Implementation-Neutral Causation

Stephen F. LeRoy
University of California, Santa Barbara

July 25, 2015

Abstract

The most basic question one can ask of a model is "What is the effect on variable y_2 of variable y_1?" Causation is "implementation neutral" when all interventions on external variables that lead to a given change in y_1 have the same effect on y_2, so that the effect of y_1 on y_2 is defined unambiguously. Familiar ideas of causal analysis do not apply when causation is implementation neutral. For example, a cause variable cannot be linked to an effect variable by both a direct path and a distinct indirect path. Discussion of empirical aspects of implementation neutrality leads to further unexpected results, such as that if one variable causes another the coefficient representing that causal link is always identified.

Keywords: causation, implementation neutrality, Cowles, Simon, interventions

Department of Economics, University of California, Santa Barbara, CA 93106, USA. Email address: leroy@ucsb.edu. web address: econ.ucsb.edu.

The most basic question one can ask of a model is "According to the model, what is the effect on variable y_2 of an intervention on variable y_1?", where y_1 and y_2 are two variables determined by the model. Two answers are possible. The first involves observing that many possible interventions on the model's external variables could have led to the assumed change in y_1, and in general the effects of these interventions on y_2 are different. Therefore the question "What is the effect of y_1 on y_2?" does not have an unambiguous answer: the information given about the intervention—its effect on y_1—is insufficient to characterize its effect on y_2.

The second answer is that even though the intervention is not completely characterized for the reason just noted, all interventions consistent with the assumed change in y_1 may map onto the same change in y_2. In that case the question "What is the effect of y_1 on y_2?" has a well-defined answer. In linear systems, to which our attention will be restricted in this paper, the effect is captured by a single constant, here labeled a_{21}. This coefficient gives the effect on y_2 of a unit change of y_1, regardless of what intervention on the external variables caused the change in y_1.

If, as in the second case above, the effect of a change in y_1 on y_2 is independent of how the change in y_1 is implemented—in other words, independent of the specific interventions on the external variables that determine the assumed change in y_1—we will say that the causation of y_2 by y_1 is *implementation neutral*, and will write $y_1 \Rightarrow y_2$. Hereafter implementation-neutral causation is abbreviated IN-causation, so that y_1 IN-causes y_2 in the specified circumstance. If the implementation neutrality condition fails we will say that y_1 causes y_2, but does not IN-cause y_2. In that case different interventions on the determinants of y_1 have different effects on y_2, implying that we cannot characterize the effect of y_1 on y_2 without knowing more about the intervention. "Knowing more about the intervention" amounts to redirecting the discussion from the causal dependence of y_2 on y_1 to the causal dependence of y_2 on the determinants of y_1.

If we know only that y_1 causes y_2—that is, if we do not have implementation neutrality—we know that interventions that affect y_1 also affect y_2, but we cannot identify a unique coefficient that gives the effect of y_1 on y_2. For many—arguably, most—scientific purposes it is useful to have implementation neutrality, so as to know that the effect of y_1 on y_2 does not depend on what caused the change in y_1.

Despite the foregoing argument, it is not necessarily essential that the causation represented in a model be implementation neutral. However, in

interpreting the model it is necessary to know whether a particular causal link is implementation neutral. If so one can make quantitative statements about causality: "The training program results in an 11 per cent increase in employment probability". Here the subtext is that this is so regardless of the fact that different trainee candidates do or do not enter the program for different reasons. In the absence of implementation neutrality the associated statement is "The extent to which the training program affects a worker's employment probability depends on his individual circumstances, so that the effect of training on the employment probability depends on why he enrolled", for example. Obviously the first answer is preferable when it is available, so it is important to know when a model supports that answer.

Use of diagrammatical methods in causal analysis has become widespread in recent years, due to work by Pearl (2001), Spirtes, Glymour and Scheines (1993), Woodward (2003), Hausman (1998), Cartwright (2007) and others. These authors do not include implementation neutrality in their definition of causation (at least not explicitly; see discussion below). As we will see, implementation-neutral causation is antisymmetric, so it can be used to define directed acyclic diagrams of the type in common use. Therefore one has the option of imposing implementation neutrality in the derivation of directed acyclic diagrams and comparing the causal diagrams so derived with diagrams obtained under characterizations of causation that do not impose implementation neutrality.

1 Characterization of IN-Causation

A distinction that is central in any model that deals with issues of causation is that between internal and external variables. *Internal variables* are those determined by the model, while *external variables* are those taken as given; that is, determined outside the model.[1] We will use y to denote internal variables and x to denote external variables.

[1] In the earlier literature the terms "endogenous" and "exogenous" were often used in place of "internal" and "external". The earlier usage is consistent with the etymology of the terms, but econometricians have implemented a change in their meaning (see Granger (1995)). To avoid ambiguity, economists now use "internal" and "external" when the earlier meaning is intended, as here.

For discussion of various definitions of exogeneity and endogeneity see Leamer (1985). For a statement of the definition of exogeneity currently favored by econometricians see Engle, Hendry and Richard (1983).

All changes in solution values of internal variables are assumed to be attributable to interventions on external variables, as opposed to alterations of equations. Implementing this attribution requires the analyst to be explicit about which hypothetical alterations in the model are permitted and which are ruled out, a specification that is essential in inquiries dealing with causation. Of course, the analyst can always model a shift in any of the equations of the model simply by specifying that the relevant equation includes an external shift variable. In that case the shift variable is a cause of any internal variable that depends on it. Doing so is not the same as converting one of the internal variables to an external variable, which constitutes an alteration of the model, and which, as discussed below, we will avoid.

External variables are assumed to be *variation free*: that is, the analyst is free to alter them independently. Independent variation corresponds to the assumption that by definition external variables are not linked by functional relations; otherwise they would be classified as internal.

The *solution form* of a model expresses each internal variable as a function of the set of external variables that determine it.[2] We will refer to the set of external variables that determine any internal variable as its *external set*, and will denote the external set for y_i as $\mathcal{E}(y_i)$. In examples we will adopt the convention that the external set for any internal variable consists of at least two external variables.[3]

There is no difficulty in defining causation when the cause variable is external: x_1 causes y_1 whenever x_1 is in the external set for y_1. In that case, by virtue of linearity, a unique constant b_{11} gives the effect of a unit change in x_1 on y_1 for any values of the external variables. If x_1 is not in the external set for y_1 the former does not cause the latter.

The ambiguity comes up when the cause variable is internal, because then an assumed change in the cause variable could come from interventions on any or all of the variables in its external set, and in general the effect on y_2 of

[2] We thus distinguish between the solution form and the reduced form, in which current-date internal variables are expressed as functions of lagged internal variables and external variables. In static models the solution form and reduced form coincide.

[3] Otherwise the internal variable is a rescaling of the external variable (assuming linearity); a model containing an internal variable the external set of which consists of one external variable can be simplified by deleting the internal variable.

Also, we will assume below that internal variables are observable and external variables are not. Allowing equations in which the external set of some internal variable consists of a single variable would raise the question whether that variable is observable or unobservable.

the interventions of the external variables of y_1 is different for each possible set of interventions. This is so even if all the contemplated interventions on external variables are restricted to have the same effect on y_1. Given this ambiguity, we cannot associate causation with a single number giving the effect of y_1 on y_2: the intervention is not described with sufficient detail to generate a clear characterization of the effect.

However, consider a special case in which two conditions are satisfied. These conditions involve two internal variables, y_1 and y_2, their external sets $\mathcal{E}(y_1)$ and $\mathcal{E}(y_2)$, and the functions relating the former to the latter. The first is the *subset condition*, which requires that the external set for y_1 be a proper subset of that of y_2. The subset condition guarantees that any external variable that affects y_1 also affects y_2, but not vice-versa. Hoover (2001) in particular emphasized this condition, which assures the antisymmetry of causation.[4] If the subset condition is satisfied we will say that y_1 causes y_2, and will write $y_1 \rightarrow y_2$.

The second is the *sufficiency condition* (the definition of which presumes satisfaction of the subset condition). The sufficiency condition states that the map from $\mathcal{E}(y_2)$ to y_2 can be expressed as two functions. The first is the composition of a function from $\mathcal{E}(y_1)$ to y_1 and a function from y_1 to y_2, while the second is a function from $\mathcal{E}(y_2) - \mathcal{E}(y_1)$ to y_2. If such functions exist then y_1 is a sufficient statistic for $\mathcal{E}(y_1)$ for the purpose of determining y_2, meaning that for the purpose of determining y_2 an intervention on any or all of the variables in $\mathcal{E}(y_1)$ is adequately characterized by the resulting induced change in y_1. If $y_1 \rightarrow y_2$ and in addition the sufficiency condition is satisfied, we will say that y_1 IN-causes y_2, and will write $y_1 \Rightarrow y_2$.

The theme of this paper is that for evaluation of the magnitude of causal

[4] In this paper the subset condition is a condition we impose on models to assure that causation is antisymmetric. Hausman (1998, Ch. 4) had a different take on what we call the subset condition. Hausman's *independence condition* states that "if a causes b ..., then b has a cause that is distinct from a and not causally connected to a." Hausman appears to view the independence condition, not as an assumption in a model, but as a proposition about the world that may or may not be true: "As a metaphysical claim about patterns of lawlike connections found in nature, [the independence condition] seems incredible, and its truth miraculous." (p. 64).

However, he went on to consider another possible interpretation, that the failure of the independence condition implies only that there may exist lawlike relations in the world that are not specifically causal relations. This is so because causality inherently involves antisymmetry, and antisymmetry may not occur if the independence condition fails. This latter interpretation is closer to the position taken here.

effects one is interested primarily in IN-causation (requiring satisfaction of both the subset and the sufficiency conditions), not just causation (requiring satisfaction of only the subset condition).

In the *IN-causal form* of a model the equations are written so as to reflect the model's IN-causal structure. Starting from the solution form of the model and having in hand a set of restrictions on the parameters of that model, one can readily derive its IN-causal form. First one derives the IN-causal ordering, which consists of determining for each i and j whether or not we have that y_i is a parent of y_j.[5] In the IN-causal form of the model, as in the solution form, each equation has one of the internal variables on the left-hand side. The equation for each internal variable y_j that has no internal variables as causal parents coincides with the corresponding equation in the solution form of the model (that is, consists of a map from $\mathcal{E}(y_j)$ to y_j). The causal form for internal variables y_j that have one or more internal variables as causal parents consists of a map from the parent, or from each of the parents, to y_j, plus a map to y_j from the elements of $\mathcal{E}(y_j)$ that are not in the external sets of any of the parents of y_j.

In the linear setting assumed here the equations of the causal form can be written in the form

$$y_j \Leftarrow a_{ji}y_i + b_{jk}x_k. \tag{1}$$

Here y_i is the (single, in this case) internal variable that is a parent of y_j, and x_k is an external variable (again, single) that is the only element of $\mathcal{E}(y_j) - \mathcal{E}(y_i)$. Here and throughout the coefficients a_{ji} and b_{jk} are constants. The cases in which y_j has more than one parent, or in which $\mathcal{E}(y_j) - \mathcal{E}(y_i)$ contains more than one external variable, are handled by expanding (1) appropriately. Note our substitution of $\Leftarrow$ for $=$; since IN-causation is irreflexive and antisymmetric it is inappropriate to use the equality relation in writing the causal form of a model, as many analysts have observed. Also, it is convenient to have notation the distinguishes the causal form of a model ($\Leftarrow$) from its structural form ($=$).

In models that are structural (in the sense that every internal variable is written as a function of the other internal variables and a subset of the external variables) it may or may not be true that the IN-causal form co-

[5] External (internal) variable x_1 (y_1) is an *ancestor* of internal variable y_2 if $x_1 \Rightarrow y_2$ ($y_1 \Rightarrow y_2$). It is a *parent* of y_2 if it is an ancestor and in addition there is no internal variable y_3 such that $x_1 \Rightarrow y_3 \Rightarrow y_2$ ($y_1 \Rightarrow y_3 \Rightarrow y_2$).

incides with the structural form. To determine whether a given structural model can be interpreted as an IN-causal model the analyst (1) computes the solution form of the model, (2) determines its IN-causal ordering by checking whether for all i, j the conditions for y_i to be a parent of y_j are satisfied, and (3) constructs the indicated causal form. If one ends with the same model that one began with, causation in the assumed structural model is implementation-neutral. In that case for each equation each right-hand side variable IN-causes the left-hand side variable. If not, one cannot necessarily interpret parameters of structural models as measuring IN-causation. For example, a structural model with simultaneous blocks can obviously not be interpreted as a model in causal form due to the antisymmetry of causation.

One can represent the causal form of a model by a causal diagram. For variables y_i without internal variables as IN-parents this consists of arrows drawn to y_i from each element of $\mathcal{E}(y_i)$, as in a diagram of the solution form. For variables with internal IN-parents the arrows run to y_j from the IN-parent(s) of y_j, and also to y_j from each variable that is an element of the external set of y_j but is not in the external sets of any of its IN-parents. Thus the IN-causal diagram corresponds exactly to the model written in IN-causal form.

Observe that under our characterization the IN-causal form does not include as arguments internal variables that are ancestors of some internal variable when these are not also parents. The corresponding convention applies to causal diagrams: no arrow directly connects variables with their ancestors when these are not direct parents. If, contrary to this specification, y_1 were entered as a separate cause for y_3 in a causal model that has $y_1 \Rightarrow y_2$ and also $y_2 \Rightarrow y_3$ the effect would be to link each element of $\mathcal{E}(y_1)$ to y_3 via both the direct effect a_{31} and the indirect effect $a_{32}a_{21}$. But we have $a_{31} = a_{32}a_{21}$, so the outcome would be a doubling of the coefficients linking elements of $\mathcal{E}(y_1)$ with y_3. This is an obvious error.

The argument just stated implies that an internal variable never has both an indirect IN-causal effect on another variable via an IN-causal chain involving one or more third variables, and also a distinct direct IN-causal effect; rather, the direct effect is always the composition of the indirect effects. In Section 3 we will point out that a different formalization of causation, that of Simon, does not share this property.

Examples will make these results clear.

1.1 Examples

Consider the following model written in solution form:

$$y_1 = b_{11}x_1 + b_{12}x_2 \tag{2}$$
$$y_2 = b_{21}x_1 + b_{22}x_2 + b_{23}x_3. \tag{3}$$

The external sets for y_1 and y_2 are $\mathcal{E}(y_1) = \{x_1, x_2\}$ and $\mathcal{E}(y_2) = \{x_1, x_2, x_3\}$. The former is a proper subset of the latter, so the subset condition is satisfied, and we have $y_1 \rightarrow y_2$.

Without parameter restrictions the sufficiency condition for $y_1 \Rightarrow y_2$ is not satisfied. However, if the condition

$$b_{21}/b_{11} = b_{22}/b_{12} \tag{4}$$

obtains the sufficiency condition is satisfied. In that case we can define a_{21} by

$$a_{21} \equiv b_{21}/b_{11} = b_{22}/b_{12}, \tag{5}$$

allowing replacement of (3) with

$$y_2 = a_{21}y_1 + b_{23}x_3. \tag{6}$$

We have $y_1 \Rightarrow y_2$.

The IN-causal form of the model is

$$y_1 \Leftarrow b_{11}x_1 + b_{12}x_2 \tag{7}$$
$$y_2 \Leftarrow a_{21}y_1 + b_{23}x_3. \tag{8}$$

The argument just presented implies that in the structural model

$$y_1 = b_{11}x_1 + b_{12}x_2 \tag{9}$$
$$y_2 = a_{21}y_1 + b_{23}x_3 \tag{10}$$

the coefficient a_{21} represents IN-causation. This is so because its structural form (9)-(10) coincides with its IN-causal form (7)-(8).

The upper panel of Figure 1 shows the causal diagram of the model under discussion if the restriction (4) is satisfied; the lower panel shows the causal diagram if the restriction is not satisfied.

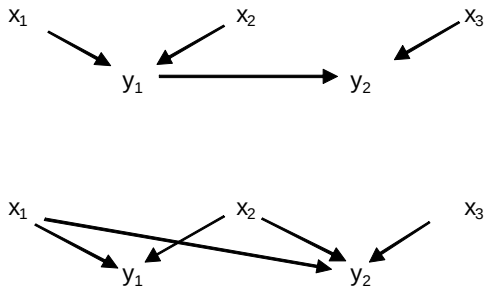

Figure 1

Upper panel: y_1 causes y_2. Lower panel: y_1 does not cause y_2 due to failure of the suffiency condition

As observed above, one can equally well begin by specifying a model in IN-causal form, as in (7)-(8). Using (2) to eliminate y_1 in (6) results in

$$y_2 = a_{21}b_{11}x_1 + a_{21}b_{12}x_2 + b_{23}x_3. \tag{11}$$

Comparing this equation with the solution equation (3) for y_2 results in $a_{21}b_{11} = b_{21}$ and $a_{21}b_{12} = b_{22}$, agreeing with (5). Thus writing a model in IN-causal form is equivalent to assuming the parameter restrictions on the solution form associated with the assumed causal ordering.

One cannot write down an arbitrary structural model and then interpret that model as if it were in IN-causal form. Some models that are acceptable as structural models are inadmissible as IN-causal models. For example, consider the model

$$\begin{aligned} y_1 &= a_{12}y_2 + b_{11}x_1 && (12)\\ y_2 &= a_{21}y_1 + b_{22}x_2 && (13)\\ y_3 &= a_{31}y_1 + a_{32}y_2 + b_{33}x_3. && (14) \end{aligned}$$

This is an acceptable block-recursive structural model, but not an acceptable IN-causal model because as such it contains both $y_1 \Rightarrow y_2$ and $y_2 \Rightarrow y_1$, violating the antisymmetry of IN-causation. The conclusion is that the model (12)-(14) is not in fact an IN-causal model.

Generically (that is, barring coefficient restrictions), the IN-causal form of the model (12)-(14) is

$$
\begin{aligned}
y_1 &\Leftarrow b_{11}x_1 + b_{12}x_2 && (15)\\
y_2 &\Leftarrow b_{21}x_1 + b_{22}x_2 && (16)\\
y_3 &\Leftarrow b_{31}x_1 + b_{32}x_2 + b_{34}x_4 + b_{33}x_3, && (17)
\end{aligned}
$$

coinciding with the solution form. The causal ordering is empty (in the sense that none of the internal variables IN-cause any other internal variables). Therefore if one begins with a model like (12)-(14) that is not interpretable as an IN-causal model one cannot view all the coefficients a_{ij} of that model as measuring IN-causal effects, although some may do so.

The causal form of any model is recursive by construction. It might be thought that the converse is also true, so that all structural models that are recursive (which the model (12)-(14) is not) would qualify as causal models. This is not so. Triangular models provide a counterexample when they have more than two internal variables. For example, consider the structural model

$$
\begin{aligned}
y_1 &= b_{11}x_1 + b_{12}x_2 && (18)\\
y_2 &= a_{21}y_1 + b_{23}x_3 && (19)\\
y_3 &= a_{31}y_1 + a_{32}y_2 + b_{34}x_4. && (20)
\end{aligned}
$$

This model does not have the IN-causal representation

$$
\begin{aligned}
y_1 &\Leftarrow b_{11}x_1 + b_{12}x_2 && (21)\\
y_2 &\Leftarrow a_{21}y_1 + b_{23}x_3 && (22)\\
y_3 &\Leftarrow a_{31}y_1 + a_{32}y_2 + b_{34}x_4. && (23)
\end{aligned}
$$

This is so because the purported cause variable y_1 is not a parent of the effect variable y_3 (although it is an ancestor), contrary to the requirement assumed for construction of IN-causal models. The model (18)-(20) can be interpreted as a causal model only under restrictions on the structural coefficients (for example, $a_{31} = 0$ or $a_{32} = 0$). In the absence of such restrictions a model like (18)-(20) has a causal ordering consisting only of $y_1 \Rightarrow y_2$, plus the equations relating internal variables to their external sets.

As another way to establish the same point, suppose that we have $y_1 \Rightarrow y_2$ and $y_2 \Rightarrow y_3$. Then implementation neutrality implies that the total effect of an intervention Δy_1 of y_1 on y_3 equals $a_{32}a_{21}\Delta y_1$, and this is so regardless of which element of $\mathcal{E}(y_1)$ caused the change in y_1. This is the total effect of y_1 on y_2 implied by the causal ordering, and it coincides with the indirect effect. There is no distinct direct effect.

2 Critiques of Implementation Neutrality

Philosophers sometimes reject this focus on settings in which causation is implementation neutral. For example, Cartwright (2007) states that "[w]e must be careful ... not to be misled by [LeRoy's] own use of the language of 'causal order' to suppose it tells us whether and how much one quantity causally contributes to another" (p. 246). Why are we misled by this supposition? How much one variable causally contributes to another is exactly what IN-causation tells us, and is exactly what we want to know. And what meaning can we attach to a purported measure of the effects of an intervention on an internal variable if the model is such that the causation is not implementation neutral, so that that measure is not well defined? In that case there is no alternative to redirecting the analysis to implementation-specific interventions on the external variables, avoiding reference to the intermediate variable—the purported cause—which in fact plays no role in the causation.

It is not difficult to find passages in the philosophy literature where the idea of implementation neutrality is implicitly introduced. Further, it is not unusual to find use of the term "causation" reserved to settings in which implementation neutrality is satisfied. For example, Woodward (2007) listed "invariance" among the requirements for causation: the effect of the cause variable on the effect variable should be invariant to interventions on other variables. He observed that "[o]ne condition for a successful intervention is that the intervention I on X [the cause variable] with respect to Y should not cause Y via a route that does not go through X, and that I should be independent of any variable Z that causes Y but not through a route that goes through I and X". If one reads I as consisting of a variable in the external set of X, then Woodward's criterion for a "successful intervention" corresponds to that for our implementation-neutral causation.

Woodward gave an example. Suppose that patients are treated or not

treated for a medical condition based on a randomized assignment mechanism such as a coin toss. So stated, the assignment mechanism is an IN-cause (assuming that the treatment is effective) of remission of the condition. But suppose that another doctor influences the outcome of the coin toss using a magnet, and does so to ensure that patients with a strong immune system get the treatment. This alteration invalidates implementation neutrality. In our terminology the state of the patient's immune system is an external variable for the use of the magnet, and the external set for the use of the magnet is a proper subset of the external set of the variable representing the assignment mechanism. The sufficiency condition for causation of the remission variable by the assignment variable is not satisfied. This is so because the variable representing the strength of the immune condition also affects the remission variable via a direct path.

Critics of the analysis of causation presented here express the view that the conception of IN-causation here unnecessarily departs from the ordinary-language usage of "causation". The opposite is the case. Under the ordinary-language usage of "causation", in settings where the conditions for IN-causation fail the answer to the question "What is the effect of y_1 on y_2?" would be "It depends on what causes the variation in y_1". This coincides exactly with the usage prescribed in this paper.

3 Comparison with Simon

It is instructive to compare the representation of causation just presented to that of Simon's classic (1953) paper. Simon characterized a structural model as a partially ordered set of self-contained sub-models, with some (or all) of the internal variables determined in each sub-model. Each sub-model contains the internal variables determined in that sub-model and, except for the lowest-ordered sub-models, also some or all of the internal variables determined in lower-ordered sub-models. Triangular models, in which each sub-model consists of a single equation, are the most extreme special case. In triangular models a complete ordering is defined on the internal variables, with the explanatory internal variables for each internal variable consisting of internal variables that are lower in the ordering.

Under Simon's definition of causation y_1 causes y_2 if y_1 enters the sub-model that determines y_2, and is determined in a lower-order sub-model. Thus for Simon causation is determined from a model's structural form, as

opposed to its solution form as here. As this difference would lead one to expect, causation according to Simon's definition is not equivalent to either causation as defined above, represented by $\rightarrow$, or IN-causation as defined above, represented by $\Rightarrow$.

Suppose first that $y_1 \rightarrow y_2$ by the definition proposed above, meaning that $\mathcal{E}(y_1)$ is a proper subset of $\mathcal{E}(y_2)$. It does not follow that y_1 causes y_2 according to Simon's definition. This is so because y_1 may not enter the structural equation that determines y_2. For example, in the model

$$y_1 = b_{11}x_1 + b_{12}x_2 \tag{24}$$
$$y_2 = b_{21}x_1 + b_{22}x_2 \tag{25}$$
$$y_3 = a_{31}y_1 + b_{33}x_3 \tag{26}$$

we have satisfaction of the subset condition, implying $y_2 \rightarrow y_3$. However, we do not have that y_2 causes y_3 under Simon's definition of causation because of the exclusion of y_2 from eq. (26).

We also have that $y_1 \rightarrow y_2$ under Simon's definition does not require satisfaction of the sufficiency condition, so that causation under Simon's definition does not imply IN-causation ($y_1 \Rightarrow y_2$). The easiest way to verify this difference is to note that Simon's definition of causation allows indirect and direct causation to coexist (again, as in the model (18)-(20)), whereas under IN-causation this cannot occur, as noted above.

In basing his definition of causation on the structural form of a model, Simon implicitly defined the intervention associated with causation to be conditional on the values of the explanatory variables other than the cause variable in the structural equation determining the effect variable. The intervention so defined can readily be translated into the implied intervention on the variables in the external set for the cause variable. This intervention will involve linear restrictions on the intervention in the external variables, so that some external variables are treated as causing other supposedly external variables. This dependence implies a violation of the variation-free condition, and therefore raises the question of what meaning can be attached to causation so defined.

The model (18)-(20) makes this clear. It is natural to interpret a_{31} as the constant associated with the causal effect of y_1 on y_3, holding constant y_2. However, this leads to problems. A nonzero intervention in y_1 is consistent with a constant value of y_2 only if x_3 is such as to offset the changes in x_1

and x_2 on y_2, via y_1. But then x_3 is implicitly assumed to be a function of x_1 and x_2 (or vice-versa), contrary to the assumption of the model that x_3 is external. Thus in the model at hand it does not make sense to consider such conditional causation. Of course, one is free to redefine x_3 as an internal variable, but then the analysis of causation should be conducted using the modified model, not the model as defined.

Simon's definition of causation differs from that analyzed here in settings where the modeler is willing to specify a structural model that is distinct from the associated solution form, and only in such settings. Defining causation in reference to the structural model is justified, if at all, only insofar as the analyst believes that the structural form is somehow superior to the solution form, in that it contains information that is lost in passing to the solution form. The Cowles economists clearly believed that this was the case, but they never succeeded in articulating clearly what this information is. It is difficult to see why performing arithmetic operations in order to pass from the structural form to the solution form should affect a model's interpretation to the point where causation has a different interpretation in the two cases.

Recognizing this, contemporary economic theorists typically do not specify a structural form distinct from the solution form. Thus the characterization of IN-causation as defined here, being based on the solution form, is consistent with current practice in a way that Simon's treatment of causation is not.

In Section 4 of his paper Simon gave three examples of application of his definition of causation in three models. It is interesting to observe that even though his definition of causation differs from ours in general, in all three examples causation turns out to be implementation-neutral by our definition.

4 Empirical Aspects of Causation

Up to this point we have considered models in which variables are specified as to their status as internal or external. We have not specified which variables are observable or what we are assuming about the probability distributions of unobserved external variables. That we could postpone discussion of observability to this point reflects the fact that, for any pair of internal variables, the existence or nonexistence of IN-causation depends only on whether the conditions for implementation neutrality are satisfied. It does not depend on which variables are observable or what is assumed about those that are

not. However, without specifying which variables are observable and characterizing the probability distribution of unobserved external variables there is no way to estimate IN-causal coefficients empirically: the correlations among internal variables implied by the model's causal structure cannot be disentangled from those induced by correlations among unobserved external variables.

The most direct way to launch an investigation of the empirical aspects of causation is to specify, first, that external variables are unobservable and internal variables are observable. This specification covers most of the cases of interest. Second, it is assumed that the external variables are statistically independent random variables. This assumption implies that whatever correlations exist among the model's internal variables are generated by the equations of the model, not by uninterpreted correlations among external variables. An analyst who is uncomfortable with the assumption that the external variables x_1 and x_2 in two equations are independent can replace x_2 with $x_2 + \lambda x_1$, which allows for correlation even if x_1 and x_2 are independent. Of course, adopting such flexible specifications results in sparse causal orderings. As always, the analyst must deal with a tradeoff between how general a model's specification is and how rich its empirical implications are.

The assumptions just listed imply that if we have $y_1 \Rightarrow y_2$ the IN-causal coefficient measuring the effect of y_1 on y_2 is identified (apart from special cases in which observability is limited, as discussed below), and can be estimated consistently using a least-squares regression of y_2 on y_1. This is so because the external variable(s) in $\mathcal{E}(y_2) - \mathcal{E}(y_1)$—the constituent(s) of the error term in the regression—is (are) independent of $\mathcal{E}(y_1)$, and therefore of y_1 itself. Therefore the conditions for the Gauss-Markov theorem of linear regression are satisfied and least-squares regression coefficients provide optimal estimators.

The contrapositive of this statement is that the existence of econometric problems in the estimation of a parameter implies that the parameter is not one associated with IN-causation. For example, consider the system

$$y_1 = b_{11}x_1 + b_{12}x_2 \tag{27}$$

$$y_2 = a_{21}y_1 + b_{21}x_1 \tag{28}$$

$$y_3 = b_{32}x_2 + b_{33}x_3. \tag{29}$$

Here the external variables x_1, x_2 and x_3 are assumed to be independently distributed. Analysis of the solution form of this model reveals that the

population parameter a_{21} does not equal $cov(y_2, y_1)/var(y_1)$, the population regression coefficient of y_2 on y_1. This is so because y_1 and x_1 are correlated due to the presence of x_1 in the external set for y_1. Therefore a_{21} is not estimated consistently by least squares on (28). Further, if y_3 and (29) are dropped from the model, then a_{21} is not even identified. This can be seen by inspection of the solution form of the model (27)-(28).

However, in the presence of y_3 and (29) we have $a_{21} = cov(y_2, y_3)/cov(y_1, y_3)$, implying that a_{21} is identified and can be estimated consistently by taking y_3 as an instrument. Here we make use of the fact that y_3 is correlated with y_1, due to the common presence of x_2 in their external sets, but not with x_1.

The result that the least-squares estimate of a_{21} is not estimated consistently by least squares reflects the fact that y_1 does not IN-cause y_2, a fact that is also easily verified directly from the definition of IN-causation. Thus the inconsistency of the least-squares estimate of a_{21} via a regression of y_2 on y_1 does not contradict our assertion that coefficients associated with IN-causal orderings are identified and estimable by least squares.

The finding that IN-causal coefficients are always identified differs from the conclusion of the Cowles economists. The reason for the difference is that, as noted, the Cowles economists used a different conception of causation—one that does not include implementation neutrality—than that we focus on here. Parameters that are causal in the Cowles sense may or may not be identified and may or may not correspond to coefficients associated with IN-causation. Here our attention is restricted to the smaller set of coefficients that are IN-causal.

The result that causal coefficients are always identified should not be taken to imply that identification is not a major problem in the analysis of causation. Obviously, there exist coefficients associated with IN-causation only when the associated variables are in fact IN-causally ordered, and whether two variables are IN-causally ordered depends on the coefficients that link observed internal variables to unobserved external variables. These coefficients, in contrast to those linking observed internal variables that are known (or assumed) to be causally ordered, are generally not identified. Therefore there may be no way to directly test models that make particular specifications of causation.

Under causation as characterized here, as with other definitions of causation, the restrictions justifying an assumed causal ordering can in principle be tested indirectly by identifying pairs of variables that are or are not statistically independent according to the model, and then determining whether

these independence implications are satisfied empirically. We now consider whether powerful empirical tests of causal models along these lines are likely to be available. It appears that they are not: only in special cases is it possible to characterize independence or the lack thereof among internal variables as testable implications of IN-causal models.

Among the few results that are available is the obvious fact that any two internal variables for which the external sets are disjoint are statistically independent. As an implication, if an internal variable has two ancestors, then either the two are statistically independent or one ancestor causes the other. To see this, suppose that $y_1 \Rightarrow y_3$ and $y_2 \Rightarrow y_3$, so that y_3 has ancestors y_1 and y_2. If $\mathcal{E}(y_1)$ and $\mathcal{E}(y_2)$ are disjoint, then y_1 and y_2 are statistically independent. Suppose instead that $\mathcal{E}(y_1)$ and $\mathcal{E}(y_2)$ have a nonempty intersection that contains external variable x. Then because (1) $x \in \mathcal{E}(y_1)$, and (2) $\mathcal{E}(y_1)$ is a proper subset of $\mathcal{E}(y_3)$, there exists a path from x to y_3 that includes y_1. Similarly, there exists a path from x to y_3 that includes y_2. These must be the same path, since if the path included y_1 but not y_2 then y_2 could not be a sufficient statistic for $\mathcal{E}(y_2)$, contradicting $y_2 \Rightarrow y_3$. Thus there is a single path connecting x and y_3, and that path includes both y_1 and y_2. This can occur only if $y_1 \Rightarrow y_2$ or $y_2 \Rightarrow y_1$.

Past this there are not many results available about correlation of variables in causal models. Assume that y_1 and y_2 have y_3 as a common ancestor. If also $y_1 \Rightarrow y_2$, then we have $y_3 \Rightarrow y_1 \Rightarrow y_2$. In that case we have that all pairs of these three variables are correlated since their external sets have a nonempty intersection (consisting of the external set for y_3). If, on the other hand, $y_1 \not\Rightarrow y_2$ the causal coefficient associated with $y_1 \Rightarrow y_2$ is not defined. In the absence of IN-causation, no inference about the correlation among variables is possible.

Despite the foregoing discussion, it happens that some of the techniques of diagrammatical analysis developed in the causation literature do carry over in the present setting. For example, it is shown in the received literature that if two internal variables are connected only by paths that are "blocked" because each contains a "collider" (a variable with incoming arrows from both directions), those variables are independent. That result appears to carry over here. An example will demonstrate this.

4.1 Example

Consider the following model:

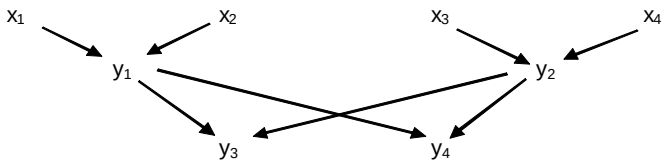

Figure 2

Paths connecting y_1 and y_2 are blocked by colliders y_3 and y_4

$$y_1 = x_1 + x_2 \tag{30}$$
$$y_2 = x_3 + x_4 \tag{31}$$
$$y_3 = x_1 + x_2 + x_3 + x_4 \tag{32}$$
$$y_4 = x_1 + x_2 - x_3 - x_4 \tag{33}$$

(note that here we have supplied specific coefficient values as well as external sets). The causal form of this model is

$$y_1 \Leftarrow x_1 + x_2 \tag{34}$$
$$y_2 \Leftarrow x_3 + x_4 \tag{35}$$
$$y_3 \Leftarrow y_1 + y_2 \tag{36}$$
$$y_4 \Leftarrow y_1 - y_2, \tag{37}$$

with Figure 2 as its causal diagram. Here y_1 and y_2 are statistically independent due to the fact that their external sets are disjoint. We have that y_1 and y_2 are parents of y_3 (and also of y_4), so the result illustrates the general fact noted above that if any internal variable has more than one ancestor, either these are independent or one ancestor causes the other.

This independence result can be generated using the diagrammatical techniques developed by Pearl and others for analysis of causation in settings where implementation neutrality is not imposed. In the example there exist two paths from y_1 to y_2, but both are blocked by the colliders y_3 and y_4. Therefore these paths do not transmit association. Independence of y_1 and y_2 results. Note that here the diagrammatical analysis applies by virtue of the assumption that the external variables are independently distributed. The result suggests that even though the conditions for causation analyzed

here are different from those in the received literature, at least some of the diagrammatical techniques for analysis of causation carry over. This is a topic that deserves further study.

The independence result does not extend to the children y_3 and y_4 except in special situations. For example, if the x_i are normally distributed and all have the same variance, y_3 and y_4 are independent. However, if x_1 and x_2 have higher (lower) variance than x_3 and x_4, then y_3 and y_4 will be positively (negatively) correlated.

5 Conditioning on Internal Variables

The result in the preceding section that the coefficient associated with any causal relation is identified and can be estimated consistently using least squares depends critically on the underlying assumption that external variables are independently distributed and internal variables are fully observable. If some internal variable y_i is observed only when it lies in a certain region, the distribution for the external variables that is relevant for determining the identifiability of causal coefficients is that conditional on this restriction, not the unconditional distribution.

The joint distribution of the external variables conditional on y_i will generally display statistical dependence even if the unconditional distribution of the external variables incorporates independence. This situation will not affect the causal ordering of the variables, but it does invalidate the result that the coefficients associated with the causal ordering can be estimated consistently by least squares. This is so because failure of independence in the external variables implies that the error term covaries with the explanatory variable in the relevant regression, inducing bias and inconsistency.

As an extreme case, suppose that the analyst only has data in which y_i takes on a single value, for some i. Obviously the coefficient associated with $y_i \Rightarrow y_j$ or $y_j \Rightarrow y_i$ for some y_j is not identified, there being no variation in the observed values of the cause variable in one case or the effect variable in the other. A more common situation occurs when the data for y_i are truncated, as by $y_i \geq 0$. In that case the sample regression coefficient associated with $y_j \Rightarrow y_k$ is not a consistent estimate of the associated causal coefficient if either y_j or y_k has an external set that overlaps with that of y_i. This is so because if y_i is subject to a restriction like $y_i \geq 0$ the relevant joint distribution of the external variables in $\mathcal{E}(y_i)$ is that conditional on $y_i \geq 0$,

and this does not generally have any independence property.

Berkson's Paradox illustrates this. Suppose, following Elwert (2013), that movie actors become famous if they are good looking or can act well, or both. Assume, probably realistically, that being good looking and being a good actor are independently distributed. If the analyst has a data set consisting only of actors who are famous, then any actor in that set who is not good looking must be a good actor, since otherwise he would not be famous. Thus in the data set of famous actors there will be a negative correlation between being good looking and being a good actor, even though by assumption there is no such correlation in the general population. Any statistical exercise that makes no allowance for this effect will be biased.

We will not discuss statistical procedures to deal with this problem since the problem does not directly involve causal issues. The point here is only to demonstrate that the attractive statistical properties of least squares in estimating causal coefficients do not apply universally when data on internal variables are not fully observed.

6 Comparison with "Fixing"

The analysis of IN-causation outlined in this paper differs in major respects from what is found in the causation literature. Most important, interventions here consist exclusively of hypothetical alterations in the assumed values of external variables. In contrast, the usual treatment in the literature (based on Haavelmo (1943) and Strotz and Wold (1960)) involves modeling policy interventions on, say, y_1 by deleting from the model the equation determining y_1 and replacing it with the specification that y_1 is external.

This practice of "fixing" internal variables and deleting equations when analyzing interventions seems misdirected. It violates the autonomy assumption (which consists of the assertion that the model equations are invariant to assumed interventions). It does not make sense to claim to analyze interventions using a model if doing so involves changing the model to accommodate the intervention. Fixing corresponds to measuring a person's height using a yardstick that expands or shrinks according to the height being measured.

Fixing internal variables involves a troubling inconsistency between how model solutions are generated in the routine operation of the model—via realizations of external variables—and how they are modeled under a policy intervention—via relabeling internal variables as external and suppressing

equations. What is it about policy interventions that motivates this difference in treatment? We are not told. As suggested above, it seems simpler and more satisfactory to be consistent about carrying over the attribution of assumed interventions on internal variables to underlying changes in the external variables that determine them, and thereby to avoid altering the equations of the model.

Besides this, there are several major problems with modeling interventions by fixing internal variables. Most obviously, doing so applies only in recursive systems, since in the presence of simultaneity y_1 is determined jointly with other variables in a group of several or many equations. In that case there does not exist any obvious way to identify which equations are to be deleted. In contrast, our analysis of IN-causation applies in non-recursive models, although of course IN-causal relations among internal variables are likely to be sparse in models with large simultaneous blocks.

The Haavelmo-Strotz-Wold procedure assumes that causal models are modular, meaning that causal relations can be modified individually without invalidating the other equations of the model (modularity has been discussed widely in the philosophical literature on causation; see, for example, Cartwright (2007) and the works cited there). Under our treatment, in contrast, the question of modularity does not come up because we are not modifying the model.

Modeling interventions by respecifying internal variables as external implies that causation is treated as if it were implementation neutral whether or not this treatment is justified. If implementation neutrality fails coefficients will be interpreted as IN-causal when they do not support that interpretation. It is far from clear why one would want to take this route. In general the answer to the question "What is the effect of y_1 on y_2?" is properly viewed as possibly, but not necessarily, depending on what brings about the change in y_1. The model encodes exactly this information in the equations determining y_1. Therefore the analyst can determine whether the question of causation has an unambiguous answer.

7 Application: Granger Causation[6]

Granger (1969) proposed a definition of causation that can be implemented empirically without relying on theoretical restrictions: a stochastic process (that is, sequence of random variables) $y_1 = \{y_{1t}\}$ *Granger-causes* another process y_2 if the optimal prediction of future values of y_2 based on past values of y_2 alone can be improved by including current and lagged values of y_1 as explanatory variables. It is asserted that if y_1 does not Granger-cause y_2, then y_{2t} can be treated as strictly exogenous with respect to y_{1t}, so that correlations between the two can be interpreted as reflecting the causal effect of y_2 on y_1. The problem here is to determine the relation between Granger-causation and IN-causation as defined in this paper.

Analysts recognized immediately that Granger-causation is not the same as causation as that term is used in ordinary discussion. For example, Granger pointed out that under the definition just stated cattle stamping their hooves before an earthquake implies that the cattle Granger-cause the earthquake. Granger termed such cases "spurious causation", implying that the question of how to define causation that is not spurious remained open.

To determine the relation between Granger causation and IN-causation, we formulate a two-variable vector autoregression generating the values of the money stock $m = \{m_t\}$ and gross domestic product $y = \{y_t\}$ (note that henceforth in this section we use y to denote GDP, not to represent a general internal variable as above):

$$m_t = a_{my}y_t + b_{mm}m_{t-1} + b_{my}y_{t-1} + x_{1t} \tag{38}$$

$$y_t = a_{ym}m_t + b_{ym}m_{t-1} + b_{yy}y_{t-1} + x_{2t}. \tag{39}$$

Here the external variables x_{1t} and x_{2t} are independent of each other, and are independent over time. The reduced form corresponding to this system is

$$m_t = c_{mm}m_{t-1} + c_{my}y_{t-1} + u_{1t} \tag{40}$$

$$y_t = c_{ym}m_{t-1} + c_{yy}y_{t-1} + u_{2t}. \tag{41}$$

GDP fails to Granger-cause the money stock if

[6] This section draws heavily on Cooley and LeRoy (1985), although some of the discussion there is altered to accommodate the treatment here of causality.

$$c_{my} = \frac{a_{my}b_{yy} + b_{my}}{1 - a_{my}a_{ym}} = 0. \tag{42}$$

The money stock is *strictly exogenous* with respect to GDP if $a_{my} = b_{my} = 0$. Strict exogeneity implies that GDP shocks do not feed back into the equation determining money, either currently or with a lag. From (42) Granger non-causation is a necessary condition for strict exogeneity, but not a sufficient condition.

We wish to know what parameter restrictions are necessary for $m_t \Rightarrow y_t$. To determine this we first write the solution form of the model under the assumption that m_t is strictly exogenous:

$$m_t = x_{1t} + b_{mm}x_{1,t-1} + ... \tag{43}$$

$$y_t = a_{ym}x_{1t} + (a_{ym}b_{mm} + b_{ym})x_{1,t-1} + x_{2t} + b_{yy}x_{2,t-1} + ... \tag{44}$$

IN-causation requires that the ratio of the coefficients of x_{1t} in determining m_t and y_t equal the corresponding ratio for $x_{1,t-1}$:

$$\frac{1}{a_{ym}} = \frac{b_{mm}}{a_{ym}b_{mm} + b_{ym}}. \tag{45}$$

Here the reasoning is exactly the same as in Subsection 1.1. This equality is satisfied if and only if $b_{ym} = 0$.

Thus even strict exogeneity of m is not a sufficient condition for interpreting the coefficient of m_t in equation (39) for y_t as the causal coefficient associated with $m_t \Rightarrow y_t$. This is so because if $b_{ym} \neq 0$ the lagged values of x_1—the external variables that determine y_t through their effect on m_t—also affect y_t via m_{t-1}. Thus we have a failure of implementation neutrality: if $b_{ym} \neq 0$ characterizing an intervention as a hypothesized change in m_t does not give enough information about the intervention to determine the resulting change in y_t. Avoiding this outcome requires imposing the implementation-neutrality condition $b_{ym} = 0$ in addition to the strict exogeneity of m, so as to shut down m_{t-1} as a determinant of y_t.

We see that to make the transition from Granger-noncausation to IN-causation, one has to make two further restrictions on the model (38)-(39), beyond $c_{my} = 0$. The first is that $c_{my} = 0$ must be strengthened to $a_{my} = b_{my} = 0$. Analysts aware of the distinction between strict exogeneity and Granger non-causality frequently state that $c_{my} = 0$ is consistent with $a_{my} =$

$b_{my} = 0$, but then incorrectly go on to treat "is consistent with" as having the same meaning as "implies". Second, as we have just seen implementation neutrality requires that one rule out m_{t-1} as an argument in the equation for y_t.

The conclusion is that Granger causation is a specialized—and, to be sure, a very useful—form of forecastability, but it cannot be directly interpreted as having anything to do with IN-causation.

It may be that we are being too narrow in trying to relate Granger-causation to causation between current values of m and y as defined here. The definition of causation here relates a single cause variable and a single effect variable at the same date, whereas Granger causation involves the stochastic processes m and y. The suggestion is that a more general notion of causation is required. If so, the task at hand for proponents of Granger causation would seem to be to propose a more general characterization of (true) causation and then relate Granger causation to that.

8 Conclusion

In this paper we distinguish between two conceptions of causation, one a restricted version of the other. As is conventional, we use the term "causation" if any intervention that produces a change in the cause variable also produces a change in the effect variable. We direct attention to a stronger meaning for causation: IN-causation. One variable IN-causes another if, in addition to causing the other in the above sense, it is the case that all interventions that produce a given change in the cause variable induce the same change in the effect variable. If both conditions are satisfied the answer to the question "What is the effect of a change in y_1 on y_2?" does not depend on what caused the assumed change in y_1. This, as argued above, captures what scientists want to know when they investigate questions dealing with causation. If the conditions for IN-causation are not satisfied one cannot identify a single number that measures the effect of y_1 on y_2. In that case one can only discuss the effects of changes in the determinants of y_1 on y_2, which is unambiguous.

The question of how to implement the definition of causation proposed here is a difficult one. At a minimum, the analysis here can play the role of raising questions about discussions of causation that use purported measures of causal magnitudes which make no attempt to justify the implicit assumption of implementation neutrality. On a more ambitious level, the

results here may provide guidance on how to justify identifying particular model parameters with causation in applied models. The underlying idea is to encourage clear communication about what exactly is involved in causal assertions. A great deal remains to be done.

Acknowledgments

Abbreviated versions of this material were presented in LeRoy (1995) and (2006) (these papers are discussed at some length in Cartwright (2007)). Here more detail is supplied, and the packaging is different. I have received helpful comments from Judea Pearl and Hrishikesh Singhania, and from two referees.

References

Cartwright, N. 2007. *Hunting Causes and Using Them.* Cambridge: Cambridge University Press.

Cooley, T. F. and S. F. LeRoy. 1985. Atheoretical macroeconometrics: A critique. *Journal of Monetary Economics* 16: 283-308.

Elwert, F. Graphical causal models. 2013. In S. Morgan, Editor, *Handbook of Causal Analysis for Social Research.* Sage Publications.

Engle, R. F., D. F. Hendry and J.-F. Richard. 1983. Exogeneity. *Econometrica* 51: 277-304.

Granger, C. W. J. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37:424-438.

Granger, C. W. J. 1995. Commentary. In K. D. Hoover, Editor, *Macroeconometrics: Developments, Tensions and Prospects.* Kluwer Academic Publishers.

Haavelmo, T. 1943. The statistical implications of a system of simultaneous equations. *Econometrica* 11:1-12.

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

Hoover, K. D. 2001. *Causality in Macroeconomics*. Cambridge: Cambridge University Press.

Leamer, E. E. 1985. Vector autoregressions for causal inference? Volume 22, *Carnegie-Rochester Conference Series on Public Policy*.

LeRoy, S. F. 1995. Causal orderings. In K. D. Hoover, Editor, *Macroeconometrics: Developments, Tensions and Prospects*. Kluwer Academic Publishers.

LeRoy, S. F. 2006. Causality in economics. Reproduced, University of California, Santa Barbara.

Pearl, J. 2001. *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.

Simon, H. A. 1953. Causal ordering and identifiability. In W. C. Hood and T. C. Koopmans, Editors, *Studies in Econometric Method*. John Wiley and Sons.

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

Strotz, R. H. and H. O. A. Wold. 1960. Recursive versus nonrecursive systems: An attempt at synthesis. *Econometrica* 28: 417-427.

Woodward, J. 2003. *Making Things Happen*. Oxford: Oxford University Press.

Woodward, J. 2007. Causation with a human face. In H. Price and R. Corry, Editors, *Causation, Physics and the Constitution of Reality: Russell's Republic Revisited*. Oxford: Oxford University Press.

Biographical Information

Stephen F. LeRoy is emeritus professor of economics at the University of California, Santa Barbara. He has written a number of papers in economics and finance journals, and is author (with Jan Werner) of *Principles of Financial Economics*. He has held research positions in the Federal Reserve System and served on the faculties of several other universities.