Explanation, Invariance, and Intervention

Jim Woodward†‡
California Institue of Technology

This paper defends a counterfactual account of explanation, according to which successful explanation requires tracing patterns of counterfactual dependence of a special sort, involving what I call active counterfactuals. Explanations having this feature must appeal to generalizations that are invariant—stable under certain sorts of changes. These ideas are illustrated by examples drawn from physics and econometrics.

1. Introduction. This essay, which derives from a longer book-length project (Woodward forthcoming b), sketches a set of ideas about explanation and then shows how these are connected to other ideas having to do with invariance and the role of interventions. I begin by setting out a certain conception of what an explanation is—one that emphasizes the importance of tracing patterns of counterfactual dependence. I then argue that this conception leads naturally to two additional ideas: first, that explanatory relations are the sorts of relations that in principle will support manipulations or interventions; and second, that explanatory relations must be invariant relations, where a relation is invariant if it remains stable or unchanged as we change various other things.

As we shall see, the notion of an invariant relationship is interestingly different from the notion of a law of nature—a notion that plays

†Division of the Humanities and Social Sciences, 228–77, California Institute of Technology, Pasadena, CA 91125, jfw@hss.caltech.edu.

‡Research for this paper was supported by the National Science Foundation. I am grateful to Fiona Cowie, Dave Hilbert, Alan Hajek, Kim Sterelny, and especially Nancy Cartwright and Dan Hausman for helpful discussion. I might add that the conception of explanation defended here is in many ways very similar in spirit to the conception in Hausman's forthcoming book, *Causal Asymmetries*, and that in particular my remarks about the explanatory significance of the econometric notion of autonomy closely parallel Hausman's ideas about the role of what he calls independent alterability in explanation.

Philosophy of Science, 64 (Proceedings) pp. S26–S41. 0031-8248/97/64supp-0003$0.00

a central role in many philosophical accounts of explanation. While laws describe invariant relationships, there are many invariant relations that do not correspond to laws, at least if we have any very demanding conception of what a law is. For example, in the social and behavioral sciences one typically finds invariant relationships but not laws. One consequence of the view that I will be describing is thus that one can explain without appealing to laws, as long as one appeals to invariant relations. I will suggest that this provides a more plausible account of how explanation works in areas of science in which there seem to be few plausible candidates for laws, than do more nomothetically based views of explanation.

2. Explanation. In what follows I will assume that, in a suitably broad sense of "causal," all explanation is causal explanation and will use the words "causes" and "explains" interchangeably. I will also focus exclusively on explanations in which what is explained is general—e.g., a regularity or the average value of some quantity within a population—rather than a particular event. This will enable us to avoid certain well-known difficulties having to do with causal preemption and overdetermination that arise in connection with the explanation of particular events. I believe that the account I will be developing can be extended to apply to such cases, but this is an argument for another day.

Consider then an explanation (Ex1) of the magnitude of the electric field created by a long, straight wire with a positive charge uniformly distributed along its length. A standard textbook account proceeds by modeling the wire as divided into a large number of small segments, each of which acts as a point charge of magnitude dq. Each makes a contribution dE to the total field E in accord with a differential form of Coulomb's law:

$$dE = (1/4\pi\varepsilon_0)(dq/s^2)$$

where s is the distance from the charge to an arbitrary point in the field. Integrating over these individual contributions yields the result that the field is at right angles to the wire and that its intensity is given by

$$E = (1/2\pi\varepsilon_0)(\lambda/r)$$

where r is the perpendicular distance to the wire and λ the charge density along the wire.

(Ex1) does exhibit something like the features to which DN theorists have drawn attention: it consists of a deductively valid argument in which a law of nature, in this case Coulomb's law, figures as an essential premise. However, I want to focus on an additional feature that plays no role in the DN model. Put abstractly the feature is this: the gener-

alization(s) cited in (Ex1) are such that they can be used not only to show that the explanandum was to be expected, given the initial and boundary conditions that actually obtained, but are also such that they can be used to show how this explanandum would change if these initial and boundary conditions were to change in various ways. In this way, (Ex1) locates its explanandum within a space of alternative possibilities and shows us how which of these alternatives is realized systematically depends upon the conditions cited in its explanans. As I will put it, (Ex1) can be used to answer a range of what if things had been different questions or counterfactual questions about its explanandum.

Some of this counterfactual information is explicit in the expression for the field intensity E. This expression makes it clear how the field would change if the charge along the wire were increased or decreased or if we were to measure the field at a different distance from the wire and, in this way, exhibits how the field depends on these factors. Moreover, in addition to this, we can use Coulomb's Law and a similar sort of strategy of integrating over the contributions made by small current elements to show how the field would have been different if the long straight wire in (Ex 1) were instead twisted into a circle of finite diameter or coiled up into a solenoid or somehow deformed into or replaced by a sphere or by two uniformly charged plates. For example, using this strategy we can deduce that the field outside a uniformly charged sphere is given by

$$E = (1/4\pi\varepsilon_0)(Q/r^2)$$

where Q is the total charge of the sphere. Such derivations show us that certain factors (e.g., the geometry of the conductor, the distribution of charge density along it, in some cases the distance from the conductor) make a systematic difference to the intensity and direction of the field and that various other factors (e.g., the specific material out of which the conductor is made, whether it is copper or iron, or its mass, or whether it is constructed by experimenter A or experimenter B) are irrelevant. This sort of information has explanatory import because it enables us to see what the field does and does not depend on. It is also crucial for explanatory relevance—a putative explanation like Salmon's birth control pills example (Salmon 1989, 50) that cites a nomologically sufficient but explanatorily irrelevant condition for an outcome will fail to exhibit the pattern described above.

The generalization (Coulomb's law) that figures in (Ex1) is, as I have said, a law of nature. However, as we shall see below (§4), generalizations that are not plausibly regarded as laws also can be used to answer a range of what-if-things-had-been-different questions as long as they have the right sort of invariance characteristics. On the account of

explanation I favor such nonlawful generalization also can be used to provide explanations. That is, I take the provision of information that answers a range of what if things had been different questions to be not just necessary but also sufficient for successful explanation. It is because there are generalizations and patterns of argument that answer such questions without citing laws that nonlawful explanation is possible. The causal models discussed in §5 provide one illustration of this.

3. Counterfactuals and Interventions. The theory I have been sketching thus ties explanatory import very closely to the provision of certain kinds of counterfactual information. But how exactly should the relevant counterfactuals be understood? The need for such an account is apparent when we consider the difficulties surrounding counterfactual theories of causation, for similar problems will face any approach that tries to connect explanation to counterfactual dependence. For example, there seems to be a perfectly good sense in which the joint effects (e.g., the reading of a barometer B and the onset of a storm S) of a common cause (a fall in atmospheric pressure A) are counterfactually dependent on one another, even though one cannot appeal to the occurrence of one effect to explain the other.

Like many other writers,[1] my response to this difficulty is that there is an interpretation of counterfactual dependence according to which the joint effects of a common cause are not counterfactually dependent on each other and that it is this notion that is relevant to explanatory import. My suggestion is that the counterfactuals that matter for explanation are counterfactuals the antecedents of which are made true by a special sort of exogenous causal process that I call an intervention. Heuristically, but only heuristically, we may think of interventions as manipulations that might be carried out by a human being in an idealized experiment. Thus, in the case of (Ex1) what we are interested in is what would happen if we (or some natural process) were to physically intervene to increase the charge density along the wire by connecting it to an appropriate source, or to change its geometry by twisting it into a circle or a solenoid. (Ex1) is explanatory because it tells us how to make the field intensity change, if only we were able to alter the relevant initial conditions—the geometry of the wire, the charge density, and so on—in the right way. A corresponding claim is not true of the barometer reading and the occurrence of the storm. Fiddling with

1. See, for example, Lewis 1973. The account that follows has obvious affinities with Lewis's. In particular, the notion of an intervention plays a role in my account that resembles the role played by "miracles" in Lewis's. But the two accounts differ in motivation and detail.

a barometer dial is not a way of bringing about or suppressing a storm, and this is why the former does not explain the latter.

However, these remarks about the connection between explanation and the results of human manipulation are intended only heuristically. It is *not* part of the theory I am proposing that causal and explanatory dependencies hold only when human intervention is possible. This can be made clearer by being more precise about the notion of an intervention itself. Suppose that I is an intervention on (or manipulation of) the variable X, where X is some property possessed by the unit U_i, the intent being to assess whether X causes or explains some other variable Y by observing whether the intervention on X produces a corresponding change in Y. My suggestion is that, ideally, I should have the following conjunction of features (M).[2]

1. I changes the value of X possessed by U_i from what it would have been in the absence of the intervention and this change in X is entirely due to I.
2. I changes Y, if at all, only through X and not directly or through some other route. That is, I does not directly cause Y and does not change any causes of Y that are distinct from X except, of course, for those causes of Y, if any, that are built into the I-X-Y connection itself; that is, except for (a) any causes of Y that are effects of X (i.e., variables that are causally between X and Y) and (b) any causes of Y that are between I and X and have no effect on Y independently of X. In addition, I does not change the causal relationships between Y and its other causes besides X. Moreover, a similar point holds for any cause Z of I itself—i.e., Z must change Y, if at all, only through X and not through some other route.
3. I is not correlated with other causes of Y besides X (either via a common cause of I and Y or for some other reason) except for those falling under (2a) and (2b) above.[3]

2. For other characterizations of the notion of an intervention that resemble the above characterization but differ in detail, see Spirtes, Glymour, and Scheines 1993, 75–81; Pearl 1995; and Hausman forthcoming.

3. In my view the need for this third condition arises because it is possible for I to be correlated with some other cause Z of Y even though there is no causal connection between I and Z and even though Y and Z have no common cause. I thus reject what Cartwright (1989, 24) calls Reichenbach's principle according to which all correlations have causal explanations—see Woodward forthcoming a for additional discussion. If this principle were correct, clause 3 would be unnecessary. I should also explicitly note that is not part of my view that if X figures in an explanation of Y, an intervention on X that changes Y must be physically possible. That is, the counterfactuals associated with successful explanation will sometimes have physically impossible an-

For reasons of space, I cannot provide a detailed defense of M and confine myself to two brief comments. First, M makes no essential reference to human beings or their activities—instead M is characterized purely in terms of notions like cause and statistical independence. Thus a purely natural process not involving human activity at any point may qualify as an intervention as long as it has the right sort of causal history as described by M. Second, a comment about circularity: the conditions M are obviously themselves stated in causal language and this means that one cannot appeal to the notion of an intervention as part of a reductive account of what it is for X to cause or figure in an explanation of Y. Nonetheless the characterization is not epistemically circular in a vicious sense: one does not already have to know whether X causes Y to determine whether X has been altered by an intervention. Instead, what one needs to know about are the causal characteristics of the process that changes X and about the various other causes of Y besides X and whether these are correlated with I and whether if present they would be causally between X and Y. In particular, we should note that requiring that I change Y if at all only through X is *not* tantamount to requiring that I does change Y through X. M goes along with a nonreductionist conception of causal inference according to which to test some causal claims we must assume the truth of others.

I will call a counterfactual, the antecedent of which is made true by an intervention an *active* counterfactual. In contrast to an "explanation" of the storm in terms of the barometer reading, (Ex1) conveys information about a pattern of active counterfactual dependence. On my view, this is why it is explanatory. (Ex1) has explanatory import because it enables us to see how an intervention on the conditions cited in its explanans would lead to corresponding changes in its explanandum.

4. Invariance. To motivate this idea, recall that the account sketched above requires that the generalization appealed to in an explanation continues to hold as we change in various ways the system whose behavior we are trying to explain. For example, the account requires that Coulomb's Law continues to hold as we alter the charge density or geometry of the wire or its spatial location.

I will say that a generalization that continues to hold or is stable in this way under some class of interventions that change the conditions described in its antecedent and that tells us how the conditions described in its consequent would change in response to these interven-

tecedents. Again, I lack the space to explain why this is more reasonable and less alarming than it sounds.

tions is *invariant* under such interventions. Invariance thus requires stability under interventions although invariant generalizations will virtually always be invariant under changes that are not interventions as well.[4] Obviously the requirement that a generalization be invariant is closely bound up with its ability to support active counterfactuals. The theory of explanation developed above thus requires that explanations appeal to generalizations (or laws or descriptions of dependency relations) that are invariant under some class of changes—in particular, the generalizations cited in an explanation must, at least, be invariant under some class of interventions that change the initial or boundary conditions cited in the explanans of the explanation.

This is interesting for several reasons. First, one finds in many areas of science, and independently of the particular account of explanation described above, the idea that causal or nomological or explanatory relations must be invariant relationships. For example, when given a so-called "active" interpretation, the symmetry requirements physicists expect laws of nature to satisfy are just invariance requirements in the sense described above. Similarly, as we shall see very shortly, in the econometrics or causal modeling literature the notion of a causal or explanatory relationship is identified with the notion of a *structural* or *autonomous* relationship and this in turn is exactly the notion of a relationship that is invariant, in the sense described above, under some relevant class of changes or interventions.

Second, if invariance is ultimately what matters in explanation, this opens up the possibility that a generalization might be invariant (and hence such that it can be used to answer a range of what if things had been different questions) even though we may not wish to regard it as a law. I think that this is exactly what one finds in the so-called special

4. Any generalization, no matter how accidental, will be stable under some changes in background conditions—for example, under changes that are causally independent of the generalization. Thus, special circumstances aside, (1) "All the coins in my pocket are dimes" will be stable under changes in the weather. The demand described above is much stronger than this. It requires invariance under changes that are interventions. In the case of (1) this amounts to the requirement that (1) must be invariant under the introduction of new coins that are not dimes into my pocket. That is, it must be the case that putting a new coin into my pocket will change it into a dime. (1) will fail this invariance test. I also caution that invariance when understood as stability under change is not the same thing as breadth of scope or applicability to a wide range of systems. A generalization can correctly describe the behavior of a wide range of different systems (and can unify or reveal what is common to the behavior of such systems) and yet fail to be invariant in the sense described. In part for this reason, the account I have sketched is not just another version of the claim that explanatory generalizations must have broad scope or must unify.

sciences-relationships that are invariant but that are not plausibly regarded as laws.

5. Invariance in Econometrics. Consider a regression equation of form

$$(2)\ Y = a_1X_1 + a_2X_2 + \ldots + a_nX_n + U$$

Here Y is some dependent variable of interest—e.g., the height of some individual plant, $X_1, \ldots X_n$ are various independent variables—e.g. , X_1 and X_2 are the quantities of fertilizer and water the plant has received, and U is a so-called error term. Y and $X_1, \ldots X_n$ are all measured variables with a joint probability distribution which we can observe. We would like to use this distribution (in conjunction with other assumptions) to infer the values of the coefficients a_i. Discussions of regression and other causal modeling techniques typically distinguish between their use to describe or represent patterns within a body of data and their use to make causal claims or to explain. My suggestion is that the account sketched above captures what this difference consists in. If (2) correctly describes a causal or explanatory relationship then (at least under a certain range of interventions—see below) if an intervention changes X_1 by the amount ΔX_1 then Y ought to change in the corresponding way represented by (2)—i.e., by $a_1 \Delta X_1$—and similarly for interventions on the other variables $X_2 \ldots X_n$. To express the idea a bit more precisely, the equation (2)—its functional form and the coefficients occurring in it—should be invariant under (some range of) interventions that change any of the independent variables X_i. If this condition is met, then (2) will exhibit something like the pattern of systematic active counterfactual dependence that we found in (Ex1) and could be used to answer a range of what if things had been different questions, showing us how changes in fertilizer amount or water would lead to systematic changes in plant height. My suggestion is that as long as one has this pattern of systematic dependence, (2) can be used to explain—even if one does not want to regard it as a law. By contrast, if (2) fails to be invariant under any range of interventions or to support active counterfactuals, it may describe a statistical relationship holding in the data but it will not be explanatory.

There is another important point about invariance and the connection between explanation and counterfactual dependence that is illustrated by (2). This is that invariance is a *relative* notion: a relationship can be invariant under one set of interventions or changes in background conditions (or as I will say under some domain or regime) and not under others. A relationship can thus be invariant within some domain without being exceptionless and universal in the way that, according to many philosophers, laws of nature are. For example, the

relationship (2) may be stable under interventions that change X_1 and X_2 within a certain range—that is, for modest changes in the amount of fertilizer and water a plant receives. However, it is obvious that (2) will not be stable under sufficiently large changes in the values of these variables or under sufficiently dramatic changes in background conditions—for example, heating the plant to 1000°C.

According to the conception of explanation I want to defend, as long as we are within the domain of invariance of a generalization we can use it to explain, even though there may be other conditions outside this domain under which it breaks down and even if these conditions are unknown to us or such that we are unable to characterize them in a theoretically perspicuous way. Thus we can use (2) to explain as long as we are in its domain of invariance even if we are unable to provide a universal exceptionless generalization describing the conditions under which the addition of varying amounts of matter and fertilizer will increase plant height. If this seems an unduly permissive conception of explanation, let me explicitly note that a similar analysis seems appropriate in connection with (Ex1) and most other physical explanations. After all, Coulomb's laws and the other laws of classical electromagnetism also hold only within a certain domain and break down outside it—for example, when distances are sufficiently small for quantum mechanical effects to become important. Most other physical laws exhibit this feature as well. The difference between Coulomb's law and (2) is one of degree rather than one of kind. The notion of invariance, which admits of degrees, and the account of explanation sketched above, are better suited to make sense of these features of both (Ex1) and (2) than are the more traditional ideas that all explanations require laws and that laws describe exceptionless regularities.[5]

6. Invariance and Explanatory Depth. This line of thought leads to a final suggestion: not only can explanatory relationships differ in the range of interventions over which they are invariant but the wider the range over which a relationship is invariant, the deeper the explana-

5. Many philosophers will respond that we should regard (2) as a "qualified" or "ceteris paribus" law and hence that (2) is not a counterexample to the thesis that all explanation requires laws. In part, the issue here is purely verbal; to the extent that it is substantive it has to do with the characteristics of (2) in virtue of which it is explanatory. The traditional view is that (2) is explanatory in virtue of being associated in some way with an exceptionless generalization—for example, in virtue of being a disguised or implicit version of such a generalization. I deny this; my claim instead is that (2) is explanatory in virtue of its invariance characteristics and its ability to support active counterfactuals. A generalization can have these features without being a disguised version of an exceptionless generalization.

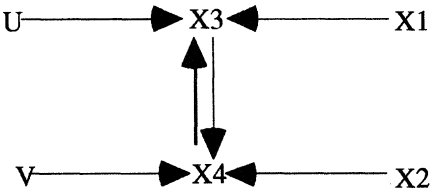
tions in which it can figure. As an illustration of this theme, consider the following system of equations which is taken from (Duncan 1975)

(3.1) $X_3 = b_{31}X_1 + b_{34}X_4 + U$

(3.2) $X_4 = b_{42}X_2 + b_{43}X_3 + V$

These correspond to the following graphical structure.

(3)



It is easy to show that the following set of equations (4.1–4.2) will be observationally equivalent to (3.1–3.2) in the sense that they will imply exactly the same facts about statistical relationships among the measured variables X_1–X_4.

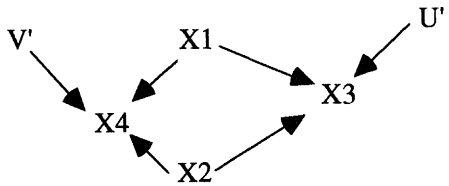
(4.1) $X_3 = a_{31}X_1 + a_{32}X_2 + U'$

(4.2) $X_4 = a_{41}X_1 + a_{42}X_2 + V'$

where (5) $a_{31} = b_{31}/\Delta$, $a_{32} = b_{34}b_{42}/\Delta$, $a_{41} = b_{43}b_{31}/\Delta$, $a_{42} = b_{42}/\Delta$, $U' = U + b_{34}V/\Delta$, $V' = b_{43}U + V/\Delta$, $\Delta = 1 - b_{34}b_{43}$.

These equations correspond to the following graphical structure:

(4)



(4.1–4.2) are the so-called reduced form equations associated with (3.1–3.2). Intuitively, the reduced form equations describe the total effect of a change in each of the exogenous variables X_1 and X_2 on the endogenous variables X_3 and X_4.

If (3) and (4) are observationally equivalent, why should we prefer one rather than the other? One answer, deriving from Tygre Haavelmo (1944) and from Duncan is this: observationally equivalent systems of equations can differ in their degree of invariance or, as Haavelmo and Duncan call it, "autonomy" and relatedly in what they imply about what would happen under various non-actual but possible interventions. That is, while (3) and (4) agree about the actual patterns of statistical dependency, they disagree about what would happen under

various counterfactual possibilities—in particular, those associated with interventions. For example, as we shall see in more detail below, (3) predicts that an intervention on X_3 will also change X_4 while (4) denies this. Models which are more autonomous (more invariant) and which yield accurate predictions about what would happen under a wider range of possible interventions provide better explanations in the sense of answering a wider range of what if things had been different questions and more accurate and perspicuous representations of causal relationships than less autonomous models and are to be preferred for this reason.

How might observationally equivalent systems differ in their invariance characteristics? We have already suggested that if a single equation or system of equations describes an invariant set of relationships then the equations themselves—their functional form and the coefficients occurring in them—should be invariant under (some range of) changes in the values of the variables occurring on the right hand side of each equation. Call this functional form invariance. In addition, there is another invariance condition that it often will be natural to expect an equation like (2) or a system of equations like (3.1–3.2) to meet: it should be possible to intervene to change each of the coefficients in these equations separately without changing any of the other coefficients. Call this condition coefficient invariance.

Applied to (3), the idea is that if (3) correctly represents a system of causal relationships, then each of the coefficients in (3) should be invariant under changes in any of the other coefficients in both equations. To see what this implies, suppose that the correct causal structure is given by (3) and that (3) satisfies coefficient invariance. Then it is fairly easy to see that the reduced form system (4) will not satisfy coefficient invariance. As we can see from the relations (5), each of the coefficients in (4) is a function of several of the coefficients of (3). Furthermore, each of the coefficients in (3) occurs in the expression for several different coefficients in (4). What this means is that if one of the coefficients in system (3)—which we are assuming describes the true causal structure—changes, then several of the coefficients in system (4) must change. If (3) describes the correct causal structure, the coefficients of (4) thus will be entangled with each other—they will not be changeable independently of each other. We will find that if we try to intervene to change one of these coefficients we will change the others as well. Alternatively, suppose instead that (4) represents the correct causal structure and satisfies coefficient invariance. Then a parallel argument will show that (3) cannot be coefficient invariant. More generally, among all of the possible different causal models that are observationally

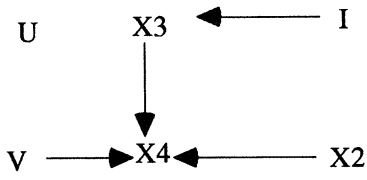
equivalent with each other, at most one will be autonomous in the sense of satisfying of coefficient and functional form invariance.

By requiring coefficient and functional form invariance we thus pick out a unique model from the class of observationally equivalent models. But why is coefficient invariance a reasonable condition to impose? Here is one way of thinking about the motivation for this requirement. When a system of equations like (3) or a single equation like (2) correctly represents a set of causal relationships, then one expects that each term or quantity in these equations should represent a distinct causal mechanism or relationship or at least a distinct quantity that is capable of changing its value independently of the other quantities in the equation and is such that we can associate a distinct causal impact with changes in the value of that quantity. For example, in the case of the single regression equation (2) we suggested that the coefficient a_1 represents the effect of quantity of water on plant height and a_2 the effect of quantity of fertilizer. Then a natural thought is that if this is the correct story about causal structure, it should make sense to think of doing something that just interferes with whatever the mechanism is by which water affects plant height or just changes the relationship between water and plant height while leaving the relationship between fertilizer and plant height undisturbed and vice versa for changes in the relationship between fertilizer and height. In other words, interfering with the relationship between Y and X_1 by altering the coefficient a_1 alone should be an allowable hypothetical experiment, even if it is one that as a practical matter we are not able to carry out. Similarly, if there is a distinct mechanism connecting X_3 and X_1 in (3), it should be possible to change this mechanism without affecting the causal relationships that hold elsewhere in the system. Coefficient invariance thus expresses the idea that causal relationships should exhibit some degree of modularity or context-independence, rather than depending in a holistic way on other relationships and mechanisms holding elsewhere in the system one is trying to model.[6] If a system is representable by a functional form and coefficient invariant set of equations, it will behave in at least some respects like a machine, with independently

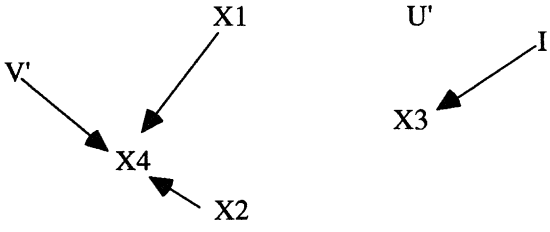
6. To guard against a possible misunderstanding: I am not claiming that only linear or additive relationships can figure in explanations or can be given a causal interpretation. Nor am I claiming that the relationship between height, water, and fertilizer in a real biological system necessarily will be linear or will satisfy coefficient invariance. My remarks are rather meant to illustrate a kind of invariance condition that is sometimes met by systems of linear equations and to illustrate within this context what it might mean to say that one system of relationships has a wider domain of invariance than another. Obviously nonlinear relationships can be invariant and we can appeal to them to explain.

changeable parts. When we can represent a system as a machine in this sense, we have achieved a kind of understanding of its behavior. It is the structure of this sort of understanding that coefficient invariance attempts to capture.

Coefficient invariance is also closely connected to another difference between (3) and (4) to which I referred above: they differ in the claims they make about what will happen under hypothetical interventions. According to (3) an intervention on X_3 will produce a change in X_4 the magnitude of which is indicated by the coefficient b_{43}. By contrast if (4) is correct, an intervention on X_3 will produce no change in X_4. This is because an intervention on X_3, as we have characterized it, is an independent exogenous causal process that changes X_3 but does not act through X_1 or X_2 and is not correlated with them. Because there is no arrow running from X_3 to X_4 i n (4), such a change will not according to (4) produce a change in X_4. As Spirtes, Glymour, and Scheines (1993, 75–81), Meek and Glymour (1994), and Pearl (1995) have recently emphasized, one may think of an intervention in graphical terms as breaking all arrows (besides that associated with intervention itself) directed *into* the variable intervened on (this corresponds to the idea that the value of the variable is now set exogenously by the intervention), while preserving all other arrows including those directed *out* of that variable. Thus if (3) is the correct structure the effect of an intervention on X_3 is to replace the graphical structure associated with (3) with



while if (4) is correct the structure after the intervention is



This suggests yet another rationale for coefficient invariance. Given that (3) is the correct structure, an intervention on X_3 sets the values of the coefficients b_{31} and b_{34} in (3.1) equal to zero, so that the value of X_3

is just set by the intervention. If the intervention is to satisfy the conditions M, this must occur without the other coefficients in (3.2) changing and this is just the idea of coefficient invariance. At least in the context of systems of linear equations like (3) and (4), something like coefficient invariance seems to be required if we are have a well-defined notion of intervention for all variables in the system, including endogenous variables, and hence well-defined and accurate answers to questions about what would happen if we were to change those variables.

By way of conclusion to this section, let me take up an issue that I skirted above. What does it mean to talk of the size of the domain of interventions over which a relationship is invariant? Where does the measure on this domain come from? While I will not attempt to give a completely general answer to this question here, we should note that in many cases, including the ones that I have been discussing, there is a ready basis for comparative judgments of degree of invariance The range of interventions over which the reduced form equations (4) are invariant are a proper subset of the range of interventions over which the structural model (3) is invariant. As a result any properly behaved measure will assign a larger domain of invariance to the latter. We can thus make comparative judgments about the size of domains of invariance and this is all that is required to motivate comparative judgments of explanatory depth of the sort we have been making. More generally, this example illustrates how we may compare the explanatory credentials of competing explanations by appealing directly to the notion of invariance rather than by appealing to the notion of a law of nature.

7. Robustness and Invariance. A good deal of recent philosophical discussion of the connection between invariance and causation has focused on an invariance condition that Redhead (1987) and Papineau (1993) call robustness. Roughly speaking, robustness requires that if A is a probabilistic cause of B then $P(B/A) = P(B/A.D_i)$ and $P(B/-A) = P(B/-A.D_i)$ for different ways D_i of bringing about A. What is the connection between this condition and the invariance conditions described above? In fact they are very different; and the objections that have recently been lodged against robustness as a necessary and /or sufficient condition on probabilistic causation do not apply to the ideas about invariance that I have defended. One way of this is to note that robustness is simply a screening off condition which is definable in terms of the *actual* joint distribution of the values of A, B and D_i in some population. To see that this is very different from the notions of functional form and coefficient invariance described above, note that a large number of different but observationally equivalent systems of equations—like (3) and (4) differing in what relationships they claim

are coefficient and functional form invariant—will fit a given probability distribution of measured variables and will imply exactly the same screening off relations among those variables. Thus the invariance related notions discussed above cannot be defined in terms of screening off, at least when screening off is understood along the actualist lines described above. Instead, functional form and coefficient invariance are *counterfactual* notions—they have to do with the relationships between families of hypothetical probability distributions, with what would change and what would remain unchanged in these distributions if we were to intervene in various ways in the system we are modeling.

As number of critics (e.g., Healey 1992) have noted, there are other problems with robustness as a necessary condition on probabalistic causal relationships. If B has other causes besides A and if the process D_i that changes A also changes these other causes or if D_i affects B directly independently of A, then $P(B/A.D_i)$ will be different for different D_is and will not in general be equal to P(B/A) and robustness will fail.[7] The invariance conditions I have defended avoid this difficulty because they are formulated in terms of the idea of stability of coefficients and functional form under various changes, rather than in terms of the stability of the probability of an effect conditional on one of its partial causes. The counterpart of the Redhead /Papineau robustness condition in the context of an equation like (2) would be the demand that the overall change in Y for a given change in X_i should be the same regardless of how this change is produced. This is obviously an inappropriate demand—a change in X_i that also changes Xj for $i \neq j$ will produce a different overall change in Y than a change in X_i alone. Assuming that (2) genuinely describes a causal relationship, what will be stable under both of these changes is instead the change in Y per unit change in X_i that is attributable to X_i alone, as reflected by the coefficient a_i. (This in turn will be just the change in Y per unit change in X_i when X_i is changed by an intervention.) It would be a mistake to conclude, from the difficulties that surround Redhead's version of robustness, that there are no interesting connections between causation, explanation, and invariance.

7. Is there a way of formulating a robustness condition for stochastic partial causes that avoids these difficulties? One possibility would be to require that P(B/A) and P(B/−A) be invariant under different *interventions* that bring about A and −A. This appears to avoid Healey-type problems but introduces an explicitly counterfactual element into the formulation of robustness. Robustness will no longer be definable in terms of the screening off relations that hold within a single probability distribution, but will rather have to do with the relationship between the different hypothetical probability distributions that would result under interventions on A and −A.

REFERENCES

Cartwright, N. (1989), *Nature's Capacities and Their Measurement.* Oxford: Oxford University Press.
Duncan, O. (1975), *Introduction to Structural Equation Models.* New York: Academic Press.
Haavelmo, T. (1944), "The Probability Approach in Econometrics", *Econometrica* 12 (Supplement): 1–15.
Hausman, D. (forthcoming), *Causal Asymmetries*.
Healey, R. (1992), "Causation, Robustness and EPR", *Philosophy of Science* 59: 282–292.
Lewis, D. (1973), "Causation", *Journal of Philosophy* 70: 556–567.
Meek, C. and C. Glymour (1994), "Conditioning and Intervening", *The British Journal for the Philosophy of Science* 45: 1001–1021.
Papineau, D.(1993),"Can We Reduce Causal Direction to Probabilities?", in D. Hull, M. Forbes, and K. Okruhlik (eds.), *PSA 1992*, v. 2. East Lansing, MI: Philosophy of Science Association, pp. 238–252.
Pearl, J. (1995), "Causal Diagrams for Experimental Research", *Biometrika* 82: 669–710.
Redhead, M. (1987), *Incompleteness, Nonlocality, and Realism: A Prolegomenon to the Philosophy of Quantum Mechanics*. Oxford: Clarendon Press.
Salmon, W. (1989), *Four Decades of Scientific Explanation*. Minneapolis: University of Minnesota Press.
Spirtes, P., C. Glymour, and R. Scheines (1993), *Causation, Prediction and Search*. New York: Springer-Verlag.
Woodward, J. (forthcoming a), "Causal Independence and Faithfulness", *Multivariate Behavioral Research.*
Woodward, J. (forthcoming b), *Explanation, Invariance and Intervention.*