Counterfactuals and Causal Structure*

Kevin D. Hoover

Department of Economics
and
Department of Philosophy
Duke University
Box 90097
Durham, NC 27708-0097

Tel. (919) 660-1876
E-mail kd.hoover@duke.edu

18 September 2010

*I thank Richard Scheines, Julian Reiss, Clark Glymour, and two anonymous referees for comments on an earlier draft.
Abstract
of

Counterfactuals and Causal Structure

by

Kevin D. Hoover

The structural account of causation derives *inter alia* from Herbert Simon’s work on causal order and was developed in Hoover’s *Causality in Macroeconomics* and earlier articles. The structural account easily connects to, enriches, and illuminates graphical or Bayes net approaches to causal representation and is able to handle modular, nonmodular, linear, and nonlinear causal systems. The representation is used to illuminate the mutual relationship between causal structure and counterfactuals, particularly addressing the role of counterfactuals in Woodward’s manipulationist account of causation and Cartwright’s attack on “impostor counterfactuals.”

**Keywords:** causation, causal order, causal structure, counterfactuals, James Woodward, Nancy Cartwright, invariance, structural account, manipulability account, causal pluralism, independence, modularity.
Counterfactuals and Causal Structure

Causality is closely related to the analysis of counterfactuals. Hume, who is often seen as having depreciated the status of causal relations, stressed their importance in political and economic contexts:

it is of consequence to know the principle whence any phenomenon arises, and to distinguish between a cause and a concomitant effect. Besides that the speculation is curious, it may frequently be of use in the conduct of public affairs. At least, it must be owned, that nothing can be of more use than to improve, by practice, the method of reasoning on these subjects, which of all others are the most important; though they are commonly treated in the loosest and most careless manners. [Hume 1754, p. 304]

The value of causal reasoning is, in part, diagnostic and retrospective: why did X happen? Such a backward looking question calls for a counterfactual inquiry: if Y had not happened, would X have happened? Causal reasoning is also prospective and related to planning: if Y were implemented, would X happen?

A central question addresses the relationship between causes and counterfactuals. David Lewis (1973), for example, defines causes reductively in terms of counterfactuals that are given an independent account. James Woodward (2003) also defines causes counterfactually, albeit nonreductively. Woodward’s account has become increasingly popular among philosophers of science, although it is not universally accepted. Nancy Cartwright (2007) attacks it, partly over the relationship of causes to counterfactuals. She objects both to defining cause in terms of counterfactual manipulations and to the subsequent use of causal knowledge so defined to evaluate counterfactuals for policy. Cartwright doubts that Woodward’s criterion for cause is generally applicable, and she

---

1 Nor is this view of the importance and utility of causal knowledge limited to Hume’s economic and political writings – see Hume (1739, pp. 73, 89; 1777, p. 76).
regards the counterfactuals supported by the supposed causal knowledge as irrelevant “imposters” (Cartwright 2007, esp. ch. 16).

Woodward’s account draws substantially on the graph-theoretic analyses of Peter Spirtes, Clark Glymour, and Richard Scheines (2001) and Judea Pearl (2000), in which causes are conceived as holding among variables that are connected through functional, but asymmetrical, relations. The graphs in these accounts are maps of the asymmetric flow of causal influence. A main purpose of the current paper is to suggest that Woodward’s version of the graph-theoretic approach implies an unnecessarily impoverished representation of causal relations and that these representations, in turn, lead him attribute too great a role for counterfactual manipulability in defining cause and to support a too highly constrained account of the structure of the relationships among causes and effects, laying his account open to many of Cartwright’s criticisms. I offer an alternative account, the structural account, built on work that long predates Woodward’s book, but which is less well known to philosophers (Hoover 1990, 1994, 2001). The structural account bears a close family resemblance to Woodward’s manipulation account. Yet, there are key differences that provide a richer set of resources, which are adequate to deal with Cartwright’s objections and to provide a basis for understanding the connection of counterfactuals to causality.

To avoid confusion, it is worth noting that the account proposed here does not fundamentally conflict with general approach of modeling causal relationships graphically, developed especially by Judea Pearl (2000) and Peter Spirtes, Clark Glymour, and Richard Scheines (2001), and used by Woodward. Rather it clarifies the relationship between graphical representations and systems of equations in a manner that
both enriches the graphical approach and demonstrates the fundamental kinship of the two approaches.

1. Woodward’s Manipulation Account
While Woodward’s account of causation relies on a counterfactual analysis, it is substantially different from the influential counterfactual account due to David Lewis (1973, 1979), which relies on a possible-worlds analysis. Woodward and I agree that Lewis unnecessarily privileges the notion of noncausal, universal laws (Hoover 2001, ch. 4, sections 4.2-4.4; Woodward 2003, p. 16, ch. 6). Furthermore, the notion of a metric for nearness of possible worlds is fundamentally vague (Woodward 2003, p. 138). Two analysts are vastly more likely to agree on a causal claim than on the truth or falsity of the counterfactual that is supposed to underwrite it or the nearness of the possible worlds that are supposed to decide that the truth value of that counterfactual.

In contrast to Lewis, who sees causal relations as connecting token events, Woodward follows Spirtes et al. (2001) and Pearl (2000) as seeing causal relations as connecting variables. Fundamentally, then, Woodward’s account is one of type-causation. Token-causation is analyzed through assessing the cases in which variables take particular values.

Causal relations among variables are represented both graphically and functionally. Thus in Figure 1, $A$ and $B$ cause $C$; and $C$ and $B$ (directly as well as indirectly), in turn, cause $D$. These relationships may be made more quantitatively precise by specifying the functional connections among the variables. For example, Figure 1 might be the graph of a system of equations:

\begin{equation}
C = \alpha_{CA}A + \alpha_{CB}B,
\end{equation}
where the $\alpha_{ij}$, $i, j = A, B, C, D$ are the coefficients that measure the strength of the causal connection between variable $j$ and variable $i$.

Neither the graphs nor the equations are dispensable, unless we abandon the symmetry of the equal sign and rule out functionally equivalent sets of equations as causally adequate. This can be done implicitly by adopting a rule: \textit{effects on the left; causes on the right}. While Woodward does not adopt such a convention, both Cartwright (2007, p. 13) and Hoover (2001, p. 40) explicitly convert the symmetrical equal sign into an assignment operator: “$\leftarrow$” for Cartwright; “$\triangleleft$” for Hoover. Thus, rewriting (1) and (2) as

\[(1') \quad C \leftarrow \alpha_{CA}A + \alpha_{CB}B,\]
\[(2') \quad D \leftarrow \alpha_{DC}C + \alpha_{DB}B,\]

in effect combines Figure 1 with equations (1) and (2).

Causal arrows indicate direct causes, a key concept in Woodward’s account:

\textbf{(DC)} A necessary and sufficient condition for $X$ to be a direct cause of $Y$ with respect to some variable set $V$ is that there be a possible intervention on $X$ that will change $Y$ (or the probability distribution of $Y$) when all other variables in $V$ besides $X$ and $Y$ are held fixed at some value by interventions. [Woodward 2003, p. 55]

Woodward (2003, p. 98) \textit{intervention variable (I)} for a variable $X$ with respect to a variable $Y$ is defined according to four criteria:

1. $I$ causes $X$;
2. $I$ acts as a switch so that when it takes the right values it can eliminate the effect of all other variables in determining $X$;
3. any causal path from $I$ to $Y$ goes through $X$;
4. $I$ is independent of any variable $Z$ that causes $Y$ otherwise than through $X$. 

\[ (2) \quad D = \alpha_{DC}C + \alpha_{DB}B, \]
An *intervention* is defined as a token realization of an intervention variable that is an actual cause of the value of $X$.

Counterfactuals play a role at two key points in Woodward’s definition of direct cause. First, the definition relies on counterfactuals in that it is enough that the contemplated interventions are possible; he does not require them to be actual. The exact modality captured in a *possible* intervention (equivalently, possible manipulation) is an open question. On the one hand, Woodward rejects as potential causes variables for which we have no notion of manipulation, as well as variables, such as race, sex, or species, for which a change would threaten the fundamental identity of the subject to which the variable is attached (Woodward 2003, pp. 113; cf. Hoover 2009a). On the other hand, Woodward rejects the notions that manipulations must be the result of human agency (naturally occurring “interventions” will suffice) or that they are necessarily practically possible (the moon causes the tides, but how can we practically manipulate the moon in the right sort of way?) (Woodward 2003, pp. 113).

The second point at which counterfactuals play an essential role is in the notion that the causal relationship is to be evaluated in isolation by holding other variables fixed. There may, in fact, be no way actually to achieve such holding fixed and, like Lewis, Woodward is willing to countenance the semantic device of “small miracles” to achieve the necessary isolation (Lewis 1973, p. 560; Woodward 2003, pp. 132, 136). And like Lewis, Woodward evaluates the counterfactual manipulation not in the actual world or, more accurately perhaps, in the actual causal graph, but in one that is different, though derived from it.
The semantic content of the assertion of Figure 1 that \( C \) is a direct cause of \( D \) is captured in the counterfactual experiment of manipulating some of its variables. Following Pearl, Woodward suggests that we consider an intervention that sets variables other than \( C \) and \( D \) to token values – in effect, “breaking” (or “wiping out”) the causal connections between variables wherever needed to achieve this. Thus Figure 1 would be replaced by Figure 2 in which the lower-case letters indicate token values for the correlative upper-case variables and in which the causal arrows into \( C \) are removed. \( C \) causes \( D \), then, if a change in \( C \), say, from \( c \) to \( c' \) results in a change in \( D \), say, from \( d \) to \( d' \). The truth of this counterfactual justifies the direction of the causal arrow from \( C \) to \( D \) in Figure 1.

Although establishing direct cause relies on the evaluation of a counterfactual, Woodward’s account, unlike Lewis’s, is not reductive. Manipulation is an admittedly causal relationship. Rather than explaining causation in terms of some more basic notion, Woodward explains the causal structure of one part of a network of variables in terms of the causal structure of other parts. While such a nonreductive account may be metaphysically unsatisfying to those unwilling to take causation as a primitive, it is very much in keeping with Cartwright’s (1989, ch. 2) slogan, “no causes in, no causes out,” and provides a framework for a causal epistemology, which explains its appeal to philosophers of science.

How one is to evaluate the truth value of counterfactuals remains an issue. Woodward rejects Lewis’s appeal to universal natural laws. In the end, he grounds the evaluation of counterfactuals in the empirical fact of invariance. The invariant connection of the manipulated cause to the effect, under the conditions set out in the
definition of direct cause (DC) is relied upon to translate the causal map given in the graphs and their associated functions into more complex counterfactual assessments that constitute Hume’s useful causal knowledge. Invariance, in Woodward’s view, is not absolute but admits of degrees. A relationship may be invariant to some sorts of interventions and not to others (Woodward 2003, ch. 6, section 6.4). And, in general, Woodward stresses that causal knowledge and the assessments of counterfactuals are deeply contextual and that causal explanation is contrastive.

Woodward’s strategy of defining direct cause through a process of counterfactual manipulation and then reconstructing causal networks out of the pieces requires, he believes, that causal relationships possess a kind of autonomy that he calls modularity:

a system of equations will be modular if it is possible to disrupt or replace (the relationships represented by) any one of the equations in the system by means of an intervention on (the magnitude corresponding to) the dependent variable in that equation, without disrupting any of the other equations. [Woodward 2003, p. 48]

The actual systems that, for example, a scientist works with may or not be modular; nonetheless, Woodward maintains

that when causal relationships are correctly and fully represented by systems of equations, each equation will correspond to a distinct causal mechanism and that the equation system will be modular. [Woodward 2003, p. 49]

Modularity, and whether it is essential to causal relationships, is a major point of dispute between Woodward and Cartwright (see Cartwright 2007, chs. chs. 7, 8; Hausman and Woodward 1999, 2004; cf. Hoover 2009a).

2. The Structural Account

While it is an alternative to Woodward’s manipulation account of causation, the structural account bears a family resemblance to it and to the related graph-theoretic
analyses of Spirtes et al. and Pearl. Its pedigree, however, can be traced back principally to J.L. Mackie’s (1980, ch. 3) INUS analysis of causation and to Herbert A. Simon’s (1953) analysis causal order in econometrics (see Hoover 2001, ch. 2). Our present focus is on Simon.

2.1. Simon on Causal Order
Simon (1953) proposes a syntax for representing causal relationships that suits the structural account very well. Consider the representation of a causal structure in a system of equations, such as (1) and (2) with the addition of

\[ A = \alpha_A, \]

and

\[ B = \alpha_B. \]

These equations, in which \( A \) and \( B \) are set equal to parameters, complete the system of equations, so that once the parameters have been assigned values, the system can be solved. Call the system (1)-(4) \( S \). In \( S \), we can solve for the value of \( A \) from equation (3) alone without knowledge of the parameters of the other equations, and we can solve for the value of \( B \) from equation (4) alone. Each is a minimal complete subsystem of \( S \); call them \( S_A \) and \( S_B \). Similarly equations (1)-(3) form a complete subsystem \( (S_C) \) in which we can solve for \( A \), \( B \), and \( C \). It is minimal for \( C \), though not for \( A \) and \( B \). Equations (1)-(4) form a complete subsystem \( (S_D = S) \) that is minimal for \( D \). For Simon, causal order is about the hierarchical relationships of minimal complete subsystems. \( A \) and \( B \) cause \( C \) because \( S_A \) and \( S_B \) are subsystems of \( S_C \) (written \( S_A \subset S_C \) and \( S_B \subset S_C \)). \( C \) causes \( D \) because \( S_C \subset S_D \). \( A \) is an indirect cause of \( D \) because \( A \) causes \( C \) and \( C \) causes \( D \) and
knowing the value of $C$ allows us to dispense with knowledge of the parameters that determine $A$ in solving for $D$. $B$ is both a direct and indirect cause of $D$ because, as with $A$, there is a chain of causation running through $C$; but, in contrast to $A$, knowledge of $C$ is not enough to allow us to dispense with knowledge of the parameters of $B$ in solving for $D$.

Simon recognizes that his syntactic approach is inadequate on its own because structures of equations can be written in equivalent forms that syntactically yield different systems of minimally complete subsystems and, therefore, different causal orderings. For example, let system $S'$ consist of equations (2), (4) and

\begin{equation}
A = \beta_A + \beta_{AB}B + \beta_{AC}C,
\end{equation}

where $\beta_A = \frac{\alpha_A}{1-\alpha_{CA}}$, $\beta_{AB} = \frac{\alpha_{CB}}{1-\alpha_{CA}}$, and $\beta_{AC} = -\frac{1}{1-\alpha_{CA}}$; and

\begin{equation}
C = \beta_C + \beta_{CB}B,
\end{equation}

where $\beta_C = \alpha_A\alpha_{CA}$ and $\beta_{CB} = \alpha_{CB}$.

By construction, systems $S$ and $S'$ have identical solutions; yet by Simon’s syntactic criteria the causal ordering of $S'$ is represented by Figure 3 – substantially different from the causal ordering of $S$ in Figure 1.

Simon’s solution to this problem of observational equivalence is to provide a semantic account of causal order (Simon 1953, pp. 24-26; 1955, p. 194). Parameters are not, on this view, fixed constants, but precisely the things that are altered by interventions or manipulations. If the $\alpha$-parameterization of $S$ were the true one, then any one of its parameters can be set to a new value without affecting any of the other parameters in the
system. However, the $\beta$-parameters of $S'$ must change in order to maintain the common solution. It works both ways, if the $\beta$-parameterization of $S'$ were the true one, then the $\alpha$-parameters of $S$ would have to change in the face of a change in one of the $\beta$-parameters.

The true causal order, then, is one that allows mutually unconstrained interventions among its parameters or, to put it another way, any change to the variables of the causal system leaves the parameterization and, therefore, the functional form of the remaining causal relations invariant. Invariance of the functional forms in the face of specific interventions is, on this view, the hallmark of a true causal representation; while failure of invariance is a key symptom of causal misrepresentation.

We can restate Simon’s semantics by defining a parameter to be one of a set of variation-free variables that represent the scope for interventions in the causal system, where variation-free means that the choice of any particular value for a variable does not constrain the admissible choices of values for other variables in the set.

Cartwright objects to this characterization of a parameterization as a set of variation-free variables that govern the values of variables that are constrained by the causal structure: “this is not generally the distinction intended between the parameters and the variables” (Cartwright 2007, p. 241). But I submit that defining parameter in this way is consistent with ordinary usage. The *Oxford American Dictionary* defines parameter as “a variable quantity or quality that restricts or gives particular form to the thing it characterizes.” Simon treats parameters as capable of taking different values and uses the parameterization to define the causal order – the form of the causal order of a system of variables.
Cartwright also suggests that this interpretation of Simon’s characterization of causal order is incorrect (Cartwright 2007, ch. 13, 14). The best rejoinder is to quote Simon at length:

The causal relationships have operational meaning, then, to the extent that particular alterations or “interventions” in the structure can be associated with specific complete subsets of equations. We can picture the situation, perhaps somewhat metaphorically, as follows. We suppose a group of persons whom we shall call “experimenters.” If we like, we may consider “nature” to be a member of the group. The experimenters, severally or separately, are able to choose the nonzero elements of the coefficient matrix of a linear structure, but they may not replace zero elements by nonzero elements or vice versa (i.e., they are restricted to a specified linear model). We may say that they control directly the values of the nonzero coefficients. Once the matrix is specified, the values of the \( n \) variables in the \( n \) linear equations of the structure are uniquely determined. Hence, the experimenters control indirectly the values of these variables. The causal ordering specifies which variables will be affected by intervention at a particular point (a complete subset) of the structure. [Simon 1953, p. 26]

What Simon refers to as “coefficients” subject to direct control – that is, able to be freely chosen by the “experimenters” – is exactly what we call parameters.² And a specific change in a parameter is the manner in which an intervention (or Woodward’s manipulation) is implemented. We can think of a causal system as a machine whose various operating characteristics are the variables which are controlled indirectly by selecting the settings for various switches and dials.

Significantly, Woodward (2003, p. 96) uses the analogy of switches as a means of explaining the breaking or wiping out of causal arrows involved in intervention. The operation of a switch is not analogous to the wiping out of a causal relationship. (This is perhaps more obvious with respect to dials that allow the setting of a continuously variable quantity. Not for the first time causal analysis is misled by philosophers’

² In light of the fact that Cartwright sees “direct control” as highly restrictive notion and the possibility that her view arises partly from an assumption that direct control requires human agency, it is worth reiterating that Simon counts “nature” as among the “experimenters” who can exercise direct control (Cartwright 2007, p. 252, fn. 27; also p. 205).
penschant for 0/1 or on/off variation.)\(^3\) Flipping a switch does not break a causal system; it operates it. Interventions that change the values of parameters maintain the topology of the causal relationships among variables, calling for variables to take different values but not altering the causal graph itself.

While Simon’s conception of causal order in one sense rejects Woodward’s approach (causal order is not best understood through the comparison of a causal system to a topologically different system), in another sense it generalizes it. For the parameters represent the scope of possible interventions in a causal system, so that the connection between interventions and outcomes for variables is clear. Simon’s conception shifts the focus away from specific token interventions to parameters that can take a variety of values. These are types, which can instantiate a variety of token manipulations. The causal structure is defined entirely at the type level.

The structural account takes the minimal causal connection between two variables as primitive, offering no deeper account. The nature even of such a primitive causal connection must be understood counterfactually. If a cause has a certain effect in the right circumstances, then we cannot sensibly assert that it has that effect when those circumstances are not actual but not when they are actual. If diamonds scratch glass, the property cannot hold only when a diamond is not actually used to scratch glass. This is the sense in which causal relationships are naturally connected to invariance and it

\(^3\) And, indeed, Simon is not free of the potential confusion; for he refers to the elimination of a causal linkage as the setting of a coefficient to zero; but there is a difference between a parameter having no value and having a range of admissible values which happens to include zero. Causal analysts frequently – and no doubt inadvertently – equivocate on the meaning of zero, failing to distinguish these two cases. See Hoover (2001, p. 45, esp. fn. 13). In the cited passage, Woodward contrasts switches and dials: switches break causal connections, while dials modulate the strength of causal connections. The structural account sees this as a distinction without a difference.
captures the meaning of what it is for a cause to be necessary in the circumstances for an effect.

Primitive causal connections reflect the natures or capacities (to use Cartwright’s 1989 preferred term) of the causes. And capacities must be understood as dispositional, subject to a counterfactual analysis (see Mackie 1973, ch. 4). Cartwright analyzes capacities as dispositions that are carried from context to context; yet they do not have to express themselves in every context (Cartwright 1989, pp. 3, 146-147, 191, passim). It is no failure of an account in terms of capacities that contextual details matter substantially in whether capacities are actualized. And a capacity account in no way presupposes modularity, which is a sort of independence from context.

While direct causal connections are primitive, they are also relative to the representation or model and may or may not be brute facts. For example, we might imagine that a drug \(D1\) is found experimentally to reduce coronary thrombosis \(CT\). The relationship may be modeled as in Figure 4.A. (Plus or minus signs next to causal arrows indicate whether the causal influence promotes or inhibits the effect.)

It is not inconsistent with such a model that further research supports a more complex model (Figure 4.B) in which \(D1\) causes \(CT\) directly and, in fact, promotes coronary thrombosis, while it also reduces blood pressure \(BP\), while high blood pressure promotes thrombosis; the net affect being to reduce thrombosis. Sometimes a model such Figure 4.B is taken to imply that the model in Figure 4.A is defective. A better interpretation is that the models operate in different contexts. If there were no known means of intervening independently on blood pressure or in such a way as would meliorate the adverse direct effect of \(D1\), then the model in Figure 4.A would be a
perfectly fine model and a guide to clinical practice. The model in Figure 4.B would simply explain the mechanism through which $D1$ inhibits thrombosis.

An advantage of the model in Figure 4.B, however, is that, in articulating the mechanism of operation, it may suggest paths toward better outcomes. For example, knowing that the positive effect of $D1$ operates through $BP$ suggests seeking another drug (say, $D2$) that would reduce blood pressure with no direct effect on coronary thrombosis (Figure 4.C). If research successfully produced such a drug, a better clinical practice might be to administer $D2$ and omit $D1$. There is a sense in which the causal arrow in Figure 4.A captures a fact that is primitive relative to the model, but which is not brute, in that it has a more complex explanation in a finer grained model. There is no guarantee that primitive causes can be explained through further refinements, though that is often the object of research.

2.2. MODULARITY AND DIFFERENCE-MAKING
The structural account agrees with Woodward (2003, e.g, p. 80) that causes are difference makers, yet it marks the difference that they make relative to an intact causal structure, not some related, but different, structure constructed by manipulations that, in effect, break the system. The difference between our approach and Woodward’s becomes important with respect to modularity. Where Woodward sees modularity as a fundamental element of a well-articulated causal system, the structural approach does not require modularity at all – a distinct advantage, since many intuitively causal systems are decidedly nonmodular.

Different notions of intervention also distinguish the structural account from Woodward’s manipulation account. A parameter can be thought of as a causal variable,
so there is no fundamental difference with Woodward’s criterion 1 for an intervention variable, cited in section 1: \( I \) causes \( X \). A significant difference arises with Woodward’s criterion 2: \( I \) acts as a switch so that when it takes the right values it can eliminate the effect of all other variables in determining \( X \). As already noted, while parameters may well act as switches or dials (i.e., instruments of continuous variation rather than simply on or off), they need not shut off the effects of other causes. The critical feature is not the breaking or wiping out but the accounting for the effects of other causes. Nor does the structural account accept criterion 3: any causal path from \( I \) to \( Y \) goes through \( X \). This criterion is closely related to modularity, and the structural account does not require modularity. A parameter may have a direct effect on \( Y \) as well as an indirect effect on \( Y \) though \( X \) without undermining a clear causal ordering of \( X \) and \( Y \). The case in the next subsection illustrates exactly this situation. The structural account suggests a different interpretation of criterion 3. When it is fulfilled for parameters (\( I \)), then we find ourselves in a particularly fortunate position to infer causal direction through an intervention. We should not, however, confuse the epistemic issue of how and when causes are inferable from data with the conceptual issue of what it means to be a cause or with the question of how to represent causal order. The structural account does not accept Woodward’s criterion 4: \( I \) is independent of any variable \( Z \) that causes \( Y \) otherwise than through \( X \). If \( I \) is interpreted as a parameter as we have defined it, then it is required to be independent only of other parameters and not of all other variables. This is the requirement that parameters be variation-free.

An intervention for the structural account is a token realization of a parameter in the same way that an intervention for Woodward is a token realization of an intervention.
variable. But unlike Woodward, the type-relations among the variables and the parameters fully determine the causal order without reference to a particular token intervention or manipulation. In relying on token interventions in the definition of direct cause, Woodward again seems to confuse the causal relationship with a strategy that supports the inference of causal relationship.

We can illustrate the issue with a typical macroeconomic model. The demand for real money balances \((m - p)\) is given by

\[ m_t - p_t = \delta - \alpha (p_{t+1} - p_t) + \nu_t, \]

where the subscripts \(t\) indicate time periods; \(m\) is the logarithm of money; \(p\), the logarithm of the price level; \(p_{t+1}\), the expectation at time \(t\) of the price level at time \(t+1\); \(\nu\), an independent random error; and \(\alpha\) and \(\delta\) are parameters. The central bank’s money supply rule is

\[ m_{t+1} = \lambda + m_t + \epsilon_t, \]

where \(\lambda\) is a parameter that governs the growth rate of money and \(\epsilon\) is an independent random error. Expectations are formed rationally:

\[ \epsilon_{t+1} = E(p_{t+1} \mid \Xi_t), \]

which says that the expectation of the price level is the mathematical expectation of actual prices conditional on all the information available at time \(t\), including the model itself. Janssen (1993, pp. 137-139) argues persuasively that “rational expectations” are not expectations at all, but a consistency criterion or solution concept analogous to the condition that markets clear – that is, that prices are assumed to take the values at which

---

4 This model is adapted from Hamilton (1995, pp. 326-332).
supply equals demand (see also Hoover 2009b). On that assumption, \( p_{t+1}^{c} \), is not a proper variable – or at least not a causally efficacious one – but an instrument for imposing a certain nonlinear restriction on the parameters of the model.

On that interpretation, the causally relevant solution to the model is given by (9) and

\[
(11) \quad p_t = m_t - \delta + \alpha \lambda + \nu_t. \quad ^5
\]

The model is nonlinear in parameters, and it is not modular. The appearance of the multiplicative coefficient \( \alpha \lambda \) in (11) results from the imposition of rational expectations – \( \lambda \) appears only because it is a parameter of the money-supply rule (9). Thus, it is impossible to perform Woodward’s manipulation test of whether money causes prices, which calls for setting \( m_t \) in (9) to a fixed value come what may and, in effect, wiping out any causal arrows, say, from \( m_t \) or \( \nu_t \) to \( m_{t+1} \); for that would remove the basis for the parameterization of (11). The causal structure reflected in (11) cannot survive such a breaking of the system.

Nevertheless, applying the definition of direct cause from the structural account tells us unequivocally that \( m_t \) causes \( p_t \). Rather than calling for miraculous token manipulations, direct cause is implied by the constraints on the variables determined by the parameterization.

The model illustrates another important point. Consider a change in the central bank’s money-supply rule – for example, an increase in the growth rate of money indicated by a higher value for \( \lambda \). As equation (11) indicates such an intervention would alter the coefficient \( \alpha \lambda \) in (11). As a result a statistical test of the relationship of money

\[^5\] See Hoover (2001, pp. 64-65) for the derivation of the solution to a slightly more general version of this model.
to prices based on (11) would be noninvariant. Woodward and others have treated
invariance under manipulation of causes as the hallmark of a causal relationship, and they
would, therefore, be tempted to reject the causal status of (11) (Woodward 2003, pp. 15-
16, ch. 6). What the example actually shows is that a more subtle approach to invariance
is necessary.

Consider an intervention that changes prices through some other instrument than
money; for example, consider an intervention that changes \$d$, then naturally (11) is
noninvariant. But (9) is invariant. And this is a general rule in nonmodular systems with
one-way causes: the causal structure that determines a cause of an effect is invariant to
interventions that alter the effect through some other mechanism than the cause in
question; while the causal structure that connects a cause to an effect is not generally
invariant to interventions that alter the cause of the effect in question (Hoover 2001, ch.
8; Cartwright 2007, pp. 99, 105). In fact, the differential invariance is diagnostic of
causal direction in nonmodular systems; and invariance in both directions is a test of
modularity (see Hoover 2001, ch. 8, section 8.1). The more subtle analysis of invariance
nonetheless supports Woodward’s view that one can alter the effect by manipulating the
cause; one cannot alter the cause by manipulating the effect.

2.4. COUNTERFACTUAL ANALYSIS
Lewis explains causal relations by an appeal to counterfactuals and evaluates the truth of
counterfactuals through a comparison of possible worlds. Mackie (1973, ch. 3) rejects
the notion that counterfactuals per se have a truth value. He interprets them as
enthymemetic or disguised argument, which can be evaluated for validity once its
structure is articulated and, additionally, for soundness once the truth of its premises is established or, at least, accepted “for the sake of argument.”

The structural account of causation rejects Lewis’s account of the counterfactual basis for causal relationships, but is compatible with Mackie’s account. A causal model can be interpreted as a map of possible worlds. Unlike the possible worlds in Lewis’s account of counterfactuals, causal models are precise about what changes are possible and what implications they have for the variables in the model. Consequently, they provide an instrument for the construction and articulation of the kind of arguments that Mackie sees as underwriting counterfactuals, and they avoid the hopeless ambiguities of Lewis’s metric for the closeness of possible worlds.

Naturally, we cannot avoid the question of the adequacy of causal models as representations of the real world or the need to choose among competing causal models. But these are the ordinary epistemological problems faced in scientific inference. They may be practically difficult and subject to other philosophical reservations; but, once we are satisfied that they have been dealt with adequately, counterfactual analysis itself is not additionally problematic.6

3. Counterfactuals and Policy Analysis

3.1. IMPOSTOR COUNTERFACTUALS
In her recent book, Cartwright (2007, ch. 16)) makes a case against the typical uses of counterfactual analysis in economics. Her theme is reflected in her title, *Hunting Causes*

---

6 Spirtes et al. (2001) provides an extensive account of the using of information encoded in conditional independence relationships of variables as a means of establishing facts relevant to causal inference. Hoover (2001, chs. 8-10) provides both a methodological account and case studies of causal inference based on interventions of a type that induces structural change in causal systems.
and Using Them; she suggests that the techniques appropriate to hunting causes are not the ones appropriate to using them in counterfactual analysis. The title, however, masks different levels of issues. It seems off the mark. In the well known story, the recipe for juggled hare (usually wrongly attributed to the famous cookbooks of Mrs. Beeton or Mrs. Glasse) begins “first, catch your hare . . .” In an obvious sense, one surely needs to hunt causes (that is, to establish the existence of causal connections empirically) before one can use them for anything.

Cartwright’s real concern is with what she calls “imposter counterfactuals” – that is, cases in which the counterfactual that received empirical warrant is not the one that would appropriately warrant a counterfactual policy analysis. One of Cartwright’s objections is simply an extension of her skepticism of modularity. Treated as an empirical strategy, Woodward’s manipulation test should reveal the causal relationship if the causal structure is, in fact, modular. Cartwright does not deny that, but stigmatizes such systems as “epistemically convenient,” suggesting that such convenience is necessarily rare.

She also objects to the breaking or wiping out of causal connections as part of Woodward’s and Pearl’s approach to evaluating a causal relationship on the ground that it is the intact system, not the broken system, that it is needed to evaluate policy counterfactuals. The difficulty is that when implementing a policy, we may in fact not be able to manipulate a cause independently of other causes, so that the counterfactual that Woodward or Pearl seeks to evaluate is simply not a counterfactual that the policy analyst can rely on.
This objection connects to the distinction that Cartwright draws between implementation-neutral and implementation-specific counterfactuals. An *implementation-neutral* counterfactual is one that implies the same effect no matter what means are used to bring about the causal antecedent. An *implementation-specific* counterfactual is one in which the effect is sensitive to the manner in which the causal antecedent is brought about. For example, in the causal structure connecting high blood pressure to coronary thrombosis in Figure 4.C, the counterfactual question, “how much would a reduction in blood pressure reduce thrombosis?,” is not well posed because the counterfactual is not implementation-neutral. A reduction of blood pressure to a particular level using drug $D_2$ will be more effective than one using drug $D_1$, since $D_1$ has a direct promoting affect on thrombosis independent of its indirect inhibiting effect operating through its effect on blood pressure.

The structural account of causation suggests that implementation-specific counterfactuals are the rule, in large measure because causal complexity and failures of modularity are the rule. What Cartwright calls implementation-neutral policies are merely policies that benefit from the special features of some causal structures that render them *robust* to the mode of implementation. In such structures, a variety of modes of determining a cause produce the same effect. Such robustness is practically useful in many cases and may be sought for that reason, but there is no reason to connect it to the existence or nonexistence of a causal relationship in general. In fact, not infrequently a lack of such robustness is desirable. Pushing up on the plastic tab causes the cap to the medicine bottle to come off, but *only when* the plastic tab is aligned with the arrow on the bottle. The lack of robustness contributes to child safety.
Cartwright (2007, p. 254) assumes that we should prefer implementation-neutral policies. Yet, such a preference is not obvious. And, even if we did prefer them, we would need a more detailed, accurate causal representation to be sure that the policies were in fact implementation-neutral. For example, a simple model of the relationship of blood pressure to coronary thrombosis, $BP \rightarrow CT$, might appear to be implementation-neutral; but, if Figure 4.C, truly represents the causal structure, whether a fall in blood pressure is brought about by drug $D1$ or $D2$ matters, the policy is implementation-specific, and we have made a mistake.

3.2. An Illustration from Monetary Economics
We can illustrate some of the key issues and how the structural account deals with them using the model in equations (8)-(11). A monetary regime is defined by the parameterization of the central bank’s money supply rule (9), so that any change in the parameter $\lambda$ represents a new regime. Imagine that the model (8)-(11) is, in fact, a true representation of the economy; it is, what econometricians sometimes call the data-generating process. Equations (9) and (11), then describe the actual dynamic process of governing the evolution of money and prices. Of course, economists do not know the data-generating process a priori. A central problem for econometrics is identification: how can we recover the parameters of the data-generating process (or of some, close enough approximation) based on observations of the variables (here, of $m$ and $p$)?

Many macroeconomists estimate so-called structural vector autoregressions. Most of the details are not important here, but a few are worth noting. The structural

---

7 Equations (8)-(11) represent a simplified version of an actual econometric model used to conduct counterfactual analyses of U.S. monetary policy (see Hoover and Jordá 2001).
vector autoregression technique gives up on learning about all of the parameters of the dynamic process, focusing instead only on those that relate the contemporaneous values of variables to each other, letting the relationships of lagged to contemporaneous variables be summarized by coefficients that may themselves be difficult-to-disentangle functions of the parameters of the data-generating process. Estimates of these contemporaneous parameters are obtained under a maintained assumption about the causal order of the variables. For example, if we assume (correctly) that in the data-generating process to hand, $m_t$ causes $p_t$, we would be able to recover good estimates of the true parameters. But most economists make the necessary assumptions about causal ordering on the basis of \textit{a priori} guesswork, so that considerable doubt hangs about their estimates. This is an area in which the graph-theoretic (or Bayes net) inferential techniques pioneered by Spirtes et al. (2001) and first applied to the problem of structural vector autoregressions by Swanson and Granger (1997) have considerable power.\footnote{See Demirap and Hoover (2003), Hoover (2005), Demiralp, Hoover, and Perez (2008); Hoover, Demiralp and Perez (2009) further development, evaluation, and applications of these techniques to economic problems.}

Typically, once an economist has estimated a structural vector autoregression, it is used to evaluate particular counterfactual questions: for example, treating (9) and (11) as the structural vector autoregression, we might ask, “what would be the path of prices ($p$) if money ($m$) were increased for a single period?” Such a one-period increase is referred to as a “monetary shock” (or “impulse”) and the path of prices is referred to as an “impulse-response function.” The shock is typically administered by setting the error term (here $\varepsilon$) to a positive value for a single period.

The effect of the shock is nonetheless the same as Woodward’s or Pearl’s experiments forcing a variable to take a particular value come what may. Typically, the
implied breaking of causal relationships is restricted to the current period, so that future values of the money are not fixed but allowed to develop in line with the causal structure. However, one could in principle offer a series of shocks that had the effect of fixing money at every future time period. That this is not often done is partly pragmatic: the own dynamics of the shocked variable are independently interesting, so economists prefer not to suppress them.

It is also partly the result of the Lucas critique, the fact that in a model with rational expectations, the dynamics are not invariant with respect to changes in the policy rule. The Lucas critique is exemplified in the point previously made that the coefficient $\alpha \lambda$ shifts with changes in the monetary-policy rule (the setting of the monetary growth rate, $\lambda$), so that to know the path of prices ($p$), we need to know not only the value of $m_t$ but also how $m_t$ was brought about – that is, the value of $\lambda$. Impulse-response analysis is sometimes thought to circumvent the Lucas critique. The idea is that a shock to the random-error term in (9) leaves the parameters untouched and, therefore, does not induce any failure of invariance in (9) or (11).

Impulse-response analysis provides a good example of what Cartwright criticizes as impostor counterfactuals. It is used to say something about the effects of monetary policy – for example, what would happen if the Federal Reserve raised the money supply by 1 percent? – and it tries to answer that question by treating the Federal Reserve’s action as a shock to a stable system. The impulse-response function does answer a well-posed counterfactual question, but not the one for which we want an answer. The question it actually poses is, what if the money supply were to rise unexpectedly and

---

9 Lucas (1976); Hoover (1988, ch. 8, section 8.3; 2001, ch. 7, section 7.4).
arbitrarily, say, by 1 percent above its dynamic path and then fall back the next period by
the same amount? It truly asks what would happen if there were a shock to the system.
But monetary policy is not delivered by shocks. The increases in the money stock that
the Federal Reserve typically delivers are reactions to economic conditions aimed at
desired goals for economic variables. Furthermore, while the impulse-response function
can trace out the effects of a random shock, it cannot trace out the effect of a change in
systematic policy, since to deliver a series of shocks in one direction, for example, in
order to force money to evolve along a desired path represents a violation of the
randomness assumptions that govern the underlying representation of the error term (9).
It is not that such systematic policy could not be analyzed; it is that it cannot be analyzed
while assuming that the causal structure of the model, which includes the model of the
random errors, is constant and – at one and the same time – changes.

In a discussion of the Great Depression, the economist Christopher Sims (1999)
recognizes that random shocks to error terms could not adequately address the
counterfactual question, what if monetary policy in the 1930s had adopted the rules that
characterized it in the 1990s? He obtained econometric estimates of a model more
complex than, but similar to, (8)-(11) for both the 1930s and the 1990s. To evaluate the
counterfactual, he, in effect, replaced the monetary-policy rule (9) for the 1930s by the
one that characterized the 1990s. For our purposes, think of this as simply changing \( \lambda \)
to a new value – while holding the estimated value of \( \alpha \lambda \) and the other parameters of (11)
constant at values appropriate to the 1930s. The counterfactual is evaluated by setting \( m \)
and \( p \) in the first period to the values that they actually took at the onset of the Great
Depression and then feeding the random error terms from the original estimates into the
model with the alternative monetary-policy rule. (Notice that if this procedure were undertaken with the original monetary-policy rule, it would necessarily have simply generated the actual path for money and prices over the Great Depression.) With such counterfactual estimates, Sims felt free to discuss whether modern central bankers would have produced better outcomes.

The ordinary impulse-response analysis was a true impostor counterfactual, in that it answered a counterfactual question, but the wrong one. In contrast, Sims’s counterfactual experiment is simply incoherent. If, as he holds, the existence of rational expectations subjects the model to the Lucas critique, then one cannot simply substitute one monetary-policy rule for another. The incoherence is displayed in simultaneously assuming that \( \lambda \) may change while \( \alpha \) and \( \alpha \lambda \) remain constant. The difficulty is the nonmodularity of the causal structure. Were the causal structure modular, as it might be if rational expectations were not an element of the data-generating process, then Sims’s method would be coherent.

Hoover and Jordá (2001) offer a different counterfactual analysis. Rather than replacing the monetary rule of the 1930s by that of the 1990s, they simply transfer the entire model of the 1990s to the 1930s by setting the initial values of the 1990s model to their values at the onset of the Great Depression and then feeding the estimated random error terms from the 1930s model into the 1990s model. This procedure makes sense if the elements of the model other than the particular shocks and the monetary-policy rule have not changed between the two time periods. The counterfactual question that it answers is well-posed and not an impostor. Essentially, the procedure is that same as if we had changed \( \lambda \) in the estimated 1930s model and, unlike Sims, allowed \( \alpha \lambda \) to take a
new value, holding $\alpha$ constant. The central message of Hoover and Jordá’s approach is that one should not ignore nonmodularity but account for it. Accounting for it raises difficult, but not necessarily insuperable, inferential problems, which – fortunately – are not our direct concern here.

3.3. INTERNAL AND EXTERNAL VALIDITY

Holding $\alpha$ constant is the emblem in our expository model for the constancy of the rest of the structure of the model between the two periods. The assumption that we are justified in doing so is by no means automatic and raises the classic question of internal versus external validity. The issue is whether a causal relationship (indeed, empirical relationships of other kinds as well) uncovered in a specific context can be transferred and assumed to hold in other contexts. Superficially, it might appear to recapitulate the distinction between implementation-specific and implementation-neutral counterfactuals. The implementation dichotomy comes down, first, to whether or not an empirically warranted causal structure supports the counterfactual; second, to whether the detail in the representation of that structure is fine enough to distinguish between alternative policy implementations; and, finally, to whether the target effects are, in fact, robust to different implementations. In contrast, external validity comes down, first, to whether there is homogeneity in the background conditions between implementations in the two situations; and, second, to the domain of possible interventions.\(^{10}\)

\(^{10}\) Hoover (2009a) frames the notion of background conditions in a manner consistent with Woodward’s (2003, pp. 145-146) emphasis on contrastive focus using John Anderson’s (1938/1962) notion of a causal field. The causal field is the set of standing conditions that, while they may themselves be causes, do not change relative in relation to our particular causal interests and, so, define the boundary conditions for a particular causal relationship. Causal relations may be evaluated differently in different causal fields. As a result, causal relationships may be represented or modeled in a variety of (ultimately noncontradictory) ways depending on our differing pragmatic aims.
In the case of the monetary-policy counterfactual, there is no implementation-neutral counterfactual possible: to know what effect a change in the money stock causes we must know how it comes about – an increase in \( m_t \) from a shock to \( \varepsilon_t \) has a different effect than one from a rise in \( \lambda \). Yet, Hoover and Jordá’s counterfactual analysis trades on external validity. In order that the 1990s may speak to the Great Depression, they assume that the actual history of the variables and the causal topology are the same in both periods. And they assume that the parameterization, except for the parameterization of the monetary-policy rule itself (the value of \( \lambda \)) is also the same.

Hoover and Jordá could easily be wrong: the conditions that underwrite external validity may fail. But that is an issue on which empirical evidence can be brought to bear. It is not a special problem for causal analysis but a more general problem for the import of empirical results derived in one set of circumstances (say, in a particular experiment) for other sets of circumstances. Our current concern is not with the problem of external validity but with the problem of using causes in situations in which the external validity of the causal model is not in question.

A good deal of Cartwright’s skepticism about causal knowledge in economics and – one presumes – in other fields is apparently generated by a lack of sufficient respect for her own distinction between hunting and using. She writes as if the manner in which causes are hunted limits their possible uses. The structural account, however, clarifies that there are a variety of things we typically need to know to have a useful representation of causal structure. We need to know the causal topology – essentially the pattern of arrows connecting variables or, equivalently, the parameterization (for example, that the parameter space includes \( \alpha \) and \( \lambda \) and not, say, \( \theta = \alpha \lambda \)). We also need to know the
functional interrelationship of the parameters, including the manner of potential nonmodularity. And we need to know, for any real-world counterfactuals, the actual values of the parameters. The structural account tells us both what we need to know and where to slot such knowledge as we have obtained into the representation of causal structure.

Cartwright (2007, p. 9, *passim*) characterizes her position as “causal pluralism.” The structural account aims at a high enough level of generality that any coherent approaches are nested within it. Yet, it is compatible with substantial methodological pluralism: different methods may supply different elements of the knowledge needed to fill in the causal structure. For example, Bayes net methods, as discussed earlier, are helpful in mapping the causal topology. But there are situations – well known to their advocates – in which they are not discriminating, and they do not directly address parameter values. Hoover (2001, ch. 8) offers methods that use patterns of invariance and noninvariance across regime changes that can sometimes resolve the equivalent causal topologies allowed by Bayes net methods. Hoover and Jordá (2001) demonstrate in an enriched version of the model (8)-(11) that knowledge of interventions in the monetary-policy rule (not even the fine details, but simply the timing of when they occur) may allow us to recover the functional relationships among parameters of nonmodular causal systems. Their approach is an empirical analogue to Woodward’s use of manipulations in the evaluation of causal counterfactuals, although it does not involve breaking of causal arrows, but in the manner of section 2 above, considers manipulations in a conserved causal topology.
These empirical methods, at various points, all involve untested assumptions – a number of which have been mentioned already with respect to Hoover and Jordá’s counterfactual experiments. But then so does all empirical investigation. These assumptions may not be tested in a particular study, but they are not necessarily untestable or, at least, not necessarily beyond empirically based criticism. They are not, however, all jointly testable at the same time. Only a thoroughly destructive skeptic would be unwilling to make some assumptions that seem reasonable and reliable until there exist reasons to doubt their truth more compelling than the mere possibility that they could be false.

3.4. EPISTEMIC OPPORTUNISM
While Cartwright takes the fact that some methods work well only for modular, “epistemically convenient” situations to be a significant drawback, from a practitioner’s point of view it is surprising, but welcome, how often the real world seems to be convenient enough to make empirical progress with relatively simple methods. Cartwright is certainly correct that modularity is not a general property of causation, but it is common enough – to a reasonable approximation – that methods that require it are often practically effective. And where modularity fails, there are other methods, such as the methods based on invariance testing advocated by Hoover (2001) and methods built on similar principles used by Hoover and Jordá (2001). Some nuts have not yet been cracked; some perhaps never will be. Rather than decrying methods that require “epistemic convenience” generally, it would be better to embrace epistemic opportunism: articulate causal models by any means necessary. The structural account gives us a systematic way to interpret what appropriate methods have accomplished.
References


Figure 1

\[ \begin{align*}
A \rightarrow C \\
C \rightarrow D
\end{align*} \]

\[ \begin{align*}
B \rightarrow C \\
C \rightarrow D
\end{align*} \]

Figure 2

\[ \begin{align*}
a \rightarrow C \\
C \rightarrow D
\end{align*} \]

\[ \begin{align*}
b \rightarrow C \\
C \rightarrow D
\end{align*} \]

Figure 3

\[ \begin{align*}
A \leftrightarrow B \\
A \rightarrow C \\
C \rightarrow D
\end{align*} \]

\[ \begin{align*}
B \rightarrow C \\
C \rightarrow D
\end{align*} \]
Figure 4

Panel A

\[
\begin{align*}
\text{Panel A} & \\
CT & \quad + \\
D1 &
\end{align*}
\]

Panel B

\[
\begin{align*}
\text{Panel B} & \\
CT & \quad + \\
BP & \quad - \\
D1 &
\end{align*}
\]

Panel C

\[
\begin{align*}
\text{Panel C} & \\
CT & \quad + \\
BP & \quad - \\
D2 & \quad - \\
D1 &
\end{align*}
\]
Figure 5

\[ A \rightarrow B \rightarrow C \]