can also be used to make or represent a causal claim. In the latter case (1) is understood as claiming that $X_1, \ldots, X_n$ are causes of Y and that U represents the combined influence of all the other causes of Y besides $X_1, \ldots, X_n$ that are not explicitly represented in (1). I will call this the natural causal interpretation of (1). My interest in this essay is in what causal claims of this sort mean – both in the case of individual equations and systems of equations. What is it that we should understand (1) as claiming when we give it its natural causal interpretation? What conditions must be satisfied if (1) can regarded as making a true causal claim?

I distinguish this interpretive issue, as I shall call it, from issues about the conditions under which one can reliably estimate the coefficients in (1). The latter issues are epistemological in character – they have to do not with what (1) means, but rather with when and how one can determine the values of the coefficients in (1). Textbooks in econometrics have a great deal to say about such estimation problems. For example, a familiar result is that if the distribution of the error term satisfies various conditions, one of which is that the error term is uncorrelated with each of the independent variables X_i, then Ordinary Least Squares (OLS) estimators for the coefficients will have various desirable properties like unbiasedness. One of my themes in what follows will be that a substantial amount of recent discussion, both among philosophers and econometricians, has run together

Synthese **121**: 199–247, 1999.

interpretive and epistemological issues in connection with equations like (1). Conditions that have to do with reliable estimation such as those regarding the uncorrelatedness of the error term mistakenly have been taken as conditions which equations like (1) must satisfy if they are to have a causal interpretation.

The interpretive account I will be defending has a long history, although it often has been ignored or misunderstood in recent discussion. The basic themes can be found in econometricians like Frisch (1938) and Haavelmo (1944), who emphasize the notions of invariance and autonomy and the role of hypothetical experiments in causal interpretation. Among contemporary writers, Judea Pearl (1993, 1995, 1997) has articulated many of the ideas I will be describing in a particularly clear and compelling way. The notion of an intervention, on which I rely below, can be found in Pearl's work and also in the work of Clark Glymour and his collaborators (Spirtes et al. 1993).

While the ideas that follow are certainly not original, there are a number of reasons why they are worth the attention of philosophers. First, many of the philosophers (e.g., Cartwright 1989, 1995; Humphreys 1989; Papineau 1991; Irzik 1996; Hausman 1998) who have recently discussed causal modeling techniques have taken a very different view of the interpretive issues. Second, a number of the ideas on which I will rely, such as the connection between causation and manipulation, are regarded as unpromising starting points for understanding causation by many philosophers. In illustrating how these ideas may be used to provide a distinctive account of the content of causal models, I hope to persuade philosophers that this assessment is mistaken. Indeed, the ideas about hypothetical experiments and invariance described below are useful and suggestive not just in understanding causal models, but in understanding causal and explanatory claims in many other areas of science.

My discussion is organized as follows: Section 1 develops the basic ideas of invariance and of an ideal experimental manipulation or intervention and applies these to individual regression equations. Section 2 applies these to systems of equations. Section 3 compares these ideas with some other accounts advocated by philosophers and social scientists. Section 4 draws some general morals.

1.

The account that I will be pursuing can be motivated by appealing to a common sense idea about the relationship between causation and manipulation – that whatever else may be true of causes, they are potential handles

or means for manipulating their effects. If we apply this thought to an equation like (1) we are led to the following suggestion: the causal claim represented by (1) is true, if and only if we (or nature) were to conduct an ideal hypothetical experiment in which the values of any of the rhs variables X_i or U were manipulated within some range, Y would change in just the way represented by (1). Thus, for example, if X_i were changed in such an experiment by amount dX_i but none of the other rhs variables in (1) were changed, then Y would change by amount $a_i dX_i$. Similarly, changes in the value of U that do not change the values of the variables X_i should also change the value of Y in the way that (1) describes. To employ standard terminology (cf. Woodward 1997) the relationship (1) should be *invariant* under such manipulations of X_i and U: it should remain stable or continue to hold under such changes. If, on the contrary, under an experimental manipulation that sets the variable X_i to some particular value x_i' the coefficient a_i or any other coefficient in (1) changes or if the functional form of (1) changes, then (1) is not invariant under this manipulation and (1) does not correctly describe the causal relationship between X_i and Y *for this value of* X_i. (See below for a discussion of what the italicized phrase means.) Since, as we shall see, there are several varieties of invariance relevant to the causal interpretation of systems of equations, I will call the notion just described level invariance, since it has to do with whether (1) is invariant under interventions that appropriately change the levels of its independent variables, including U.

1.1. *Interventions*

The connection between causation and manipulation just described is very rough: it needs to be qualified and stated much more carefully if it is to be remotely defensible. The remainder of this section attempts to do this. First, what do we mean by an ideal hypothetical experiment? Clearly not all ways of manipulating X_i for the purposes of determining whether X_i causes Y will satisfy the conditions for such an experiment. For example, if the event M which changes X_i also directly causes a change in Y, then changes in X_i might be accompanied by changes in Y in a way that satisfies (1) even though there is no causal relationship at all between X_i and Y. A similar possibility will be present if the event M which changes X_i also changes one of the other variables $X_j (j \neq i)$ which is also a cause of Y, or if the change in X_i produced by the manipulation is merely correlated with changes in other causes of Y. What we need to do is to define the notion of an ideal experimental manipulation so as to exclude these and other possibilities.

A serviceable approximation to the notion we want is this: an ideal experimental manipulation M of X with respect to Y (or for the purposes of determining whether X causes Y) is an exogenous change in the value of X which changes the value of Y if at all only through X and through whatever variables are causally between X and Y and not through some other route. Moreover, M must not be correlated with changes in Y that proceed through some other route (that, is through a route that does not go through X and those variables are causally between X and Y). I will call a process or event which changes X in a way that satisfies these conditions an *intervention* on X with respect to Y. We assume that if it is true that Y changes or would change under such an intervention on X in the way specified by some putative causal relationship, this can only be because the relationship really is causal. On the other hand, if a relationship fails to be invariant under any interventions at all on its independent variables (i.e., in the variables it represents as causes) the causal facts are not as the relationship represents them as being.

There are now a number of characterizations of the notion of an intervention in the philosophical literature which are broadly similar to the notion just described, although they differ in detail (see, e.g., Cartwright and Jones 1991; Meek and Glymour 1994; Pearl 1995; Hausman 1998). Fortunately, most of what I will have to say in what follows will not turn on these details – the reader can simply think of an intervention on X with respect to Y as a process that satisfies whatever conditions he or she thinks must be satisfied in an ideal experimental manipulation of X for the purposes of determining whether X causes Y.

1.2. *Clarification*

There are a number of other respects in which connection between causation and intervention advocated above requires clarification. As a point of departure, I should emphasize that the above remarks are put forward as an account of what it is for a particular kind of equation or mathematical relationship – namely, a regression equation of form (1) – to correctly represent a quantitative causal relationship. They are *not* put forward as a completely general account of what it is for a relationship to be causal. Obviously, it can be true that X_i causes Y even if (1) is not invariant under interventions on X_i. This will happen if the relationship between X_i and Y is causal but does not conform, in its quantitative behavior, to (1) – for example, if the relationship is nonlinear. However, it would be a mistake to conclude from this that the account offered above fails to capture what must be true for an equation like (1) to describe a causal relationship. While (1) says, when interpreted causally, that each of the independent

variables X_i causes Y, it does not *just* say this. Rather it makes a much more specific quantitative claim about the causal relationship between X_i and Y. The ideas sketched above are put forward as an account of how this quantitative claim is to be understood. Although I believe that the general outlines of the account sketched above can be extended (with appropriate modifications) to other sorts of causal relationships, I will not try to argue for this contention here.[1]

A second important point, which is implicit in my discussion above, is that a relationship such as (1) may be invariant may correctly describe the results of hypothetical interventions – for some range of values of its independent variables that are set by interventions but not others. In fact, it seems plausible that many if not almost all of the relationships described by regression equations in the social and biological sciences will behave in this way. For example, even if it is true that (1) is invariant under some range of interventions on the amount of water that a plant receives it is clearly not invariant under all such interventions – one cannot make a plant grow arbitrarily tall by putting arbitrarily large amounts of water on it. A similar point holds for many other social, biological and physical relationships that we regard as causal. Consider the relationship between the restoring force F exerted by a particular kind of spring and its extension X. This may closely conform to Hooke's law (2) $F = -kX$ but only for (interventions that will produce) a certain range of extensions. If one stretches the spring too much, (2) will break down.[2]

One possible response to these observations is that if there are any values of their independent variables produced by interventions for which (1) and (2) fail to be invariant, then they do not represent genuine causal relationships. Bona-fide causal relationships, it may be said, must be invariant under all possible interventions that set their independent variables to all possible values. Given the connection between causation and manipulation defended above, this position seems unreasonably stringent and unmotivated. As the examples of (1) and (2) above show, as long as a relationship is stable under interventions that set some values of their independent variables, one can use it (with respect to those values) to manipulate, even if it is not stable under all interventions. Hence, there is a natural motivation for saying that relationships that are invariant under some but not all interventions can qualify as causal.

We can accommodate the above observations by explicitly relativizing the notion of invariance (and, correlatively the connection between manipulation and causation) to a range of values of variables set by interventions or, to express the same idea more compactly, to a range of interventions. I will thus say that a relationship like (1) is invariant un-

der some interventions but not others, and correctly describes the results of some hypothetical experiments but not others. To qualify as a correct causal description, a relationship must be invariant under some range of interventions but it need not be (and in the case of the sorts of causal relationships typically described by regression equations will not be) invariant under all interventions. Thus for example, (1) may qualify as a true causal generalization (with respect to those values of its independent variables for which it is invariant) if it is invariant under interventions that set the value of the amount of water that the plants in some population receive to 1, 2 or 3 liters, even if (1) would break down if the plants were to receive 1000 liters of water. In such circumstances (1) does correctly describe how, by manipulating the amount of water that a plant receives within the 1–3 liter range, one may manipulate its height within a certain range and this establishes that (1) is a correct causal description within this range.

I turn now to another issue concerning how the notion of invariance is to be understood. My argument has been that a necessary and sufficient condition for an equation like (1) to correctly represent a causal relationship is that it be invariant under some range of interventions on its rhs variables. However, I do *not* claim that when (1) represents a causal relationship it will be invariant *only* under interventions on such variables. Typically, when (1) represents a causal relationship, it will be invariant under many other sorts of changes as well. It is analytically useful to separate such "other changes" into two categories. First, there are changes in what we may call background conditions. These are conditions or factors that are not included in (1), either among the measured variables or in the error term. Second, there are changes in variables that explicitly occur in (1) where those changes do not involve interventions. I will first explore the significance of stability under the first category of change and then turn to the second category.

As an illustration of the notion of background conditions, return to the interpretation of Y in (1) as measuring the height of individual plants in some population, and X_1 and X_2 as measuring the amount of water and fertilizer that individual plants receive. Then background conditions for (1) will include changes in the position of Mars, changes in the color of the shirt worn by the person administering the water, and changes in the day of the week. One expects that (1) will be stable under many changes in background conditions of this sort, although of course it is an empirical matter exactly which such changes will or will not disrupt (1).

Should we conclude from this that it is only a necessary but not a sufficient condition for an equation such as (1) to describe a causal relationship, that it be invariant under some range of interventions on its independent variables changes, and that an additional necessary condition is required – namely that it be invariant under some range of changes in background conditions as well? Once we have decided to relativize the notion of invariance to a range of interventions, this additional move appears to be unnecessary. Any intervention must occur in some background circumstances or other (e.g., it will be an intervention in which an experimenter in a red shirt puts a liter of water on each of the plants on a Tuesday afternoon while the position of Mars is such and such) and we can always incorporate reference to these into the characterization of the intervention. If (1) is invariant under an intervention occurring in certain background conditions that sets the variable X_i to some value x^*, but not invariant under other interventions occurring under different background conditions that also set $X_i = x^*$, we can convey this fact by being explicit about exactly which interventions in which background conditions fall within the range of invariance of (1). In general, rather than trying to formulate an additional necessary condition having to do with invariance across changes in background conditions for (1) to represent a causal relationship (and hence struggling with the problem of finding a non-arbitrary answer to the question of *which* background conditions) it seems preferable and less arbitrary to say simply that different causal claims will differ in the range of changes in background conditions under which they are invariant, and that we can spell out the content of different casual claims by being explicit about the range of interventions and other changes over which they are invariant. This having been said, we should also note that while a relationship can certainly count as causal if it fails to be invariant under some changes in background conditions, relationships that are, so to speak, almost endlessly sensitive to changes in such conditions – that are altered or disrupted in indefinitely many ways as background conditions shift – are regarded as of little scientific interest, even if they appear to be invariant under interventions in some highly specialized background circumstances.

The second possibility described above concerns changes that are not produced by interventions, but which do occur in variables that explicitly figure in a relationship. Consider a casual process that alters both X_1 and X_2 in (1). Such a process will not count as an intervention on X_1, since the change produced in X_1 will be correlated with the change produced in another cause of Y, namely X_2 and by parallel reasoning also will not count as an intervention on X_2. Nonetheless, we expect that if (1) describes a causal relationship, it will be stable or invariant under some range of such changes. It is important to understand what this means: in saying that (1) will be invariant under such changes what we mean is that the change in Y will be what (1) says it will be, given the changes in X_1 and X_2. Thus

if X_1 is changed by amount dX_1 and X_2 by amount dX_2, the total change in Y will be $a_1dX_1 + a_2dX_2$ if (1) is invariant under this change. This contrasts with the change in Y that would be produced by an intervention on X_1 which is just a_1dX_1. To put the point slightly differently, we may interpret the individual coefficients in a linear regression equation as telling us what change in Y would be produced by interventions on the associated rhs variables – thus the coefficient $a_1(a_2)$ tells us what the change in Y would be under an intervention that produces a unit change in $X_1(X_2)$. (Recall that an intervention on X_1 will change X_1 but not any of the other rhs variables in (1).) When several of the rhs variables in (1) are changed, (1) will often remain invariant but in such cases, the total change in Y will be the net effect or sum of the contributions made by the changes in each of the independent variables where each of these contributions is the change in Y that would have occurred if an intervention had occurred on just that variable. Thus what is "the same" across cases in which X_1 is altered by amount dX_1 by an intervention, and cases in which X_1 is altered by amount dX_1 and X_2 is altered by amount dX_2 is the *contribution* (namely a_1dX_1) made to the total change in Y by the change in X_1 or, alternatively, the relationship (1). Both of these need to be distinguished from the total change in Y which will not of course be the same in these two cases.

I have belabored this point because a number of recent criticisms of the idea that causal relationships are invariant relationships turn on misunderstandings about what should be expected to be invariant in cases in which an effect has multiple causes. Both Richard Healey (1992) and Dan Hausman (1998) note that in cases in which both X_1 and X_2 are causes of Y, the total change in Y will be different depending on the whether a change occurs just in X_1 or in both X_1 and X_2. Hausman takes this to spell trouble for standard formulations of the thesis that causal relationships are invariant (Hausman 1998, 222ff.). In doing so, he interprets the thesis that causal relationships are invariant as the thesis that the total change in Y will be the same, regardless of whether X_1 alone is changed by an intervention or both X_1 and X_2 are changed. However, when invariance is understood in the way advocated above and when we recognize the important difference between the claim that the *relationship* (1) is invariant and the (obviously false) claim that the total change in Y per unit change in X_i will be the same regardless of what else changes along with X_i, we see that cases involving multiple causes pose no threat to the connection between causation and invariance that I advocate.

The claim that (1) is invariant under some interventions on its independent variables and the claim that it is invariant under some non-interventions

that change those variables are logically or conceptually distinct. It appears to be logically possible to have one sort of invariance without the other. However, it is hard to think of realistic cases of causal relationships in which this happens. As before, rather than looking for some additional invariance condition, having to do with invariance across changes that are not interventions, that will distinguish those generalizations that describe causal truths from those that do not, it seems more natural and less arbitrary to say that typical causal generalizations are invariant not just under a range of interventions but also under a range of changes in their independent variables that do not result from interventions as well as under a range of changes in background conditions and that other things being equal we prefer relationships that are invariant under a larger or more important range of interventions and changes.

I can further clarify the connection between causation and invariance that I advocate by contrasting my views briefly with those of Nancy Cartwright. In recent work (e.g., 1995) Cartwright suggests that we should distinguish between two questions – whether a relationship is causal and the extent to which it is stable or invariant across various sorts of changes. She contrasts what she calls "mere causal relationships" with "capacities that will be stable across a range of envisioned changes" or as she also describes them, "*super*causal relations" which "remain invariant across a range of envisioned interventions" (1995, 55). A relationship can be causal without being "supercausal" or "invariant" (1995, 56). Situations meeting the conditions for a controlled experiment will sometimes allow us to establish "what causal relationships obtain in that situation" but, because causation is different from invariance, we cannot infer from this "what causal relationships will obtain outside that situation". Additional assumptions about capacities or invariance are required to "export" causal conclusions to different situations.

Where Cartwright sees a sharp distinction between two questions (Is this relationship causal? Is it invariant?), I see something more closely resembling a continuum. If one accepts the view that causal relationships tell us something about the results of hypothetical experiments, then it seems to follow that some measure of invariance under interventions is necessary for a relationship to be causal at all. On any view that connects causation and manipulation, relationships that are so unstable that they are disrupted by all possible attempts to use them to manipulate will not qualify as causal. From this perspective, what relatively stable or invariant relationships have is not some additional feature that is not present at all in "mere" causal relationships but rather more of the same feature – invariance and stability under a larger or more important range of inter-

ventions and changes than is present in mere causal relationships. This is not to deny Cartwright's point about exportability, which is correct and important, but rather to re-express it in terms of the idea that a relationship can be invariant under a relatively limited or narrow range of changes and interventions and hence causal, but can fail to be invariant under other changes and interventions which may be of interest to us.

1.3. *Comparison with Traditional Manipulability Theories*

The idea that the existence of a causal relationship between X and Y has to do with the behavior of the relationship between X and Y under interventions on X figures centrally in the so-called manipulability or agency theories of causation that have been developed by philosophers like von Wright (1971) and Price and Menzies (1993). Such theories are subject to a number of serious criticisms, and because of this, many philosophers have concluded that any view that connects causation and manipulation is a non-starter. For this reason, it is important to see that the position sketched above is not subject to the difficulties that face traditional manipulablity theories. The latter theories are reductionist in aspiration and avowedly anthropocentric. Their basic strategy is to appeal to some antecedently understood concept of human agency or manipulation and to use this to provide a non-circular, reductive account of what it is for one event to cause another. For such a reduction to work, the relevant notion of agency must not itself be a causal notion, or at least must not presuppose all of the features of the notion of causation we are trying to explicate. Such theories are thus naturally led to take human agency as a primitive, irreducible feature of the world that stands outside or behind the rest of the natural causal order, rather than just one variety of causal transaction among others (von Wright 1971, 74; Price and Menzies 1993, 190ff.).

There are many reasons why theories of this sort are unpromising. To begin with, if our discussion above is on the right track, the notion of an intervention that is required for a successful explication of the connection between causation and manipulation is a thoroughly causal notion, and this completely undercuts the reductive project. This is so not only in the most obvious respect – namely that when we say that an intervention I changes X, this must be understood as meaning that I "causes X to change" but in a variety of more subtle respects as well. As we have seen, the notion of an intervention I on X with respect to Y requires reference as well to the presence or absence of various other causal relationships besides the causal relationship between I and X. For example, there must be no direct *causal* connection between I and Y and I must not *cause* changes in or be correlated with other *causes* of Y besides X (except, as noted above, for

those that are causally between I and X or X and Y). This use of causal language is ineliminable – we can not replace it with causally uncommitted talk of correlations. Nor can we replace it with notions having to do with human agency. For one thing, manipulations carried out by human beings can (and frequently do) fail to meet the causal conditions for an intervention described above. For example, a human experimenter can carry out a manipulation and have the concept or experience acting "freely" as a human agent, but if her manipulation of the treatment variable is correlated with other causes of the effect variable, her action will not qualify as an intervention and her causal inferences probably will be mistaken. It is a fundamental objection to traditional agency theories that it is not the experimenter's agency per se that underpins her causal inferences, but rather whether her actions satisfy the causal conditions on interventions described above.

The connection between causation and manipulation described above differs from traditional manipulability theories in several important respects. First, it is non-reductionist in aspiration. The notion of an intervention to which I have appealed is a straightforwardly causal notion which is not tied to human activities in any essential way. The idea is to use background information about some casual and correlational relationships – information about the presence of a causal relationship between I and X, about the absence of a casual relationships between I and Y that does not go through X and so on – to say what it is for a *different* casual relationship – a causal relationship between X and Y – to hold. The connection between causation and manipulation to which I have appealed is thus not viciously circular in the sense that to say what it is for an intervention to occur on X with respect to Y, one has to have already decided whether there is a causal relationship between X and Y.[3] Because this connection is non-reductionist, it avoids those criticisms of manipulabilty theories that are premised on their reductionist aspirations. And because the notion of an intervention is characterized in explicitly causal terms, we avoid any appeal to the antinaturalist idea that human action is somehow special and outside the causal order of the rest of nature.

Another criticism lodged against traditional manipulability theories is that they are objectionably anthropocentric in the sense that they tie the existence of causal relationships much too closely to facts about the manipulations that human beings can actually carry out. Because such theories hold that our concept of what it is for a relationship to be causal is logically or conceptually tied to the concept or experience of human agency, they face great difficulties in explicating causal claims (e.g., "changes in the position of the moon cause changes in the tides on earth") for which the

relevant human manipulations are impossible. Advocates of manipulability theories often respond by claiming that ascriptions of causal relationships in such cases involve a projection or non-literal, analogical extension of notions that apply literally only when human manipulation is feasible, but this strikes most critics as implausible. The connection between causation and intervention advocated above avoids this difficulty for two reasons. First, while the characterization of an intervention employs causal notions, it makes no reference to human beings or their activities. A natural causal process that does not involve human beings or their activities will count as an intervention as long as it has the right sort of causal history. So-called "natural experiments" illustrate this possibility. Human interventions count as interventions because (or if) they have the right sorts of causal characteristics, not because there is anything special about human agency per se. Second, the connection advocated above is hypothetical or counterfactual – the claim is that when a relationship like (1) is a correct causal description then *if* an intervention were to occur that alters X, Y would change in the way described by (1). It is not required that the intervention in question actually occur or even that it be physically possible in the sense of being consistent with the laws and initial conditions that actually obtain.[4]

1.4. *The Error Term*

There is a final implication of the ideas about invariance that I have been defending that is worth special emphasis at this point. This has to do with the role of the error term in (1). As I remarked above, when we claim that a regression equation is level invariant, this implies not just that the relationship is invariant under (some range) of interventions that change the level of the independent variables X_i that explicitly figure in (1), but also that a similar invariance claim is true for changes in the value of the error term U. That is, it is assumed that (again at least within a certain range) changes in the value of U will not affect the functional relationship (1). This means, among other things, that this functional relationship should be invariant under changes in the *distribution* of the error term. This in turn has the important consequence that, contrary to what many writers maintain, it is *not* necessary, if a regression equation is to have a causal interpretation or to accurately describe a set of causal relationships, that the error term in the equation be uncorrelated with the each of the independent variables in that equation. With respect to the error term, what matters is not its actual distribution but rather whether the equation is invariant under changes in its distribution. Suppose that we are estimating a regression equation of form (1) for a certain population and that the uncorrelatedness assumption does turn out to be satisfied. The idea that if this equation correctly represents a

causal relationship, it should be invariant under changes in the distribution of the error term means that if the distribution of the omitted causes U of Y changes in such a way that the error term now is correlated with one or more of the independent variables, the relationship (1) between Y and $X_1, \ldots, X_n$, U should nonetheless continue to hold (in the sense that this relationship should continue to describe what would happen to Y under interventions on $X_1, \ldots, X_n, U$). Because the error term is now correlated, one will no longer be able to use OLS estimation techniques (although one may still be able to use other estimating techniques – see below) but assuming that (1) is invariant under this change, it will be just as much a correct causal representation as before. I will return to this topic in Section 3 below

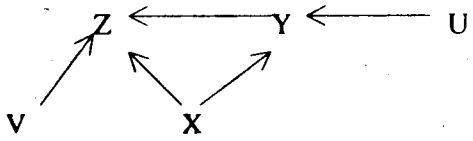
2.

Regression equations represent a particularly simple sort of causal structure in which a single dependent variable is represented as causally influenced by a set of independent variables but in which no causal relationships are represented as holding among the independent variables themselves and no reciprocal or cyclic causal links back from the dependent variable onto the independent variables are represented. Often, however, social scientists and other users of causal modeling techniques want to represent more complex structures. This is accomplished through the use of systems of equations. The conventions for interpreting such equations causally parallel those for single regression equations. Each equation in a system contains a single dependent variable on the left hand side and one or more right hand side variables which are interpreted as the direct causes of the lhs variable. If one wishes to represent causal relationships among the rhs variables one adds additional equations conforming to the convention just described. For example, in the system of equations

$$(3) \qquad Y = aX + U$$

$$(4) \qquad Z = bX + cY + V$$

(3) says that X is a direct cause of Y, and (4) says that X and Y are direct causes of Z (as before, U and V are error terms that represent causes of the dependent variable in each equation that are unmeasured and not explicitly represented). We may also represent the structure (3)–(4) by means of a directed graph, following the convention that an arrow directed out of one vertex and into another means that the former is a direct cause of the latter.



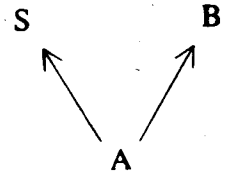
2.1. *Modularity*

In view of our earlier discussion, it is natural to impose the requirement that if a system of equations like (3)–(4) correctly represents the causal facts, then each individual Equation (3) and (4) must be level invariant under some range of interventions on the rhs variables of that equation. However, there is also a second, distinct invariance requirement which it is natural to impose on systems of equations. To motivate this requirement, consider a simple illustration. Atmospheric pressure A is a common cause of the reading B of a barometer and the occurrence or non-occurrence of a storm S but neither B nor S are causes of each other. We can represent this by means of the following equations (where I have suppressed the error terms since they are not essential for the point that follows)

(5) $S = cA$

(6) $B = dA$

or by means of the following diagram



How can we interpret this system in terms of the results of hypothetical experiments? Suppose that we intervene on A. Then if (5) and (6) correctly describe the casual structure of the system, the value of S should change in the way described in (5), and similarly the value of B should change in the way described by (6). Suppose, on the other hand, a set of interventions I occur on the position of the barometer dial B. This might be accomplished by consulting a randomized device , the output of which is causally independent of and uncorrelated with A (or any other causes of B) and, depending on this output, manually fixing the dial at one position or another. Of course what we expect to find, if the causal structure of this system is correctly specified by (5)–(6), is that there will be no correlation at all between changes in S and changes in B when produced by I. This

corresponds to the fact that B is not a cause of S. However, the logic of this reasoning depends crucially on an assumption that has not yet been made explicit. An intervention on B of the sort described above changes the causal relationship between A and B. When the interventions I are carried out, the value of B is entirely determined by the interventions and is no longer causally affected by the value of the atmospheric pressure A. In effect, the intervention replaces (6) with a new equation (6*) $B = I$.

The crucial assumption that we made above is that it is possible to replace (6) with (6*) without changing or disrupting Equation (5) that when an intervention on B changes the relationship between A and B, it does not automatically change (or need not change) the causal relationship between A and S. To see the importance of this assumption, suppose that any possible intervention on B does alter the causal relationship between A and S, as expressed by (5) – for concreteness suppose that it changes (5) to (5*) $S = c^*A$ where $c^* \neq c$. Then under any such intervention on B, the value of S would systematically change, since the effect of a given value of A on S changes as the coefficient in (5) changes. This is exactly the sort of behavior that we would ordinarily take to show (on a manipulationist conception of causation) that B causes S but (5)–(6) do not say that there is any causal relationship between B and S.

As an another illustration, consider an intervention on Y in the equation system (3)–(4). We can think of this as replacing (3) with a new Equation (3*) $Y = I$ which represents the fact that the value of Y is now determined by the intervention variable I rather than, as was previously the case, by the variables X and U. If the result of any such intervention is that Equation (4) is disrupted, then the value of Z will not be influenced by Y in the way that (4) claims. More generally we can say that if any intervention that changes one equation also alters other equations in a system, then the system will be misspecified in the sense that it will fail to correctly and completely represent the causal structure that it purports to model – variables will change in response to interventions on other variables even though the equations represent the variables as causally unrelated or, alternatively, will fail to change under interventions in the way that the equations suggest that they should. In either case there will be a mismatch between what the equations say and what will in fact happen under interventions.

What is it that justifies us in thinking that in the case of the atmospheric pressure, storm, barometer system it should be possible to manipulate the barometer without altering the relationship between the atmospheric pressure and the occurrence of the storm? A very natural answer which is implicit in a great deal of econometric thinking is this: we suppose that the mechanism (or casual route or relationship) by which the atmospheric

pressure affects the barometer when the latter is operating normally is different or distinct from the mechanism by which the atmospheric pressure influences whether or not a storm occurs. Because these mechanisms are distinct, it makes sense to suppose that one mechanism can be changed or interfered with or without a change necessarily occurring in the other. If, on the contrary, there is no possible way of intervening on B without at the same time altering the relationship between (or the mechanism linking) A and S, we have grounds for doubting that these mechanisms are really distinct. We would also have no well-defined answer to questions about whether (and to what extent) S would change just in response to interventions on B, and it is just this answer which characterizes the causal effect of B on S.

These considerations can be used to motivate the following proposal: When a system of equations is used to represent causal structure, then not only should each equation in the system be level invariant but each equation in the system should represent or correspond to a distinct causal mechanism or relationship (or to a set of mechanisms or relationships) which are distinct from the mechanisms or relationships corresponding to other equations.[5] Mechanisms M_1 and M_2 are distinct when it is possible in principle for M_1 to change or be interfered with without M_2 changing or being interfered with and vice-versa. It follows that, in a system of equations that correctly represents causal structure, not only should each equation in the system be level invariant, but that the system should also satisfy the following additional invariance condition: for each equation there should be some possible change that alters that equation or replaces it with another equation while leaving the other equations in the system unaffected. Here, changing an equation means changing the mechanism(s) or relationship(s) represented by it, and we can think of this as a matter of intervening on the lhs (dependent) variable in the equation so that the value of that variable is now fixed by the intervention rather than by whatever variables previously determined its value. When we say that the other equations are unaffected, we mean that they would continue to hold and continue to be level invariant under this change. Somewhat more concisely: each equation should be invariant not only under (some range of) interventions on its independent variables but also under some possible changes in the other equations in the system. When a system of equations possesses this feature, I will say that it is *modular* or *equation – invariant*. We thus have:

MODULARITY. A system of equations is modular iff (i) each equation is level invariant under some range of interventions and (ii) for each equation

there is an intervention or the dependent variable that changes only that equation while the other equations in the system remain unchanged and level invariant.

I have defined modularity in such a way that if a system is modular, then each equation in the system must be level invariant. Nonetheless, modularity and level invariance are distinct concepts. In particular, it is not true that if each equation in a system is level invariant the system must be modular. (As we shall see below, the so-called reduced form equations associated with a system of equations will always be level invariant but need not be modular.) Intuitively, level invariance is a condition that applies within individual equations and concerns whether an individual equation is invariant under interventions on its rhs variables. By contrast, modularity is an invariance condition that also applies between equations and has to do with whether each equation is invariant under changes in other equations.

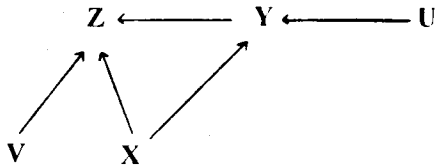
We can bring out more clearly what modularity involves by re-writing the system of Equations (3) and (4) as follows

(3) $Y = aX + U$

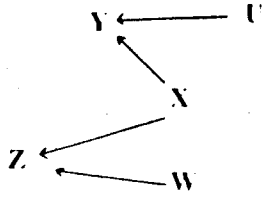
(7) $Z = dX + W$

where $d = b + ac$ and $W = cU + V$.

Since (7) is obtained by substituting (3) into (4), the system (3)–(7) has exactly the same solutions in X, Y, and Z as the system (3)–(4). Since X, Y and Z are the only measured variables, (3)–(4) and (3)–(7) are in a sense observationally equivalent – they imply or represent exactly the same facts about the patterns of correlations among these measured variables. Nonetheless, by the rules given above for interpreting systems of equations, these two systems correspond to different causal structures. (3)–(4) says that X is a direct cause of Y and that X and Y are direct causes of Z. By contrast, (3)–(7) says that X is a direct cause of Y and that X is a direct cause of Z but says nothing about a causal relation between Y and Z. This difference is also reflected in the graphical representation associated with the two systems:



(3)-(4)



(3)-(7)

We can gain some additional insight into what modularity means by noting that, despite their observational equivalence, if (3)–(4) is modular, then (3)–(7) cannot be (and vice-versa). To see this, consider an intervention on the variable Y in (3). The result of this will be to replace (3) with a new Equation (3*) $Y = I$, specifying that the value of Y is now determined by the intervention variable, rather than by X. In effect, the coefficient a in (3) is set equal to zero by the intervention. If the system (3)–(4) is modular, (4) will continue to hold under this change in (1). By contrast, if (3)–(4) is modular, (7) must change under this intervention since, as we have seen, its effect is to change the value of the coefficient a in (3) and the coefficient d in (7) is a function of a. Thus changing a in (3) will change d and hence (7). This corresponds to our judgment that if (3)–(4) is a correct representation of the causal facts then (3)–(7) collapses or mixes together distinct mechanisms or causal routes – the influence of X on Z that occurs because X directly influences Z (this is represented by the coefficient b) and the influence which occurs because X influences Y which in turn influences Z (this is represented by the product ac) – into a single overall mechanism linking X and Z, which is represented by the coefficient d. This failure to correctly segregate the system being modeled into distinct mechanisms is directly reflected in the non-modularity of (3)–(7).

We can also bring out the difference between (3)–(4) and (3)–(7) in a slightly different way. Suppose that an intervention occurs on Y. Then if (3)–(4) is a correct representation of the causal facts (and hence is modular), we know that while (3) will be disrupted, (4) will not be. Hence, if Y is set to some new value k it will continue to make just the contribution to Z that is indicated by (4) – i.e., the contribution will be ck. Thus, according to (3)–(4), an intervention on Y will change the value of Z. By contrast, according to (3)–(7), Y does not cause Z and hence an intervention on Y should not change Z. (Recall that an intervention on Y should be uncorrelated with other causes of Y such as X). Of course what will happen under such an intervention, if (3)–(4) is the correct structure, is (as we have seen) that the coefficient d will change in (7) so as to reflect the change in the value of Z that is really produced by the change in the value of Y, but this fact about the dependence of Z on Y is not represented in the Equations (3)–(7). Thus, assuming that (3)–(4) is modular, (3)–(7) fails to correctly represent what will happen under hypothetical interventions on Y.

There are many other systems of equations besides (3)–(7) that may be obtained from (3)–(4) by equality-preserving algebraic transformations and are in this sense observationally equivalent to the latter. For example, we can also rewrite (3)–(4) as

(3) $\quad Y = aX + U$

(8) $\quad Z = eY + R$ where $e = (b/a + c)$ and $R = V - (b/a)U$

Again, if (3)–(4) is modular, (3)–(8) will not be, since changing a will change the value of e. Again, (3)–(8) represents a different set of causal claims from (3)–(4) – something which is born out in the fact that (3)–(8) makes different predictions about what will happen under various hypothetical predictions. If one accepts that, despite their observational equivalence, at most one of these systems of equations can correctly represent the causal facts, there must be some additional constraint that is satisfied by the correct system. Modularity is the natural candidate for this constraint. The idea that among all of the observationally equivalent representations, we should prefer the one that is modular (because it will be the one that correctly represents the causal relationships) picks out, as Alderich (1989) puts it, a "privileged parameterization".

The Equations (3)–(7) are the reduced form equations associated with the system (3)–(4). In general, one forms the reduced form equations associated with a system by first identifying the exogenous variables in the system – i.e., the variables that are not themselves caused by any of the other variables in the system and do not have arrows directed into them in the graphical representation of the system. One then substitutes into the equations in the system in such a way that one is left with a set of equations, one for each endogenous variable, which have the endogenous variable on their lhs and only exogenous variables (and an error term) on their rhs. It is always possible to do this, and the resulting reduced form system will always be observationally equivalent to the original system regardless of its structure. Moreover, the error term in each equation will always be uncorrelated with the rhs variables in that equation and hence one may always estimate the values of the coefficients in the reduced form equations by OLS. By contrast, as we shall see in more detail below, it is not true for all systems of equations that the values of the coefficients are estimable from statistical data about the measured variables – in the jargon of econometrics, some of the coefficients may be *unidentifiable*. In addition to this, as long as no changes occur in the coefficients in the original system (either because of interventions on the endogenous variables or for some other reason), the reduced form equations will be level invariant (under interventions on the exogenous variables in each equation). If we care only about producing an observationally adequate representation of the pattern of correlations among the measured variables, or if we care only about finding an observationally adequate representation that is level invariant, we may just use the reduced form equations.

However, as the preceding discussion illustrates, the reduced form equations need not be modular. For example, on the assumption that (3)–

(4) are modular, the associated reduced form Equations (3)–(7) will not be. We thus see that, as claimed above, it is possible to have a system of equations, all of which are level invariant but which fail to be modular. Again, if researchers are not always content with the reduced form representation – and it is clear that they are not – this can only be because they value something else (modularity) besides level invariance and observational adequacy.

2.2. *Autonomy*

I can further bring out the significance of this last point by connecting it with a frequently-cited passage from Tygre Haavelmo's monograph (1944) in which he introduces the notion of autonomy. It will be recalled that Haavelmo envisions a researcher who investigates the relationship – call it (R) – between the maximum speed attained by a particular make of car and the depression of the gas pedal as this "experiment" is repeated under exactly the same conditions – the same road conditions, fuel mixture, state of the engine and so on. It is plausible that, provided these conditions continue to obtain, (R) will be level-invariant under some range of interventions that depress the gas pedal. Because of this – because (R) describes how, within this range, one can use the pedal to manipulate the speed – (R) qualifies as a genuine causal relationship. Nonetheless, as Haavelmo says, (R) strikes us explanatorily shallow and as scientifically uninteresting. Haavelmo contrasts (R) in this respect with an engineering style theory – call it E – in which the operation of the car is decomposed into a number of distinct mechanisms and the principles governing these. We expect that (E) will be modular in the sense that it should consist of independently changeable representations of the operations of mechanisms that are in fact distinct from one another, so that when we change the representation in E of the operation of one mechanism (e.g., the spark plugs) this does not automatically necessitate changes in the representation of the operation of other mechanisms (e.g., the relationship between the air pressure in the tires and friction with the road surface). It is this feature which will allow us to trace out the implications of hypothetical changes in the operation of the various components of the car, just as the modular representation (3)–(4) allows us to trace out the implications of changes in the individual mechanisms it represents. A theory like E should thus enable one to see how, if the operation of any of the mechanisms making up the car were to change (e.g., the spark plugs are cleaned or more air is put in the tires) or if various factors in the environment (the grade of the road, the head wind etc.) were to change, the operation of the car and the relationship between the gas pedal and the speed would change. While each of the individual

equations in (E) will be invariant under changes in the other equations, (R) will fail to be invariant under most such changes. Haavelmo says that because (R) is invariant under a smaller set of changes and interventions than the equations in E, it is less *autonomous* than those equations, and he links this to the fact that (R) is less satisfactory from the point of view of explanation and causal understanding than E.

We can think of (R) as like a non-modular reduced form equation. Just as the coefficient d in Equation (7) (on the assumption that (3)-(4) is modular and (3)–(7) is not), represents a sum or mixture of the coefficients corresponding to several distinct mechanisms, so (R) summarizes an overall relationship between speed and gas pedal position that is the combined upshot of the operation of many different mechanisms making up the car. Like (R), (7) will be invariant under some interventions that change the level of its exogenous variable X_1 and hence will qualify as causal. But both (R) and (7) will continue to hold only as long as none of the many causal relationships or mechanisms that contribute to the overall relationship they describe are altered. For example, a change in any one of the coefficients b, a or c will change the relationship described by (7), just as a change in the any one of the mechanisms making up the car engine will disrupt (R). Just as (R) is less autonomous than (E), so (7) is thus less autonomous than, say, (4) in the sense that interventions that disrupt (7) will also disrupt (4), but there are interventions – for example, interventions that disrupt (3) – that will disrupt (7) but not (4). Just as we believe that E is more satisfactory than (R) from the point of view of causal explanation, so we should prefer (3)–(4) to (3)–(7).

2.3. *Clarifications*

Before proceeding, let me address some possible misunderstandings of the argument of the preceding paragraphs. First, there are of course many causal systems for which there are at present no technologically feasible methods that allow for separate interference with all of the distinct mechanisms that compose them. As with level invariance, when we ask whether there is an intervention on one equation that would leave other equations unchanged, what we are interested in is what would happen under certain hypothetical possibilities, and not whether such interventions are practically possible. In the example discussed above, what makes (3)-(7) non-modular (if (3)–(4) is modular) are the mathematical relationships between the coefficients of the two systems. It is these that insure that if the coefficient a in (3) changes, the coefficient in (7) must change – i.e., that there is not even a hypothetical intervention that changes (3) without changing (7).

Second, I empasize that the argument is conditional in character: *if* (3)–(4) is modular, then (3)–(7) will not be modular and will fail to correctly represent the results of various hypothetical interventions. Since the coefficients a, b and c in (3)–(4) can also be written as functions of a and d in (3)–(7), a parallel argument could also be used to show that *if* (3)–(7) is modular, then (3)–(4) will not be. What the argument shows is thus that, at most, one among the various alternative systems of equations relating X, Y and Z that are observationally equivalent can be modular, and hence that modularity represents a real constraint on the choice of a system of equations. However, the argument does not purport to tell us which, if any, of these alternative systems is in fact the modular one. It is nature or the world – and, in particular, facts about what the causal mechanisms are and about what would happen under different hypothetical changes – that determines this. In other words, researchers must first determine, in some independent way, which mechanisms are distinct and what would happen under various hypothetical interventions. They may then represent this information by a system of equations, guided by conventions of the sort described above: that different equations should correspond to distinct mechanisms, that if an intervention on X would change Y, then X should occur on the rhs of an equation and Y on the lhs and so on.

These observations may also help to address another concern that may trouble some readers: (3)–(7) has been obtained from (3)–(4) by a series of equality-preserving algebraic transformations. In view of this, how can it be true that these two systems of equations represent different causal claims? Aren't the two systems "mathematically equivalent"? If (3)–(4) and (3)–(7) said only that the quantities X, Y and Z are regularly associated or correlated in a certain pattern, then they would indeed be interchangeable representations of the same set of correlational facts. However, when interpreted causally, systems like (3)–(4) say more than this. As Judea Pearl (1998) has emphasized, what we may think of as the syntactic form of these equations also conveys information – information about what would happen under hypothetical interventions. For example, when we write Y on the lhs of an equation and X on the rhs, we convey the information that an intervention on X would change Y, but not the information that an intervention on Y will change X. This is so even though we can derive (9) $X = (1/b)Z-(c/b)Y$ from (4) $Z = bX+cY$ and vice-versa. Similarly, writing (3)–(4) rather than (3)–(7) conveys, through the syntactic convention, that different equations should represent distinct mechanisms, one set of claims about what the causal mechanisms are rather than another and an accompanying set of claims about what will happen under hypothetical interventions. It is because such syntactic information is lost or changed

under algebraic manipulation that it is possible for these two systems to represent different sets of causal facts.

I said above that (3)–(4) and (3)–(7) were in a sense observationally equivalent. The preceding paragraphs enable us to be somewhat more precise about what this means and to understand how, despite this fact, they can represent different causal structures. Assuming that their coefficients and error terms are related in the way described, the two systems agree, so to speak, about what has actually happened – about the actual patterns of correlations among the measured variables that obtain so far. What they disagree about is modal or counterfactual: they make inconsistent predictions about what would happen should various changes or interventions occur. It is entirely possible, of course, that these changes will not occur, in which case the two systems will continue to agree about what will be observed. When we take the two systems to disagree nonetheless about what the causal facts are, we accept that there is a fact of the matter about what would happen under various hypothetical changes even if these never occur.

2.4. *Modularity and Graphs*

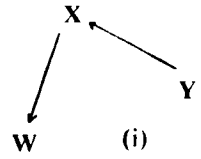
It is also worth noting that these ideas about modularity just described are closely connected to a set of ideas about the graphical representation of interventions that have been developed by Judea Pearl (Pearl 1995) and by Peter Spirtes et al. (1993). According to these writers, an intervention on a variable X may be represented by drawing an arrow directed into X from the intervention variable and removing all other arrows directed into X. All other arrows in the graphical representation, including all arrows directed out of X are preserved. Thus, for example, interventions on Y and B in the systems

replace them respectively with

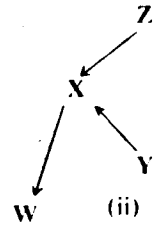
Directed graphs provide a qualitative representation of causal relationships – they represent whether or not there is a causal relationship between two variables (whether the coefficient linking the variables is zero or non-zero) but, unlike systems of equations, do not represent what the quantitative strength (the numerical value of the coefficient) of that relationship is. But apart from this difference, there is a close correspondence between this graphical interpretation of interventions and the ideas about level invariance and modularity described above. The intuition behind the idea that interventions on a variable break all other arrows directed into it is that the value of the variable is now completely determined exogenously by the intervention rather than by the endogenous causal influences that previously determined its value. The idea that arrows directed out of the variable intervened on are preserved represents a qualitative version of level invariance – the thought is that if a directed arrow from X to Y represents a genuine causal relationship, then that relationship, and hence the arrow, should be preserved under interventions that change the value of X. The idea that all other arrows in the graph (i.e., those that are neither into or out of X) are preserved is a qualitative version of the modularity requirement. As before, we may motivate this idea by appealing to assumptions concerning the distinctness of mechanisms. Each set of arrows directed into a variable X is associated with a distinct equation, and hence ought to represent a mechanism or set of mechanisms that is distinct from the mechanisms represented by arrows directed into other variables. Because arrows directed into different variables represent distinct mechanisms, it is possible to intervene on X and to disrupt the mechanism that determines the value of X without disrupting any other mechanism. The idea that when we intervene on X, we break arrows directed into X while preserving arrows directed into other variables requires that it be possible to disrupt the equation in which X occurs as a dependent variable without affecting the equations corresponding to other sets of arrows directed into other variables, and this of course is just the idea of equation invariance or modularity. If a system of equations is not modular then the corresponding graph will exhibit action at a distance in the sense that the graph will behave as though interventions on some variables change arrows directed into or out of other variables. We may take this to indicate that the graph does not accurately and completely represent the causal relationships among those variables.

Let me conclude by noting that, while it is natural to represent many interventions in terms of arrow-breaking or the complete replacement of one equation by another, the main claims of this section concerning modularity do not require this assumption. Consider a situation in which there

is a causal connection from Y to X and in which a new exogenous source of variation Z is supplied to X but without disrupting the Y–X connection. As long as Y and Z are uncorrelated this would also supply to X the kind of independent variation that is demanded of an intervention. Graphically, this would amount to replacing (i) with (ii):



(i)



(ii)

Algebraically we might represent such a change by adding a new term to the equation $X = aY$, to yield $X = aY + Z$. Some readers may find this a more natural way of representing at least some interventions.

The important point for our purposes is that, regardless of how interventions are represented, we still require the notion of modularity for causal interpretation. It still must be the case that when we replace (i) with (ii) the addition of a new arrow into Y does not alter arrows elsewhere in the graph directed into other variables. Similarly, the addition of a new term to an equation is a change in that equation and must not change other equations in the system if it is to be fully causally interpretable.

3.

Our discussion so far has focused on examples involving linear equations, but it is readily generalized to equations involving arbitrary functional forms. By doing this we can be more explicit than we have hitherto been about what the notion of an intervention involves, and we can also connect the ideas described above to an explicitly probabilistic framework for thinking about causation that may seem more familiar to philosophers. Following Pearl (1998b), let us define a causal theory T as a four-tuple $T = \langle V, U, P(u), \{f_i\}\rangle$ where

(i) $V = \{X_1, \ldots, X_n\}$ is a set of measured variables
(ii) $U = \{U_1, \ldots, U_n\}$ is a set of exogenous error variables
(iii) $P(u)$ is a probability distribution over $U_1, \ldots, U_n$.
(iv) $\{fi\}$ is a set of n functions, each having the form $X_i = f_i$ (Parents X_i, U_i) for $i = 1, \ldots, n$.

Intuitively, the parents X_i are those variables in V that represent the direct causes of X_i. The probability distribution $P(u)$, together with the functions f_i, determine a probability distribution over all the variables in V.

Within this framework to say that some individual equation $X_i = f_i$ (Parents X_i, U_i) is level invariant is just to say that the function f_i (whatever it may be) is invariant or continues to hold on some range of interventions on the variables Parents X_i and U_i. As before, we may think of each individual equation as corresponding to or representing a distinct causal mechanism. As before, we assume that if mechanisms are distinct, then it ought to be possible to disrupt or change the corresponding equation without changing other equations in T. As before, if T meets this condition, we may describe it as modular or equation-invariant.

Again following Pearl, let us introduce an operator set $(X = x)$ to represent the fact that the value of the variable X has been set equal to x by an intervention where this notion is understood along the lines described above. We may then represent the effect of an intervention which sets $X_i = x$ in the following way: we delete (or "wipe out") from T all equations f_i in which X_i is the dependent variable, and replace them instead with equations which specify $X_i = x$. All other equations are left undisturbed. By then substituting $X_i = x$ for every occurrence of X_i among the independent variables (parents) in each of the f_i we may trace out the effects of the intervention $X_i = x$.

If we are willing to assume that set X is itself a random variable with a well-defined probability distribution (as would be the case if, for example, the values of set X are determined by some appropriate randomizing device as in the storm–barometer example of Section 2), then we may talk about the probability distribution of some other variable Y conditional on X's being set to various values – that is, $P(Y/\text{set } X)$. In this connection, it is worth emphasizing that it is not true in general that $P(Y/\text{set } X) = P(Y/X)$ when $P(Y/X)$ is understood as the probability of Y conditional on X's being observed to have various values. For example, the probability of the occurrence of a storm conditional on the barometer's being observed to have some value is quite different from the probability of the storm conditional on the barometer's being set to that value. This is just the well-known distinction between conditioning and intervening (cf. Meek and Glymour 1994).

Given this understanding of the "set" operator and its connection with the "wipe out" procedure for equations described above, various rules or requirements governing its behavior follow. For example, we have the

following probabilistic analogue to MODULARITY which we may call PROBABALISTIC MODULARITY (PM).

(PM) $P(X/\text{Parents}(X)) = P(X/\text{Parents } X, \text{set } Z)$ where Z is any set of variables distinct from X.

(PM) expresses the idea that once one conditions on the full set of causes of X, setting any other variable should make no additional difference to the probability of X. Each conditional probability $\Pr(X/\text{Parents } X)$ is determined by the corresponding equation $X = f_i$ (parents X, U) and the probability distribution over U, and hence we may think of each conditional probability (like each equation) as corresponding to a distinct mechanism. Hence, (PM) is another way of expressing the idea that we may disrupt any other mechanism in the system (by setting the dependent variable for that mechanism equal to some exogenously determined value) without disrupting the conditional probability $\Pr(X/\text{Parents } X)$. Note also that it will not in general be true that $P(X/\text{Parents } X) = P(X/\text{Parents } X, Z)$ – that is, if one replaces set Z in (PM) with Z, the resulting statement will no longer be true. For example, conditioning on an effect Z of X may provide information about the value of X even when the values of the parents of X are taken into account. By contrast, if the value of the same variable Z is set by an intervention, then it will provide no information about the value of X, since the effect of the intervention is to break the previously existing causal relationship between X and Z. This provides a further illustration of the difference between conditioning and intervening.

We also have an obvious probabilistic analogue of level invariance:

PROBABALISTIC LEVEL INVARIANCE (PLI) $P(X/\text{Parents } X) = P(X/\text{set parents } X)$.

This represents the idea that if Parents X are the causes of X, then the conditional probability $P(X/\text{Parents } X)$ should be invariant under interventions that change the value of any of the variables in Parents (X).

If we are willing to assume in addition that the causal theory with which we are dealing satisfies the Causal Markov condition (which says that conditional on its parents every variable in V is independent of every other variable except its effects), then a number of additional rules governing the set operator will hold. Two such rules, discussed in Hausman and Woodward (forthcoming) and labeled by us (PM2) and (PM3) are

(PM2) When X and Y are distinct, $P(X/\text{set [Parents } Y\text{]}, Y) = P(X/\text{set [Parents } Y\text{]}, \text{set } Y)$

(PM3) If X does not cause Y, then $P(X/\text{Parents } X) \;\&\; \text{set } Y) = P(X/\text{Parents } X \& Y)$.

Additional rules governing the set operator and a much more detailed and systematic discussion connecting it to graph-theoretical representations of causal relationships are given in Pearl (1998b). Rules of this sort connect the set operator (and the notions of intervention and invariance) to the probability calculus and to more familiar mathematical operations like conditioning, and they also make explicit when we may move from causal knowledge and information derived from passive observations (recorded in the probability distribution P) to predictions about what will happen under interventions. They ought to reassure the reader that the notions of an intervention and of invariance are not obscure metaphysics but can be given a precise mathematical characterizaion. As Pearl has put it (private correspondence), they show that that notion of an intervention is useful for calculating with and not just thinking about causal relationships.

4.

The account just described, which emphasizes the role of invariance and interventions in the causal interpretation of systems of equations, contrasts with another position which is held by a number of philosophers and contributors to the causal modeling literature. This alternative account focuses instead on the role of assumptions about the uncorrelatedness of the error term in causal interpretation. As we have seen, these assumptions take two forms. First, there is the assumption – call it (U1) – that the error term in an individual equation or, in the case of systems of equations, the error term in each individual equation, is uncorrelated with each of the rhs variables in that equation. Second, in the case of systems of equations, there is the assumption (U2) that the error terms in different equations are uncorrelated with each other. It will be important to keep the distinction between (U1) and (U2) in mind in what follows, since neither entails the other and since philosophers who have discussed the role of "the" uncorrelatedness assumption in causal inference have not always made it clear whether they have in mind (U1) or (U2) or both.

4.1. *Uncorrelatedness and Causal Interpretation*

The guiding idea of those who emphasize the role of the uncorrelatedness of the error term is that satisfaction of one or both of the assumptions (U1) and (U2) is necessary and/or sufficient for a system of equations to provide a correct description of causal structure. While this suggestion is in itself quite interesting, it is sometimes accompanied by an additional, stronger suggestion: that the relevant uncorrelatedness assumptions are, as it were,

purely correlational and can be formulated without making use of causal assumptions at any point. If so, and if satisfaction of these assumptions is necessary and sufficient for the causal claims expressed by a set of equations to be correct, we appear to be well on the way to a successful Humean reduction of causal claims to claims about the presence or absence of correlations. This idea (or something very close to it) has recently been endorsed by David Papineau (1991). Papineau writes

> ... take the first two equations in the normal triangular array:
>
> (10) $X_1 = U_1$
>
> (11) $X_2 = a_{21}X_1 + U_2$
>
> Now these two equations are indeed algebraically equivalent to:
>
> $$X_2 = a_{21}U_1 + U_2$$
>
> $$X_1 = 1/a_{21}X_2 - 1/a_{21}U_2$$
>
> which represents X_2 as the independent variable X_1 as dependent on X_2. However, if the error terms in the original equations are probabilistically independent, as required, then the 'error terms' in the re-written equations won't be: the 'error term' in the second equation $-1/a_{21}U_2$ will be negatively correlated with the 'error term' in the first, $a_{21}U_1 + U_2$, and also with the other exogenous variable in the second equation, X_2.[6]

Papineau's suggestion is that, among all the systems of equations that are observationally equivalent to (10)–(11), we should regard that system in which the uncorrelatedness assumptions (U1) and (U2) are satisfied as the correct representation of causal structure. If the uncorrelatedness assumptions are satisfied for the first system of equations but not for the second, this shows that X_1 causes X_2 rather than vice-versa. As he puts it:

> So the requirement of independent error terms is a real constraint, which ensures that the ordering of variables in a set of regression equations isn't just an arbitrary importation of prior causal assumptions ... from a metaphysical point of view, there is nothing to stop us regarding the probabilistic independence of the error terms as a basic and objective fact, from which the causal ordering derives. (1991, 400)

A number of other writers eschew the reductionist aspect of Papineau's program, but endorse or are sympathetic to the idea that satisfaction of the uncorrelatedness assumption is a sufficient and/or necessary condition for a system of equations to represent a correct description of some causal structure. For example, while the views expressed by Nancy Cartwright (1989) and subsequent papers are subtle and complex, it seems fair to say that she also regards assumptions about the uncorrelatedness of the error terms as central to the question of when a system of equations can be given a causal interpretation. In a broadly similar vein, Gurol Irzik advocates a non-Humean account of causation, but writes that

a crucial assumption of causal modeling is that an error term is uncorrelated not only with other error terms, but also for each equation, with the causes in that equation and also with each of the earlier causes in other equations. All models – path or structural – endorse this assumption one way or another (1996, 255, italics in original).

Irzik goes on to argue that this assumption plays "an indispensable role in causal interpretation". Another illustration is provided by the economists T. F. Cooley and S. F. Leroy. In their influential paper (1985), they endorse the connection between invariance, claims about the outcomes of hypothetical experiments, and causation advocated in sections **1** and **2**, but then go on to suggest that satisfaction of an uncorrelatedness assumption concerning exogenous variables is necessary for such invariance claims to have a clear meaning:

> In order for the required invariance under intervention to have unambiguous meaning in all contexts, particularly large systems of equations, the assumption that all exogenous variables be uncorrelated is required. (1985, 293)

Elsewhere they suggest that "the role of the uncorrelatedness assumption" is "essential to specify(ing) precisely what is invariant under hypothesized intervention(s)" on the exogenous variables.

Why think that satisfaction of one or both of the uncorrelatedness assumptions is crucial for causal interpretation? One apparently natural line of thought, described but not endorsed by Cartwright, and advocated by Papineau and Irzik, is that violation of the uncorrelatedness assumptions will occur when and only when some sort of specification error is present (a mistake about causal direction, omission of a relevant variable, etc.). Cartwright puts the idea this way:

> Specifically, why should one think that the independent variables are causes of the dependent variables so long as the errors satisfy the no-correlation assumptions? One immediate answer invokes Reichenbach's principle of the common cause: if two variables are correlated and neither is a cause of the other, they must share a common cause. If the independent variables and the error term were correlated, that would mean that the model was missing some essential variables, common causes which could account for the correlation, and this omission might affect the causal structure in significant ways.[7] (1989, 24)

Papineau endorses a similar line of thought:

> If an error term *were* correlated with one of the other causes, then this would indicate some kind of hidden causal connection which would invalidate the causal order postulated by the system of equations. (1989, 401)

As Cartwright notes, this emphasis on the uncorrelatedness assumptions brings together the two sets of considerations that I separated at the beginning of this essay: semantical issues about causal interpretabillty (what does a system of equations mean when it has a casual interpretation,

what conditions are necessary and/or sufficient for the system to correctly describe a set of causal relationships) and epistemological issues having to do with estimation and identifiability. As observed earlier, when the error terms within each equation are uncorrelated with the rhs variables in that equation, OLS estimates of the coefficients in each equation will have desirable features like unbiasedness. All such coefficients will be identifiable – given these assumptions we can infer unique values for the coefficients from the statistical data. Thus, if the claims described above are correct, the uncorrelatedness assumptions play both a central epistemological role and a central role in causal interpretation. Cartwright describes the idea this way:

> the very same condition (regarding uncorrelatedness) that ensures the identifiability of the parameters in an equation like that for X_e [where X_e is the dependent variable in an equation in a recursive system] also ensures that the equation can be given its natural causal reading – so long as a generalized version of Reichenbach's principle of the common cause can be assumed. (1989, 31)

These claims about the role of the uncorrelatedness assumptions in causal interpretation present an important challenge to the position defended in sections **1–2**. If the uncorrelatedness assumptions are necessary and/or sufficient for causal interpretation but pick out the same relationships as causal as the invariance–based account, then it may be that the former is more fundamental or that, as Cooley and Leroy suggest, the latter account requires the former if its central notions are to be well-defined. And if the two approaches pick out different relationships as causal, we need to ask which is correct.

In what follows, I will argue for the following conclusions. First, to avoid trivialization, the uncorrelatedness assumptions must be understood in causal rather than purely correlational terms. This undercuts Papineau's reductionist proposal. Moreover, even when the error terms are interpreted causally, satisfaction of (U1) and/or (U2) is neither necessary nor sufficient for a set of equations to correctly represent causal structure or to have a well-defined causal interpretation. Instead, as urged in Section 1, the role of the uncorrelatedness assumptions is purely epistemological: they are simply one of a set of conditions that are jointly sufficient, in systems of equations that have a certain structure, for OLS estimation to have desirable properties like unbiasedness. Moreover, although it is sufficient for this epistemological role, satisfaction of the uncorrelatedness assumptions is not necessary: depending on the structure of the equations in question there are non – OLS techniques by which one can reliably estimate the coefficients even if neither uncorrelatedness assumption is satisfied. In such cases, the equations can certainly have a causal inter-

pretation. Indeed, a system of equations can have a causal interpretation even if there is no way at all to reliably estimate their coefficients from the statistical data. To suppose otherwise is to assume that, if there is no way of determining which of several competing causal structures is correct from available data, there is no correct structure, or that if the statistical data is unable to discriminate between competing systems of equations, the equations themselves lack a coherent causal interpretation. In addition, the assumptions (U1) and (U2) will only be satisfied in one particular sort of causal structure – so-called recursive systems. In non-recursive systems, which are employed extensively, the uncorrelatedness assumption (U1) is regularly violated. If, as I will argue, such systems can be given a coherent causal interpretation, this will again show that (UI) is not necessary for causal interpretability. Finally, it is also mistaken to argue that satisfaction of either (U1) or (U2) is required for the notion of invariance under intervention to be well-defined In general, then, it is invariance and not uncorrelatedness which is the key to causal interpretation. The separation of epistemological and interpretive issues advocated at the beginning of this essay is well motivated.

4.2. *Uncorelatedness and Reduction*

I begin with the reductionist version of Papineau's proposal. A fundamental difficulty with his suggestion is that if the uncorrelatedness assumptions are to have any determinate content at all, they cannot be formulated in purely correlational terms. Instead, the error terms must themselves be given a substantively causal interpretation – they must be understood to represent the net effect of omitted *causes* of the lhs variables in the equations in which they occur, rather than omitted variables which are merely correlated with the lhs variables. Moreover, the rhs variables in each equation that, according to assumption (U1), must be uncorrelated with the omitted causes represented by the error terms must be *causes* of the dependent variable in each equation, rather than merely being correlated with the dependent variable.[8]

One way of seeing this is to note that, given a body of statistical data for the variables $Y, \ldots, X_1, \ldots, X_n$, if we pick any one of them as the dependent variable and the remainder as independent or rhs variables, and then take the coefficients in the resulting regression equation to be given by their OLS estimators, the error term automatically will be uncorrelated with the rhs variables by construction. Since this can be done for any choice of dependent and independent variables, the fact that the error term when constructed in this way is uncorrelated with the independent variables cannot possibly guarantee that the resulting equation is a correct

causal description. To illustrate this point in the simplest possible case – univariate regression – consider the equation

$$(3) \qquad Y = aX + U$$

and define a as the regression coefficient estimated by OLS, i.e., $a = E(YX)/E(XX)$. Then

$$U = Y - aX$$

$$XU = YX - a(XX)$$

$$E(XU) = E(YX) - a(XX) = 0$$

i.e., when a is so defined, U is uncorrelated with X. By contrast, if we had begun instead with the equation

$$(12) \qquad X = bY + V,$$

and defined b as $b = E(XY)/E(YY)$, then a parallel argument would show that V is uncorrelated with Y. In effect, our choice of a defines the error U in (3) in such a way that it is equal to the "residual" $Y - aX$, which by construction must be uncorrelated with X. Similarly, our choice of b defines a residual $V = X - bY$ that must be uncorrelated with Y in (12). It should be clear, however, the fact that it is always possible to do this does *not* show that either Equation (3) or Equation (12) is a correct casual description or that a in (3) has its natural causal interpretation – that it tells us what would happen to Y if we were to intervene to change the value of X. It is only if a has this causal interpretation, and the error term is understood as $Y - aX$ when a is so interpreted, that there is a sensible, nontrivial question about whether the error term is correlated with X.[9]

4.3. *The Irrelevance of Uncorrelatedness to Causal Interpretation*

The above argument undercuts attempts to use the uncorrelatedness assumptions as part of a program to reduce causal claims to claims about correlations, but it leaves open the possibility that satisfaction of one or both of the uncorrelatedness assumptions (U1) and (U2), when these are given some substantive causal interpretation, is necessary and/or sufficient for a system of equations to represent causal structure correctly. Suppose that we confine ourselves to equations in which the error terms represent omitted causal factors. Does it follow that representations in which the error terms are uncorrelated are in some way preferable or privileged with

respect to causal interpretation? I think that there are a number of reasons for denying this.

First, there are a number of procedures that allow for estimation of the coefficients in a system of equations when one or both of (U1) and (U2) are violated – it is just that such procedures do not involve the use of OLS estimators. For example, if a system is hierarchical – i.e., contains no causal cycles or loops – then, depending on details of the system, it may be possible to use instrumental variables or indirect least squares to estimate the coefficients even if (U1) or (U2) is violated.[10] In such cases, we may still think of the coefficients in each equation as having the causal interpretation advocated above – i.e., each coefficient describes how the dependent variable in that equation will change in response to an intervention on the independent variable associated with that coefficient. This seems to show that satisfaction of (U1)–(U2) is not necessary for causal interpretability and that their role is purely epistemological – they are just one of a set of conditions that permit the use of a certain estimating technique, which is just one such technique among several.

Second, consider again the reduced form equations formed from some arbitrary system of equations. As we noted, the reduced form equations will always satisfy the uncorrelatedness assumption (U1), although not necessarily (U2). Since the reduced form equations sometimes incorrectly represent causal structure, it follows that satisfaction of (U1) is not sufficient for the correct representation of causal structure. Since OLS estimators will be unbiased for the coefficients in the reduced form equations and the coefficients will always be identifiable, it follows that it is not true in general that the conditions that insure unbiased estimation and identifiability also insure correctness of causal representation.

4.4. *Non-Recursive Systems*

There is yet another reason for thinking that the uncorrelatedness assumption (U1) is not necessary for causal interpretability. This has to do with the existence of situations whose causal structure cannot be represented by any recursive systems of equations. For our purposes, we may think of a recursive system as a system which is (a) hierarchical – there are no causal loops or cycles in which X causally influences Y and Y in turn causally influences X – and which (b) satisfies the uncorrelatedness assumptions (U1) and (U2). In a system which is non-hierarchical, some of the error terms *must* be correlated with the independent variables in the equations in which they occur, in violation of (U1). A simple example is provided by the following system.

$$(13) \quad X_3 = aX_1 + bX_4 + U$$

$$(14) \quad X_4 = cX_2 + dX_3 + V$$

The error term U enters into the equation (13) for X_3, and hence is correlated with X_3. But since according to (14), X_3 is also a cause of X_4, X_3 is also correlated with X_4. Hence, U will be correlated X_4 – i.e., U will be correlated with one of the r.h.s. variables in (13). I will suggest below that structures like (13)–(14) can be given a coherent causal interpretation, and if this is correct, it follows that the satisfaction of the uncorrelatedness assumption (U1) cannot possibly be a necessary condition for causal interpretability. Similarly, Papineau's claim that once we have written down a system of equations in a form in which the error terms occurring in each equation are uncorrelated with the r.h.s. variables in that equation and the error terms are uncorrelated across equations, we have found or fixed the right causal ordering cannot possibly be correct if nonhierarchical systems are possible. In general, it is only in hierarchical systems that the assumptions (U1) and (U2) can both be satisfied. In assuming that the correct causal structure must be one in which (U1) and (U2) are satisfied, we are assuming, on *a priori* grounds, that the correct causal model must be hierarchical. For the same reason, Irzik is mistaken in claiming that all causal models endorse assumptions (U1) and (U2).

Can non-recursive models be given a coherent causal interpretation? Are there real world causal systems that are represented by such models? While I will not try to explore these questions in detail, a few general remarks about the status and interpretation of such models may be helpful. First, such models are employed very widely in economics and other areas of social science. Many social scientists believe that such models can be given a causal interpretation, and that many of the systems they investigate are appropriately described by such models. There are a number of reasons for the popularity of non-recursive models. For one thing, many theories in the social sciences attempt to describe equilibrium outcomes and these are naturally modeled by systems of equations that contain reciprocal links. For example, a system like (13)–(14) might be used to represent a supply and demand equilibrium: we can think of (13) as a demand equation in which X_3 is the quantity demanded of some good, X_4 its price, and X_1 is some other variable assumed to influence demand. (14) is the supply equation relating the price of the good X_4 to the quantity supplied X_3 (at equilibrium equal to the quantity demanded) and some other factor X_2 assumed to influence supply. While it may seem natural to suppose that underlying (13)–(14) is some more complex dynamical process that might be modeled by a hierarchical, non-cyclic system with time indexed variables (cf. Cartwright 1989, 16–17), theories of the dynamics underlying the attainment of equilibria are often underdeveloped in social science.

While it is certainly desirable to develop more adequate dynamical models, equilibrium models are widely regarded as having considerable explanatory power and empirical support. It seems misguided to reject them simply because they are nonrecursive.[11]

How then are the reciprocal links in a model like (13)–(14) to be understood? I don't want to claim that there is any single answer to this question, for it seems plausible that social scientists use non-recursive systems to model many different kinds of processes, and different interpretations seem plausible in different cases. Nonetheless, as Judea Pearl (1993) has recently emphasized, the ideas about invariance and interventions described above provide a natural way of understanding at least some non-recursive models. In particular, we can apply the ideas about modularity developed above to non-recursive as well as recursive systems. If the system (13)–(14) is modular, then the result of an intervention on X_3 will be to replace (13) with the Equation (13*) $X_3 = k$ while leaving (14) unaffected. Similarly an intervention on X_4 will replace (14) with Equation (14*) $X_4 = k'$ while leaving (13) intact. As before, if the system (13)–(14) is a correct causal description, each individual equation will correctly describe how its dependent variable will respond to interventions on its right hand side variables as the other equation is altered or disrupted. Thus an intervention on X_3, which sets its value equal to k, will lead according to the undisturbed Equation (14) to a value of X_4 that is equal to $X_4 = cX_2 + dk + V$. Thus the intervention on X_3 will help to determine the value of X_4. Consider, by contrast, the reduced form system $[X_3 = eX_1 + fX_2 + U', X_4 = gX_1 + hX_2 + V']$ which is observationally equivalent to (13)–(14). This reduced form system differs from (13)–(14) in its predictions about what will happen under some hypothetical interventions. For example, (13) predicts that an intervention on X_3 will change X_4 while the reduced form denies this. Of course, as emphasized above, this sort of thought experiment will only make sense or yield well-defined predictions if the two equations represent distinct mechanisms that can be changed independently of each other. In the supply and demand interpretation of (13)–(14), it is arguable that this condition is satisfied. As Thomas Rothenberg puts it in an entry on "Simultaneous Equations Models" in the New Palgrave Dictionary which connects the ideas of independent changeability of equations, distinctness of mechanisms, and the notion of autonomy:

> In the supply-demand model, it is easy to contemplate changes in the behavior of consumers that leave the supply curve unchanged. For example, a shift in tastes may modify demand elasticities but have no effect on the cost conditions of firms. In that case the supply curve is said to be autonomous with respect to this intervention in the causal mechanism [described by the demand curve]. (1987, 233)

It is also worth noting that, just as with hierarchical systems with correlated errors, there are techniques (e.g., two stage least squares, as well as the use of exclusion restrictions and information about the details of the error variance-covariance matrix) other than OLS that, under the right circumstances, permit estimation of the coefficients in non-recursive systems. This again seems to reinforce the point that the uncorrelatedness assumptions have more to do with OLS estimation than with causal interpretation.

Parallel remarks apply to the notion of identifiability. While all recursive models are identifiable, some non-recursive models are not: the statistical data generated by the model may not allow us, even in principle, to reliably estimate the coefficients in the model. Instead the values of the coefficients will be underdetermined by all such information. A very simple example of such a model is the following structure

$$(15) \quad Y = aX$$

$$X = bY$$

Here the statistical data generated by this model does not allow us to estimate the coefficients a and b. (Intuitively, this is because all sorts of different combinations of the coefficients a and b will generate the observed association between X and Y.) A nonidentified model is certainly a disappointment to the investigator, but there is no good reason to suppose that such models are impossible or lack a coherent causal interpretation. To say that the non-identified model correctly describes some system is just to say that the causal structure of the situation (at least as matters stand at present) is such that the statistical associations produced among the measured variables are not such that we can use them to fully discriminate among different possible models of the structure. As emphasized above, this is not to say that no possible observations can discriminate among the models, since different models will always make different predictions about what will happen under some hypothetical changes or intervention, but rather simply that in the absence of such changes, we cannot discriminate among them. For example, in the case of the model (15) immediately above, with X representing quantity demanded or supplied and Y price, it is well-known that if we could observe shifts in the supply curve (the first equation) caused by some factor which influences supply but not demand, such as rainfall, we could identify the coefficient in the second equation. (That it is at least possible that changes in the supply curve should be produced by factors that do not at the same time influence demand is of course bound up, in the way described above, with the idea that the

supply and demand equations describe distinct mechanisms.) It would be an extravagant kind of operationalism to suppose that a model like (15) only becomes meaningful if such a shift actually takes place or only when changes occur that allow us to estimate the coefficients in (15). This gets things backwards – it is because a model like (15) already has a coherent causal interpretation which differs from the interpretation associated with various observationally equivalent alternative models that it makes sense to ask what sorts of data would discriminate between it and these alternatives and under what circumstances nature will produce this data.

4.5. *Invariance and Uncorrelatedness*

Next let me turn to Cooley and Leroy's contention that for the notion of invariance under interventions to be well-defined, exogenous variables must be uncorrelated. The argument that they give in support of this claim has recently been sympathetically discussed by Cartwright (1995) and is worth quoting in some detail. They write:

> Suppose that we start with the model
>
> (16) $M = v$
>
> (17) $P = aM + u,$
>
> where u and v are correlated. If the analyst were willing to assume that the correlation between v and u occurred because v determines a component of u, i.e.,
>
> (18) $u = bv + w$
>
> then (16)–(17) could be rewritten as
>
> (19) $M = v$
>
> (20) $P = aM + bv + w,$
>
> with v and w uncorrelated. Since M is exogenous in this setup, the effect of a change in M on P is well-defined: $dP = (a + b)dv$. Here $a + b$ could be estimated by regression; a and b, of course, are not separately identified.
>
> If, on the other hand, the analysts were willing to specify that the correlation between v and u of (16)–(17) owes to a causal link in the reverse direction, the system would be rewritten
>
> (21) $M = e + cu,$
>
> (22) $P = aM + u,$
>
> with u and e uncorrelated. Now the question 'What is the effect of M on P? is not well-posed, since the answer depends on whether the assumed shift in M is due to an underlying change in e (in which case the answer is $dP = ade$) or in u (in which case the answer is $dP = (ac + 1)du$). (1985, 291–2, I have renumbered the equations and changed some symbols)

The conclusion they draw from this is that

> the notion of exogeneity involves the idea of intervention: a change in an exogenous variable is envisaged, and the effect of this intervention on the endogenous variables is calculated. Exogeneity also involves the idea of invariance under intervention: a cet. par. [ceteris paribus] assumption is made, and this restriction must have unambiguous meaning so that the hypothesized intervention is clearly defined. In order that the required invariance under intervention have unambiguous meaning in all contexts, particularly large systems, the assumption that all exogenous variables be uncorrelated is required. (1985, 292–3)

If the ideas defended in Sections 1 and 2 are correct, this argument is fundamentally misguided. In my view, when the analyst writes down equations like (20) or (22), intending that they be understood as a causal claims rather than as mere claims about patterns of association, their causal interpretation does not depend on whether the correlation between v and u occurs because v determines a component of u or vice-versa. What one means by "the effect of a change in M on P" is just "the change in P that would result from an intervention on M" and this effect is the same and is equally well-defined for both (20) and (22). In both cases, the effect on P of a change in M is given just by the coefficient a – i.e., if the change in M is dM, the effect of this change on P is adM. To see this, note that an intervention on M disrupts equation (19), replacing it with an equation in which the value of M is determined by the intervention or, what is the same thing breaks the arrow directed into M from v in the associated graphical representation. However, the value of v itself is not changed by this intervention and the value w is unchanged as well. As a result, according to (20), the only change in P is due to the change in M and is reflected in the value of a. In the case of (21)–(22), an intervention on M similarly breaks arrows directed into M from but leaves the values of these variables themselves unchanged. Hence the change in P due to a change dM in M is again just adM.

The expression that Cooley and Leroy give for the effect of a change in M on P in connection with (19) and (20) – $dP = (a + b)dv$ – is in fact the expression for the total effect of a change in v on M. What (19) and (20) tell us is that v affects P through two different routes – both directly and by affecting M which in turn affects P. This total effect of v on P should be distinguished from the direct effect of M on P which is what the coefficient a tells us about. Similarly, in connection with (21)–(22) the expression ade is the total effect on P of a shift in e which affects P only indirectly by affecting M. By contrast, the expression $(ac + 1)du$ is the total change that would result in P from a change in u, which is the sum of two components – a direct effect from u to P and an indirect effect of u on P through M. Of course these two expressions are different but this shouldn't be alarming since they describe different total effects.

Both of these total effects are different from the direct effect of M on P which is again just adM. Contrary to what Cooley and Leroy suppose, the fact that the total effect of a change in e on P is different from the total effect of a change in u on P does not show at all that the change in P that would result in an intervention on M (which has to do with still another quantity, namely adM) is ambiguous or ill-defined. It is of course perfectly true that if I observe a change in M and want to predict the total change in P that will occur, it makes a difference whether the change in M is due to a change in e or a change in u. However, it is just a mistake to identify the total change in P that it would be reasonable to predict after observing a change in M with the change in P that would result from an intervention on M.[12] In the former case, it is legitimate to "backtrack" – to make use of whatever information the state of M conveys about its causal antecedents – and to then take account of how these antecedents may affect P through some other route besides M. By contrast, because we think of an intervention on M as an exogenous change in M, a change in M that results from an intervention conveys no information about the state of its causal antecedents and hence we cannot reason backward to these antecedents and then forward to the state of P.

The reasoning described above, in which the effect of M on P is specified by the coefficient a, plainly depends on invariance assumptions – for example, that while interventions on M disrupt (16) and (21), they do not disrupt (20) or (22). These are what we earlier called modularity assumptions. However, once we have these invariance assumptions, we do not require further assumptions about the uncorrelatedness of exogenous variables for the invariance assumptions to be well-defined or to give unambiguous answers about what will happen under interventions. Instead it is one of the virtues of the invariance assumptions that they yield well-defined answers to such questions regardless of whether or not the exogenous variables are uncorrelated.

4.6. *Misspecification and Uncorrelatedness*

Finally, let me comment briefly on another argument described above for the significance of the uncorrelatedness assumptions. This is the argument that violation of the uncorrelatedness assumptions indicates a mistake of some kind about causal structure. There are at least two problems with this argument. The first is that the uncorrelatedness assumption ($U1$) will always be violated in nonrecursive models. If such models are sometimes correct, not all violations of the uncorrelatedness assumptions are due to mistakes about causal structure. A second and more fundamental problem is that even in the case of hierarchical models and assuming Reichenbach's principle, it does not follow from the fact that the uncorrelatedness assumptions are violated that the model is mistaken in its claims about causal structure – it may instead be that the model is merely incomplete. For our purposes, a system of equations is mistaken, if for the variables that explicitly figure or are included in the system, it postulates causal relationships that do not exist or fails to postulate relationships that do exist between those variables. A system is incomplete if there are other variables, not included in the system, that are causally related to the variables included in the system and whose causal relationships to the included variables are omitted by the model. An example of a mistaken model is a regression model which represents X as a cause of Y, when in fact X is not a cause of Y. An example of an incomplete model is a regression model (23) $Y = aX_1 + bX_2 + U$ in which X_1, X_2, and U are indeed causes of Y, but in which X_2 and U are correlated because they are joint effects of some additional variable Z which has been left out of the model. As I have argued, in this second case the model need not be mistaken in the causal claims it makes, although we cannot use OLS to estimate it. In the passages quoted above, Cartwright seems to recognize that unexplained correlations need not indicate a mistake about causal structure (she claims only that they "might" have significant implications for causal structure). By contrast, Papineau seems to suppose that a correlated error term must indicate a mistake in causal structure.

Some incompleteness will be a feature of any causal model with correlated exogenous variables – which is to say, virtually all causal models. Consider, for example, a version of the regression model (23) in which U is uncorrelated with X_1 and X_2 but X_1 and X_2 are correlated with each other. Assuming Reichenbach's principle, this model must be incomplete, since there must be some additional common cause or causes, not represented in the model, which accounts for the correlation between X_1 and X_2. Still, it would be crazy to conclude from the fact that X_1 and X_2 are correlated that the model is mistaken in the causal claims it does makes – indeed we can estimate it by OLS, assuming that we know the variances and covariances of Y, X_1 and X_2. Once we recognize the distinction between mistakes and incompleteness, we see that the presence of a correlation between exogenous variables, whether or not one of these is the error term, does not automatically indicate a mistake in causal structure.

4.7. *Two Additional Arguments*

By way of conclusion to this section, I want to comment more specifically on two additional arguments linking assumptions about the uncorrelatedness of the error term to causal structure. The first is due to Gurol

Irzik (1996). He asks us to consider a simple regression Equation (3) $Y = aX + U$, where U is not just the residual but rather has a causal interpretation as the causes of Y in addition to X that have been omitted from the equation. Following an argument of Herbert Simon's (1954), Irzik contends that that if (a) X is uncorrelated with U, (b) Y does not cause X, (c) Reichenbach's principle holds – namely, if X and Y are correlated, then either X causes Y, Y causes X or X and Y have a common cause or causes, and (d) X and Y are correlated, then it follows that (e) X causes Y. While, in fact, (e) does not quite follow from these premises, I believe there is a suitably reconstructed version of the argument that is sound: from (a), (b), (d) and a suitably reformulated version of Reichenbach's principle, (e) does indeed follow.

What does this show? The first thing to note is that argument does nothing to establish that satisfaction of the uncorrelatedness assumption (a) is *necessary* for (3) to have a causal interpretation. At best the argument shows that the conditions (a)–(d) are jointly (not individually) sufficient for it to be true that X causes Y. According to the position adopted in this essay, if (a) holds and (3) correctly describes a causal relationship, then (3) should continue to hold under interventions that change the distribution of U so that it is correlated with X. That is, changing the distribution of U should not change the value of a or the way in which Y responds to an intervention on X. If, on the contrary, changing the distribution of U did result in the change in the value of a or in the way that X responds to interventions on Y, then Equation (3) would be misspecified and not a correct representation of the causal facts. *Thus, rather than satisfaction of* (a) *being a necessary condition for (24) to have an interpretation as a true causal claim, my view is that if (3) only held under* (a), *it would not describe a true causal claim because it would fail to have the right invariance characteristics*. Of course it is true that if U is correlated with X in (3) and one attempts to estimate the value of a by OLS, then, whether or not (3) represents a true causal claim, the resulting estimate will be biased but this shows only that changing the distribution of U (from uncorrelated to correlated with X) will change the OLS estimate of a, not that it will change the value of a itself.

What about the suggestion that, even if the uncorrelatedness assumption (a) is not necessary for causal interpretabilty, it is at least sufficient? Even if we put aside the point that what the above argument shows is that (a) in conjunction with a number of other assumptions is sufficient for causal interpretability and not that (a) alone is, this interpretation of the argument strikes me as misleading. For one thing,[13] Irzik's argument does not generalize to many other contexts in which structural equations are used to make causal claims. Consider again the equation system

(15a) $Y = aX$

(15b) $X = bY$

As already noted, this can be given a natural causal interpretation within the manipulationist framework described above. An intervention on X that changes it by amount dX will disrupt (15b) while leaving (15a) undisturbed; hence Y will change by amount adX. Parallel remarks apply to the results of an intervention on Y. Because there are no error terms in (15a)–(15b) and because in the absence of any interventions, X and Y will be correlated, we cannot appeal to assumptions about the uncorrelatedness of the error term or other variables (even in conjunction with other assumptions) to explain what it is for this system to have a causal interpretation. Instead we must appeal to some other set of ideas (such as the manipulationist framework) to supply this interpretation. However, if we are going to employ the manipulationist framework in this case, why not also employ it in connection with (3)? In short, the interpretive framework defended in this essay is both more general and more unified than any framework which assigns a central role in causal interpretation to assumptions about the uncorrelatedness of the error term.[14]

I turn next to an argument of Nancy Cartwright's. Referring back to Equations (16)–(17), Cartwright claims

> In order to read off experimental results from (16)–(17) we must not only know that v and u are uncorrelated but we must also know that (i) v causes M (ii) u represents all causes of P other than those that operate through M and (iii) neither u nor v causes the other nor do they have causes in common. (1995, 52)

She then observes that (16)–(17) can be rewritten as (16*)–(17*)

(16*) $P = t$

(17*) $M = bP + w$

and shows that if the conditions (i)–(iii) are satisfied for v and u in Equations (16)–(17) they cannot also be satisfied for t and w in Equations (16*)–(17*).

I see no incompatibility between these claims (at least taken in themselves) and the conclusions about the role of the error term defended above. First, at least in this passage, Cartwright does not claim that satisfaction of conditions (i)–(iii) is necessary for (16)–(17) to be interpretable as making

true causal claims. Second, in agreement with my position, Cartwright is explicit that the mere uncorrelatedness of v (or M) and u is not sufficient to "read off experimental results" from (16) and (17) – other assumptions must be satisfied as well. Finally, what is the intuition or motivation that underlies conditions (i)–(iii)? From my perspective, there is an obvious answer: these are the conditions that must be satisfied if v is to count as an intervention variable for M with respect to P. What Cartwright in effect shows is that if M and P are correlated ($a \neq 0$) under an intervention on P, then we may legitimately think of M as causing P or at least that we may assign (16) its obvious interpretation in terms of the outcome of a hypothetical experiment. This is my view as well. Cartwright's proof that (i)–(iii) cannot be satisfied for both sets of equations (16)–(17) and (16*)–(17*) relies on the further assumption that the structure under investigation is acyclic. Relative to this assumption, what she shows (in effect) is that, if it is possible to carry out an intervention on M with respect to P (where this requires among other things that the intervention be uncorrelated with other causes of M) and this intervention changes M, then one cannot also carry out an intervention on P with respect to M that changes M. From my point of view, this is just to say that if M causes P, then P does not cause M and vice-versa.

5.

My concern in this essay has been to illustrate how ideas having to do with hypothetical experiments and invariance can be used to elucidate causal claims in a very specific context – systems of equations of the sort studied in the causal modeling literature. I believe, however, that these ideas also can be used to develop a much more general account of causation and explanation, applicable to many other areas of science. For example, the idea that causal claims should be interpretable as claims about the outcomes of hypothetical experiments fits very naturally with – indeed is an alternative way of expressing – the idea, which I have defended elsewhere, that explanations work by answering "what if things had been different questions" about their explananda (Woodward 1979, 1984, 1997). Similarly, the notion of invariance provides a natural and plausible framework for thinking about causal and explanatory generalizations that contrast with the usual tendency to assimilate such generalizations to "laws of nature". While the notion of law is usually understood to admit just two possibilities (a generalization is either a law or it is "accidental"), invariance comes in degrees. And while philosophical tradition regards laws as exceptionless, a generalization can, as we have seen, be invariant within a certain domain and break down outside of it. For both of these reasons, invariance is a more promising notion for understanding the status of exception-ridden explanatory generalizations in the special sciences than accounts which treat such generalizations as laws of nature (Woodward 1993, 1995, forthcoming).

Finally, the notions of modularity and of the independent changeability of distinct mechanisms also are also widely applicable outside of causal modeling contexts. Consider a block of mass m sliding down an inclined plane in a downward directed terrestial field that produces constant gravitational acceleration g. The influence of the gravitational force on the motion of the block down the incline is given by (24) $mg \sin \theta$ where θ is the angle between the incline and the horizontal. This is opposed by a frictional force that is proportional to the velocity of the block. Gravity and friction are distinct forces, deriving from distinct physical mechanisms or relationships. We accordingly think that it should be possible in principle to change one of these forces without changing the other. For example, we might alter the relationship between the frictional force and the velocity of the block, but not the relationship (24) for the influence of the gravitational field on the block, by greasing the surface of the plane, and we might alter the relationship (24) by moving the block to the surface of the moon without altering the relationship between friction and velocity. When we adopt the usual physical analysis in which the total force on the block is decomposed into distinct components due to gravity and friction, we implicitly respect the ideas about modularity and independent changeabilty described above. We think of the correct decomposition of a total force into components as just the decomposition in which the components can be changed independently of each other or in which the force law for each component is invariant under changes in the other component. Just as in Haavelmo's example, it will be this representation which will be best suited to tracing the results of hypothetical changes and answering "what if things had been different" questions; hence the most perspicuous representation from the point of view of explanation. The notions of hypothetical experimentation and invariance are central to understanding causation and explanation everywhere in science and not just in causal modeling contexts.

NOTES

* Many of the ideas in this paper were worked out and clarified in collaboration with Dan Hausman. A forthcoming co-authored paper (Hausman and Woodward, 'Independence, Invariance and the Causal Markov Condition') explores the implications of the view of causation defended below for the status of the Causal Markov Condition, which has figured

prominently in recent discussions of causation (see, e.g., Pearl 1998c and Spirtes et al. 1993). I am heavily indebted to Hausman for extensive discussion and criticism. I have also received extremely helpful comments from Nancy Cartwright, Gurol Irzik and Judea Pearl. Research for this paper was supported in part by a grant from the National Science Foundation (SBR-9320097).

[1] Very roughly, my view is that whenever a generic causal claim of the form *Cs* cause *Es* is true, there must be *some* relationship between *Cs* and *Es* that is stable or invariant under some range of interventions on C, but which relationship is invariant and under which range of interventions will vary with the causal claim in question. There is no single invariant relationship which is common to all cases of causation and for this reason it is simply a mistake to look for a single "invariance condition" which is necessary and/or sufficient for the truth of all causal claims. The claim that *Cs* cause *Es* is a highly generic, non-specific claim covering many possible more specific relationships between C and E. In general, one can spell out the content of a causal claim by being more specific about which relationship is claimed to be invariant and over what range of interventions.

As an illustration, consider the probabalistic theories of causation favored by many philosophers. These typically focus on causal relationships between dichotomous variables or at least variables that are measurable only on a nominal scale. It is also assumed that there is some well-defined joint probability distribution for these variables. Obviously, one cannot represent causal relationships between such variables by means of linear relationships like (1), and thus one cannot capture what it is for a relationship between dichotomous variables to be causal by talking about the invariance of (1). However, there are many other possible candiates for invariant relationships involving C and E. If both C and E dichotomous variables taking the values 0 and 1, and if C is the only factor which is causally relevant to E, then the existence of a causal relationship between C and E may show itself in the fact that the conditional probabilities $P(E = 1/C = 1)$ and $P(E = 1/C = 0)$ are not equal and are invariant under some range of interventions that change $P(C)$. Under such conditions, one can manipulate $P(E)$ by intervening on $P(C)$ and hence it makes sense, on the connection between manipulation and causation advocated above, to talk about a causal relationship between C and E. If E has multiple causes $C_1, \ldots, C_n$, the existence of a causal relationship between C_i and E may show itself in the invariance of the conditional probabilities $P(E/C_1, \ldots, C_n)$ under interventions that change $P(C_i)$. Alternatively, it might be the case that these conditional probabilities are not invariant under interventions on $P(C)$ but that the inequality $P(E = 1/C = 1) > P(E = 1/C = 0)$ is invariant or continues to hold under some range of interventions on $P(C)$ —i.e., changing the value C for some units in the population of interest from 0 to 1 always increases the probability of $E = 1$ for those units even though the numerical values for the conditional probabilities are not stable for different units or for different levels of $P(C)$. In this case too, one can manipulate the value of E (or at least the frequency with which different values of E occur) by manipulating the frequency of different values of C, and hence it will be appropriate to think of C as a cause of E, although of course the relationship which is invariant and hence casual will be different from the previous case. In still other cases, the relevant variables may be continuous or interval-valued and may not be governed by any well-defined joint probability distribution or probability density. (Think of fundamental physical laws like Maxwell's equations). Here the framework associated with probabilistic theories of causation will not be applicable but we can still think of such laws as describing invariant relationships and as causal for this reason (cf. Woodward 1992).

[2] For a more detailed treatment of the application of the notion of invariance to explanatory generalizations in the biological and social sciences see Woodward, forthcoming and Hitchcock and Woodward, forthcoming.

[3] We know that it must be possible to specify what it is for an intervention to occur on X without presupposing whether or not it is true that there is a causal relationship between X and Y – if this were not possible we could never learn about causal relationships from experiments.

[4] Under determinism, every intervention or change that does not occur will be "impossible" in the sense that its occurrence will be inconsistent with the laws and actually obtaining initial conditions.

[5] We don't preclude the possibility that a single equation may represent several distinct mechanisms or causal relationships but only the possibility that different equations represent same or non-distinct mechanism(s).

[6] To see that the error term in $X_1 = 1/a_{21}X_2 - 1/a_{21}U_2$ is indeed correlated with X_2, on the assumption that the error term in (12) is uncorrelated with X_1, multiply both sides of this equation by U_2 and take expectations:

$$U_2X_1 = 1/a_{21}U_2X_2 - 1/a_{21}U_2U_2$$

$$E(U_2X_1) = 1/a_{21}(U_2X_2) - 1/a_{21}E(U_2U_2)$$

Since U_2 and X_1 are uncorrelated $E(U_2X_1) = 0$, from which it follows, since $E(U_2U_2)$ is not zero, that $E(U_2X_2)$ must be non-zero.

[7] Cartwright goes on to reject this formulation in favor of a more complex argument connecting causes and probabilities, involving her "Open Back Path Condition" (1989, 29ff.). However, her original formulation is taken up and endorsed by Papineau and Irzik.

[8] Similar arguments and conclusions can be found in Cartwright (1995), and Pearl (1998). Irzik (1996) provides a somewhat different argument for a similar conclusion.

[9] Thus even the epistemological role of the uncorrelatedness requirement in estimation and identification only has a non-trivial application when a regression equation is interpreted causally (cf. Cartwright 1995, 66, for a similar observation).

[10] Although instrumental variables estimators and indirect least squares will not be unbiased, they will be consistent. By contrast, when (U1) is violated OLS estimators are both biased and inconsistent. Hence the former are preferable in such contexts.

[11] Another consideration favoring the use of non-recursive models, recently emphasized by Kevin Hoover (1993), has to do with nature of the variables employed in social science theories: for both conceptual reasons and reasons having to do both with the practicalities of measurement, such variables may not be fine-grained enough temporally to allow us to replace non-recursive models with recursive ones.

[12] This is just the contrast between "conditioning" and "intervening" described earlier.

[13] Another reason why it seems to me to be a mistake to take the argument as showing that (a) in conjunction with other assumptions is sufficient for causal interpretability has to do with the character of these additional assumptions – in particular, Reichenbach's principle (c). I do not deny that when suitably qualified and reformulated some version of Reichenbach's principle is probably correct. The problem rather has to do with understanding what the principle means and the grounds on which one should accept it. In particular, to understand the principle (and to assess whether it is true) it looks as though one must already understand what it means to claim that (i.e., one must already have an

assigned an interpretation to) "X causes Y" since this locution figures in the statement of the principle. But then one must have some basis, independent of the argument Irzik describes, for assigning a causal interpretation to (3). To put the point a bit differently, the claim that the above argument supplies a set of conditions that are jointly sufficient for (3) to have a causal interpretation gets matters backwards. To understand Reichenbach's principle we must first assign an interpretation to the causal locutions that appear on the rhs of the principle. It is only then that we may ask what reasons we have, given this interpretation, to accept the principle. Hausman and Woodward (forthcoming) attempt to do answer this question for a suitable generalization of the principle when "X causes Y" is given the manipulationist interpretation defended above.

[14] There is a related issue which has been raised by Irzik in correspondence and that is worth mentioning at this point. An intervention on X involves an exogenous causal process that breaks the causal connection between X and its previous causes and is uncorrelated with those causes. In appealing to the notion of an intervention to give a causal interpretation to a set of equations don't we thus reintroduce the idea of an uncorrelated additional cause of X which is tantamount to an uncorrelated error term directed into X? In my view the answer to this question is "no". The key point is that the notion of an intervention is a purely hypothetical notion. To say that a relationship is invariant under some intervention is to say that the relationship would continue to hold if the intervention were to occur. It is not necessary that the intervention in question actually occur or that some variable in the system in question actually satisfy the conditions for an intervention variable.Thus, for example, in the case of the equations (15a)–(15b) talking about what would happen to Y under an intervention on X does not commit one to the claim that there must actually be some additional variable U that is directed into X and yet uncorrelated with Y.

REFERENCES

Alderich, J.: 1989, 'Autonomy', in N. di Marchi and C. Gilbert (eds.), *History and Methodology of Econometrics*, Oxford University Press, Oxford.

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

Cartwright, N.: 1995, 'Probabilities and Experiments', *Journal of Econometrics* **67**, 47–59.

Cartwright, N. and Jones, M.: 1991, 'How to Hunt Quantum Causes', *Erkenntnis* **35**, 205–231.

Cooley, T. and S. Leroy: 1985, 'Atheoretical Macroeconometrics: A Critique', *Journal of Monetary Economics* **16**, 283–308.

Frisch, R.: 1938, 'Autonomy of Economic Relations' (unpublished), reprinted in D. F. Hendry and M. S. Morgan (eds.), *The Foundations of Econometric Analysis*, Cambridge University Press, Cambridge, 1995.

Haavelmo, T.: 1944, 'The Probability Approach in Econometrics', *Econometrica* **12** (Supplement).

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

Hausman, D.and J. Woodward: forthcoming, 'Independence, Invariance and the Causal Markov Condition', *The British Journal for the Philosophy of Science*, December, 1999.

Healey, R.: 1992, 'Discussion: Causation, Robustness, and EPR', *Philosophy of Science* **59**, 282–292.

Hitchcock, C. and J. Woodward: forthcoming, 'Explanatory Generalizations Deep and Shallow'.

Hoover, K.: 'Causality and Temporal Order in Macroeconomics or Why Even Economists Don't Know How to Get Causes from Probabilities', *British Journal for the Philosophy of Science* **44**, 693–710.

Humphreys, P.: 1989, *The Chances of Explanation*, Princeton University Press, Princeton.

Irzik, G.: 1996, 'Can Causes be Reduced to Correlations?', *British Journal for the Philosophy of Science* **47**, 259–270.

Meek, C. and C. Glymour: 1994, 'Conditioning and Intervening', *British Journal for the Philosophy of Science* **45**, 1001–1021.

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

Papineau, D.: 1991, 'Correlations and Causes', *British Journal for the Philosophy of Science* **42**, 397–412.

Pearl, J.: 1993, 'On the Statistical Interpretation of Structural Equations', Technical Report, R-200.

Pearl, J.: 1995, 'Causal Diagrams for Empirical Research', *Biometrika* **82**, 669–688.

Pearl, J.: 1996, 'Causation, Action and Counterfactuals', Technical Report, R-223-T.

Pearl, J.: 1998a, 'Tetrad and SEM', *Multivariate Behavioral Research* **33**, 129–148.

Pearl, J.: 1998b, 'Graphical Models for Probabilistic and Causal Reasoning', in D. Gabbay and Ph. Smuts (eds.), *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, Vol. 1, Kluwer, Dordrecht, The Netherlands, pp. 367–389.

Pearl, J.: 1998c, 'Graphs, Causality and Structural Equation Models', to appear in *Sociological Methods and Research*.

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

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

Woodward, J.: 1979, 'Scientific Explanation', *British Journal for the Philosophy of Science* **30**, 41–67.

Woodward, J.: 1984, 'A Theory of Singular Causal Explanation', *Erkenntnis* **21**, 231–262. Reprinted in D. Ruben (ed.), *Explanation*, Oxford University Press, Oxford, 1993.

Woodward, J.: 1992, 'Realism About Laws', *Erkenntnis* **36**, 181–218.

Woodward, J.: 1995, 'Causality and Explanation in Econometrics', in D. Little (ed.), *On the Reliability of Economic Models*, Kluwer, Dordrecht, The Netherlands.

Woodward, J.: 1997, 'Explanation, Invariance and Intervention', in L. Darden (ed.), *PSA 1996*, University of Chicago Press, Chicago.

Woodward, J.: forthcoming, 'Explanation and Invariance in the Special Sciences', *The British Journal for the Philosophy of Science*, March, 2000.

Division of the Humanities and Social Sciences
The California Institute of Technology
Pasadena, CA 91125
USA