In cases where we tend to assert categorically that the flow of causation in a feedback loop goes clockwise, this assertion is normally based on the relative magnitudes of forces. For example, turning the faucet would lower the water level in the water tank, but there is practically nothing we can do to the water tank that would turn the faucet. When such information is available, causal directionality is determined by appealing, again, to the notion of hypothetical intervention and asking whether an external control over one variable in the mechanism necessarily affects the others. This consideration then constitutes the operational semantics for identifying the dependent variables V_i in nonrecursive causal models (Definition 7.1.1).

The asymmetry that characterizes causal relationships in no way conflicts with the symmetry of physical equations. By saying that "X causes Y and Y does not cause X," we mean to say that changing a mechanism in which X is normally the dependent variable has a different effect on the world than changing a mechanism in which Y is normally the dependent variable. Because two separate mechanisms are involved, the statement stands in perfect harmony with the symmetry we find in the equations of physics.

Simon's theory of causal ordering has profound repercussions on Hume's problem of causal induction, that is, how causal knowledge is acquired from experience (see Chapter 2). The ability to deduce causal directionality from an assembly of symmetrical mechanisms (together with a selection of a set of endogenous variables) means that the acquisition of causal relationships is no different than the acquisition (e.g., by experiments) of ordinary physical laws, such as Hooke's law of suspended springs or Newton's law of acceleration. This does not imply that acquiring physical laws is a trivial task, free of methodological and philosophical subtleties. It does imply that the problem of causal induction – one of the toughest in the history of philosophy – can be reduced to the more familiar problem of scientific induction.

7.3 AXIOMATIC CHARACTERIZATION

Axioms play important roles in the characterization of formal systems. They provide a parsimonious description of the essential properties of the system, thus allowing comparisons among alternative formulations and easy tests of equivalence or subsumption among such alternatives. Additionally, axioms can often be used as rules of inference for deriving (or verifying) new relationships from a given set of premises. In the next subsection, we will establish a set of axioms that characterize the relationships among counterfactual sentences of the form $Y_x(u) = y$ in both recursive and nonrecursive systems. Using these axioms, we will then demonstrate (in Section 7.3.2) how the identification of causal effects can be verified by symbolic means, paralleling the derivations of Chapter 3 (Section 3.4). Finally, Section 7.3.3 establishes axioms for the notion of *causal relevance,* contrasting those that capture informational relevance.

7.3.1 The Axioms of Structural Counterfactuals

We present three properties of counterfactuals – composition, effectiveness, and reversibility – that hold in all causal models.

Property 1 (Composition)
For any three sets of endogenous variables X, Y, and W in a causal model, we have

$$W_x(u) = w \implies Y_{xw}(u) = Y_x(u). \tag{7.19}$$

Composition states that, if we force a variable (W) to a value w that it would have had without our intervention, then the intervention will have no effect on other variables in the system. That invariance holds in all fixed conditions $do(X = x)$.

Since composition allows for the removal of a subscript (i.e., reducing $Y_{xw}(u)$ to $Y_x(u)$), we need an interpretation for a variable with an empty set of subscripts, which (naturally) we identify with the variable under no interventions.

Definition 7.3.1 (Null Action)

$$Y_\emptyset(u) \triangleq Y(u).$$

Corollary 7.3.2 (Consistency)
For any set of variables Y and X in a causal model, we have

$$X(u) = x \implies Y(u) = Y_x(u). \tag{7.20}$$

Proof
Substituting X for W and $\emptyset$ for X in (7.19), we obtain $X_\emptyset(u) = x \implies Y_\emptyset(u) = Y_x(u)$. Null action (Definition 7.3.1) allows us to drop the $\emptyset$, leaving $X(u) = x \implies Y(u) = Y_x(u)$. □

The implication in (7.20) was called "consistency" by Robins (1987).[13]

Property 2 (Effectiveness)
For all sets of variables X and W, $X_{xw}(u) = x$.

Effectiveness specifies the effect of an intervention on the manipulated variable itself – namely, that if we force a variable X to have the value x, then X will indeed take on the value x.

Property 3 (Reversibility)
For any two variables Y and W and any set of variables X,

$$(Y_{xw}(u) = y) \ \& \ (W_{xy}(u) = w) \implies Y_x(u) = y. \tag{7.21}$$

Reversibility precludes multiple solutions due to feedback loops. If setting W to a value w results in a value y for Y, and if setting Y to the value y results in W achieving the

[13] Consistency and composition are used routinely in economics (Manski 1990; Heckman 1996) and statistics (Rosenbaum 1995) within the potential-outcome framework (Section 3.6.3). Consistency was stated formally by Gibbard and Harper (1976, p. 156) and Robins (1987) (see equation (3.52)). Composition is stated in Holland (1986, p. 968) and was brought to my attention by J. Robins.

value w, then W and Y will naturally obtain the values w and y (respectively), without any external setting. In recursive systems, reversibility follows directly from composition. This can easily be seen by noting that, in a recursive system, either $Y_{xw}(u) = Y_x(u)$ or $W_{xy}(u) = W_x(u)$. Thus, reversibility reduces to $(Y_{xw}(u) = y)\ \&\ (W_x(u) = w) \Longrightarrow Y_x(u) = y$ (another form of composition) or to $(Y_x(u) = y)\ \&\ (W_{xy}(u) = w) \Longrightarrow Y_x(u) = y$ (which is trivially true).

Reversibility reflects "memoryless" behavior: the state of the system, V, tracks the state of U regardless of U's history. A typical example of irreversibility is a system of two agents who adhere to a "tit-for-tat" strategy (e.g., the prisoners' dilemma). Such a system has two stable solutions – cooperation and defection – under the same external conditions U, and thus it does not satisfy the reversibility condition; forcing either one of the agents to cooperate results in the other agent's cooperation $(Y_w(u) = y, W_y(u) = w)$, yet this does not guarantee cooperation from the start $(Y(u) = y, W(u) = w)$. In such systems, irreversibility is a product of using a state description that is too coarse, one where not all of the factors that determine the ultimate state of the system are included in U. In a tit-for-tat system, a complete state description should include factors such as the previous actions of the players, and reversibility is restored once the missing factors are included.

In general, the properties of composition, effectiveness, and reversibility are independent – none is a consequence of the other two. This can be shown (Galles and Pearl 1997) by constructing specific models in which two of the properties hold and the third does not. In recursive systems, composition and effectiveness are independent while reversibility holds trivially, as just shown.

The next theorem asserts the *soundness*[14] of properties 1–3, that is, their validity.

Theorem 7.3.3 (Soundness)
Composition, effectiveness, and reversibility are sound *in structural model semantics; that is, they hold in all causal models.*

A proof of Theorem 7.3.3 is given in Galles and Pearl (1997).

Our next theorem establishes the *completeness* of the three properties when treated as axioms or rules of inference. Completeness amounts to sufficiency; *all* other properties of counterfactual statements follow from these three. Another interpretation of completeness is as follows: Given any set S of counterfactual statements that is consistent with properties 1–3, there exists a causal model M in which S holds true.

A formal proof of completeness requires the explication of two technical properties – existence and uniqueness – that are implicit in the definition of causal models (Definition 7.1.1).

Property 4 (Existence)
For any variable X and set of variables Y,

$$\exists x \in X \text{ s.t. } X_y(u) = x. \tag{7.22}$$

[14] The terms *soundness* and *completeness* are sometimes referred to as *necessity* and *sufficiency*, respectively.

Property 5 (Uniqueness)
For every variable X and set of variables Y,

$$X_y(u) = x\ \&\ X_y(u) = x' \Longrightarrow x = x'. \tag{7.23}$$

Definition 7.3.4 (Recursiveness)
A model M is recursive *if, for any two variables Y and W and for any set of variables X, we have*

$$Y_{xw}(u) = Y_x(u) \quad \text{or} \quad W_{xy}(u) = W_x(u). \tag{7.24}$$

In words, recursiveness means that either Y does not affect W or W does not affect Y. Clearly, any model M for which the causal diagram $G(M)$ is acyclic must be recursive.

Theorem 7.3.5 (Recursive Completeness)
Composition, effectiveness, and recursiveness are complete (Galles and Pearl 1998; Halpern 1998).[15]

Theorem 7.3.6 (Completeness)
Composition, effectiveness, and reversibility are complete for all causal models (Halpern 1998).

The practical importance of soundness and completeness surfaces when we attempt to test whether a certain set of conditions is sufficient for the identifiability of some counterfactual quantity Q. Soundness, in this context, guarantees that if we symbolically manipulate Q using the three axioms and manage to reduce it to an expression that involves ordinary probabilities (free of counterfactual terms), then Q is identifiable (in the sense of Definition 3.2.3). Completeness guarantees the converse: if we do not succeed in reducing Q to a probabilistic expression, then Q is nonidentifiable – our three axioms are as powerful as can be.

The next section demonstrates a proof of identifiability that uses effectiveness and decomposition as axioms.

7.3.2 Causal Effects from Counterfactual Logic: An Example

We revisit the smoking–cancer example analyzed in Section 3.4.3. The model associated with this example is assumed to have the following structure (see Figure 7.5):

$$V = \{X \text{ (smoking)}, Y \text{ (lung cancer)}, Z \text{ (tar in lungs)}\},$$
$$U = \{U_1, U_2\},\ U_1 \perp\!\!\!\perp U_2,$$

[15] Galles and Pearl (1997) proved recursive completeness assuming that, for any two variables, one knows which of the two (if any) is an ancestor of the other. Halpern (1998) proved recursive completeness without this assumption, provided only that (7.24) is known to hold for any two variables in the model. Halpern further provided a set of axioms for cases where the solution of $Y_x(u)$ is not unique or does not exist.

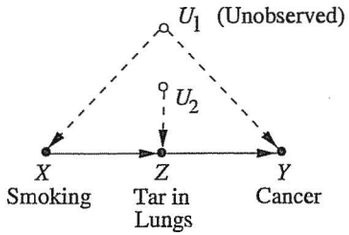

Figure 7.5 Causal diagram illustrating the effect of smoking on lung cancer.

$$x = f_1(u_1),$$
$$z = f_2(x, u_2),$$
$$y = f_3(z, u_1).$$

This model embodies several assumptions, all of which are represented in the diagram of Figure 7.5. The missing link between X and Y represents the assumption that the effect of smoking cigarettes (X) on the production of lung cancer (Y) is entirely mediated through tar deposits in the lungs. The missing connection between U_1 and U_2 represents the assumption that even if a genotype (U_1) is aggravating the production of lung cancer, it nevertheless has no effect on the amount of tar in the lungs except indirectly (through cigarette smoking). We wish to use the assumptions embodied in the model to derive an estimable expression for the causal effect $P(Y = y \mid do(x)) \triangleq P(Y_x = y)$ that is based on the joint distribution $P(x, y, z)$.

This problem was solved in Section 3.4.3 by a graphical method, using the axioms of *do* calculus (Theorem 3.4.1). Here we show how the counterfactual expression $P(Y_x = y)$ can be reduced to ordinary probabilistic expression (involving no counterfactuals) by purely symbolic operations, using only probability calculus and two rules of inference: effectiveness and composition. Toward this end, we first need to translate the assumptions embodied in the graphical model into the language of counterfactuals. In Section 3.6.3 it was shown that the translation can be accomplished systematically, using two simple rules (Pearl 1995a, p. 704).

Rule 1 (*exclusion restrictions*): For every variable Y having parents PA_Y and for every set of variables $Z \subset V$ disjoint of PA_Y, we have

$$Y_{pa_Y}(u) = Y_{pa_Y z}(u). \tag{7.25}$$

Rule 2 (*independence restrictions*): If $Z_1, \ldots, Z_k$ is any set of nodes in V not connected to Y via paths containing only U variables, we have

$$Y_{pa_Y} \perp\!\!\!\perp \{Z_{1pa_{Z_1}}, \ldots, Z_{kpa_{Z_k}}\}. \tag{7.26}$$

Equivalently, (7.26) holds if the corresponding U terms $(U_{Z_1}, \ldots, U_{Z_k})$ are jointly independent of U_Y.

Rule 1 reflects the insensitivity of Y to any manipulation in V, once its direct causes PA_Y are held constant; it follows from the identity $v_i = f_i(pa_i, u_i)$ in Definition 7.1.1. Rule 2 interprets independencies among U variables as independencies between the counterfactuals of the corresponding V variables, with their parents held fixed. Indeed, the statistics

of Y_{pa_Y} is governed by the equation $Y = f_Y(pa_Y, u_Y)$; therefore, once we hold PA_Y fixed, the residual variations of Y are governed solely by the variations in U_Y.

Applying these two rules to our example, we see that the causal diagram in Figure 7.5 encodes the following assumptions:

$$Z_x(u) = Z_{yx}(u), \tag{7.27}$$
$$X_y(u) = X_{zy}(u) = X_z(u) = X(u), \tag{7.28}$$
$$Y_z(u) = Y_{zx}(u), \tag{7.29}$$
$$Z_x \perp\!\!\!\perp \{Y_z, X\}. \tag{7.30}$$

Equations (7.27)–(7.29) follow from the exclusion restrictions of (7.25), using

$$PA_X = \emptyset, \quad PA_Y = \{Z\}, \quad \text{and} \quad PA_Z = \{X\}.$$

Equation (7.27), for instance, represents the absence of a causal link from Y to Z, while (7.28) represents the absence of a causal link from Z or Y to X. In contrast, (7.30) follows from the independence restriction of (7.26), since the lack of a connection between (i.e., the independence of) U_1 and U_2 rules out any path between Z and $\{X, Y\}$ that contains only U variables.

We now use these assumptions (which embody recursiveness), together with the properties of composition and effectiveness, to compute the tasks analyzed in Section 3.4.3.

Task 1

Compute $P(Z_x = z)$ (i.e., the causal effect of smoking on tar).

$$\begin{aligned} P(Z_x = z) &= P(Z_x = z \mid x) && \text{from (7.30)} \\ &= P(Z = z \mid x) && \text{by composition} \\ &= P(z \mid x). \end{aligned} \tag{7.31}$$

Task 2

Compute $P(Y_z = y)$ (i.e., the causal effect of tar on cancer).

$$P(Y_z = y) = \sum_x P(Y_z = y \mid x) P(x). \tag{7.32}$$

Since (7.30) implies $Y_z \perp\!\!\!\perp Z_x \mid X$, we can write

$$\begin{aligned} P(Y_z = y \mid x) &= P(Y_z = y \mid x, Z_x = z) && \text{from (7.30)} \\ &= P(Y_z = y \mid x, z) && \text{by composition} \\ &= P(y \mid x, z). && \text{by composition} \end{aligned} \tag{7.33}$$

Substituting (7.33) into (7.32) yields

$$P(Y_z = y) = \sum_x P(y \mid x, z) P(x). \tag{7.34}$$

Task 3

Compute $P(Y_x = y)$ (i.e., the causal effect of smoking on cancer).

For any variable Z, by composition we have

$$Y_x(u) = Y_{xz}(u) \quad \text{if } Z_x(u) = z.$$

Since $Y_{xz}(u) = Y_z(u)$ (from (7.29)),

$$Y_x(u) = Y_{xz_x}(u) = Y_z(u), \quad \text{where } z_x = Z_x(u). \tag{7.35}$$

Thus,

$$\begin{aligned} P(Y_x = y) &= P(Y_{z_x} = y) && \text{from (7.35)} \\ &= \sum_z P(Y_{z_x} = y \mid Z_x = z) P(Z_x = z) \\ &= \sum_z P(Y_z = y \mid Z_x = z) P(Z_x = z) && \text{by composition} \\ &= \sum_z P(Y_z = y) P(Z_x = z) && \text{from (7.30)} \end{aligned} \tag{7.36}$$

The probabilities $P(Y_z = y)$ and $P(Z_x = z)$ were computed in (7.34) and (7.31), respectively. Substituting gives us

$$P(Y_x = y) = \sum_z P(z \mid x) \sum_{x'} P(y \mid z, x') P(x'). \tag{7.37}$$

The right-hand side of (7.37) can be computed from $P(x, y, z)$ and coincides with the front-door formula derived in Section 3.4.3 (equation (3.42)).

Thus, $P(Y_x = y)$ can be reduced to expressions involving probabilities of observed variables and is therefore identifiable. More generally, our completeness result (Theorem 7.3.5) implies that *any* identifiable counterfactual quantity can be reduced to the correct expression by repeated application of composition and effectiveness (assuming recursiveness).

7.3.3 Axioms of Causal Relevance

In Section 1.2 we presented a set of axioms for a class of relations called *graphoids* (Pearl and Paz 1987; Geiger et al. 1990) that characterize informational relevance.[16] We now develop a parallel set of axioms for *causal relevance,* that is, the tendency of certain events to affect the occurrence of other events in the physical world, independent of the observer–reasoner. Informational relevance is concerned with questions of the form: "Given that we know Z, would gaining information about X gives us new information about Y?" Causal relevance is concerned with questions of the form: "Given that Z is fixed, would changing X alter Y?" We show that causal relevance complies with all the axioms of path interception in directed graphs except transitivity.

[16] "Relevance" will be used primarily as a generic name for the relationship of being relevant or irrelevant. It will be clear from the context when "relevance" is intended to negate "irrelevance."

The notion of causal relevance has its roots in the philosophical works of Suppes (1970) and Salmon (1984), who attempted to give probabilistic interpretations to cause–effect relationships and recognized the need to distinguish causal from statistical relevance (see Section 7.5). Although these attempts did not produce a probabilistic definition of causal relevance, they led to methods for testing the consistency of relevance statements against a given probability distribution and a given temporal ordering among the variables (see Section 7.5.2). Here we aim at axiomatizing relevance statements in themselves – with no reference to underlying probabilities or temporal orderings.

The axiomization of causal relevance may be useful to experimental researchers in domains where exact causal models do not exist. If we know, through experimentation, that some variables have no causal influence on others in a system, then we may wish to determine whether other variables will exert causal influence (perhaps under different experimental conditions) or may ask what additional experiments could provide such information. For example, suppose we find that a rat's diet has no effect on tumor growth while the amount of exercise is kept constant and, conversely, that exercise has no effect on tumor growth while diet is kept constant. We would like to be able to infer that controlling only diet (while paying no attention to exercise) would still have no influence on tumor growth. A more subtle inference problem is deciding whether changing the ambient temperature in the cage would have an effect on the rat's physical activity, given that we have established that temperature has no effect on activity when diet is kept constant and that temperature has no effect on (the rat's choice of) diet when activity is kept constant.

Galles and Pearl (1997) analyzed both probabilistic and deterministic interpretations of causal irrelevance. The probabilistic interpretation, which equates causal irrelevance with inability to change the probability of the effect variable, has intuitive appeal but is inferentially very weak; it does not support a very expressive set of axioms unless further assumptions are made about the underlying causal model. If we add the stability assumption (i.e., that no irrelevance can be destroyed by changing the nature of the individual processes in the system), then we obtain the same set of axioms for probabilistic causal irrelevance as the set governing path interception in directed graphs.

In this section we analyze a deterministic interpretation that equates causal irrelevance with inability to change the effect variable in any state u of the world. This interpretation is governed by a rich set of axioms without our making any assumptions about the causal model: many of the path interception properties in directed graphs hold for deterministic causal irrelevance.

Definition 7.3.7 (Causal Irrelevance)

A variable X is causally irrelevant *to Y, given Z (written $X \not\to Y \mid Z$) if, for every set W disjoint of $X \cup Y \cup Z$, we have*

$$\forall(u, z, x, x', w), \quad Y_{xzw}(u) = Y_{x'zw}(u), \tag{7.38}$$

where x and x' are two distinct values of X.

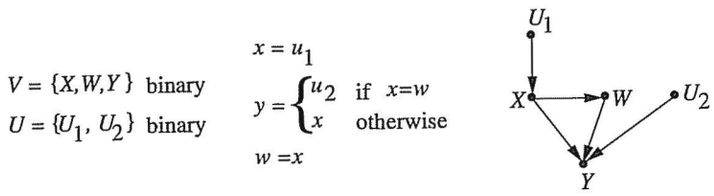

Figure 7.6 Example of a causal model that requires the examination of submodels to determine causal relevance.

This definition captures the intuition "If X is causally irrelevant to Y, then X cannot affect Y under any circumstance u or under any modification of the model that includes $do(Z = z)$."

To see why we require the equality $Y_{xzw}(u) = Y_{x'zw}(u)$ to hold in every context $W = w$, consider the causal model of Figure 7.6. In this example, $Z = \emptyset$, W follows X, and hence Y follows X; that is, $Y_{X=0}(u) = Y_{X=1}(u) = u_2$. However, since $y(x, w, u_2)$ is a nontrivial function of x, X is perceived to be causally relevant to Y. Only holding W constant would reveal the causal influence of X on Y. To capture this intuition, we must consider all contexts $W = w$ in Definition 7.3.7.

With this definition of causal irrelevance, we have the following theorem.

Theorem 7.3.8

For any causal model, the following sentences must hold.

Weak Right Decomposition:[17]

$$(X \not\to YW \mid Z) \;\&\; (X \to Y \mid ZW) \implies (X \not\to Y \mid Z).$$

Left Decomposition:

$$(XW \not\to Y \mid Z) \implies (X \not\to Y \mid Z) \;\&\; (W \not\to Y \mid Z).$$

Strong Union:

$$(X \not\to Y \mid Z) \implies (X \not\to Y \mid ZW)\ \forall W.$$

Right Intersection:

$$(X \not\to Y \mid ZW) \;\&\; (X \not\to W \mid ZY) \implies (X \not\to YW \mid Z).$$

Left Intersection:

$$(X \not\to Y \mid ZW) \;\&\; (W \not\to Y \mid ZX) \implies (XW \not\to Y \mid Z).$$

This set of axioms bears a striking resemblance to the properties of path interception in a directed graph. Paz and Pearl (1994) showed that the axioms of Theorem 7.3.8, together with transitivity and right decomposition, constitute a complete characterization of the

[17] Galles and Pearl (1997) used a stronger version of right decomposition: $(X \not\to YW \mid Z) \implies (X \not\to Y \mid Z)$. But Bonet (1999) showed that it must be weakened to render the axiom system sound.

relation $(X \not\to Y \mid Z)_G$ when interpreted to mean that every directed path from X to Y in a directed graph G contains at least one node in Z (see also Paz et al. 1996).

Galles and Pearl (1997) showed that, despite the absence of transitivity, Theorem 7.3.8 permits one to infer certain properties of causal irrelevance from properties of directed graphs. For example, suppose we wish to validate a generic statement such as: "If X has an effect on Y, but ceases to have an effect when we fix Z, then Z must have an effect on Y." That statement can be proven from the fact that, in any directed graph, if all paths from X to Y are intercepted by Z and there are no paths from Z to Y, then there is no path from X to Y.

Remark on the Transitivity of Causal Dependence

That causal dependence is not transitive is clear from Figure 7.6. In any state of (U_1, U_2), X is capable of changing the state of W and W is capable of changing Y, yet X is incapable of changing Y. Galles and Pearl (1997) gave examples where causal relevance in the weak sense of Definition 7.3.7 is also nontransitive, even for binary variables. The question naturally arises as to why transitivity is so often conceived of as an inherent property of causal dependence or, more formally, what assumptions we tacitly make when we classify causal dependence as transitive.

One plausible answer is that we normally interpret transitivity to mean the following: "If (1) X causes Y and (2) Y causes Z regardless of X, then (3) X causes Z." The suggestion is that questions about transitivity bring to mind chainlike processes, where X influences Y and Y influences Z but where X does not have a *direct* influence over Z. With this qualification, transitivity for binary variables can be proven immediately from composition (equation (7.19)) as follows.

Let the sentence "$X = x$ causes $Y = y$," denoted $x \to y$, be interpreted as the joint condition $\{X(u) = x, Y(u) = y, Y_{x'}(u) = y' \neq y\}$ (in words, x and y hold, but changing x to x' would change y to y'). We can now prove that if X has no direct effect on Z, that is, if

$$Z_{y'x'} = Z_{y'}, \tag{7.39}$$

then

$$x \to y \;\&\; y \to z \implies x \to z. \tag{7.40}$$

Proof

The l.h.s. of (7.40) reads

$$X(u) = x, \quad Y(u) = y, \quad Z(u) = z, \quad Y_{x'}(u) = y', \quad Z_{y'}(u) = z'.$$

From (7.39) we can rewrite the last term as $Z_{y'x'}(u) = z'$. Composition further permits us to write

$$Y_{x'}(u) = y' \;\&\; Z_{y'x'}(u) = z' \implies Z_{x'}(u) = z',$$

which, together with $X(u) = x$ and $Z(u) = z$, implies $x \to z$. □

Weaker forms of causal transitivity are discussed in Chapter 9 (Lemmas 9.2.7 and 9.2.8).

7.4 STRUCTURAL AND SIMILARITY-BASED COUNTERFACTUALS

7.4.1 Relations to Lewis's Counterfactuals

Causality from Counterfactuals

In one of his most quoted sentences, David Hume tied together two aspects of causation, regularity of succession and counterfactual dependency:

> we may define a cause to be an object followed by another, and where all the objects, similar to the first, are followed by object similar to the second, Or, in other words, where, if the first object had not been, the second never had existed. (Hume 1748/1959, sec. VII).

This two-faceted definition is puzzling on several accounts. First, regularity of succession, or "correlation" in modern terminology, is not sufficient for causation, as even nonstatisticians know by now. Second, the expression "in other words" is a too strong, considering that regularity rests on observations whereas counterfactuals rest on mental exercise. Third, Hume had introduced the regularity criterion nine years earlier,[18] and one wonders what jolted him into supplementing it with a counterfactual companion. Evidently, Hume was not completely happy with the regularity account, and must have felt that the counterfactual criterion is less problematic and more illuminating. But how can convoluted expressions of the type "if the first object had not been, the second never had existed" illuminate simple commonplace expressions like "A caused B"?

The idea of basing causality on counterfactuals is further echoed by John Stuart Mill (1843), and it reached fruition in the works of David Lewis (1973b, 1986). Lewis called for abandoning the regularity account altogether and for interpreting "A has caused B" as "B would not have occurred if it were not for A." Lewis (1986, p. 161) asked: "Why not take counterfactuals at face value: as statements about possible alternatives to the actual situation ...?"

Implicit in this proposal lies a claim that counterfactual expressions are less ambiguous to our mind than causal expressions. Why else would the expression "B would be false if it were not for A" be considered an *explication* of "A caused B," and not the other way around, unless we could discern the truth of the former with greater certitude than that of the latter? Taken literally, discerning the truth of counterfactuals requires generating and examining possible alternatives to the actual situation as well as testing whether certain propositions hold in those alternatives – a mental task of nonnegligible proportions. Nonetheless, Hume, Mill, and Lewis apparently believed that going through this mental exercise is simpler than intuiting directly on whether it was A that caused B. How can this be done? What mental representation allows humans to process counterfactuals so swiftly and reliably, and what logic governs that process so as to maintain uniform standards of coherence and plausibility?

[18] In *Treatise of Human Nature,* Hume wrote: "We remember to have had frequent instances of the existence of one species of objects; and also remember, that the individuals of another species of objects have always attended them, and have existed in a regular order of contiguity and succession with regard to them" (Hume 1739, p. 156).

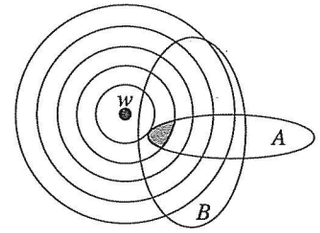

Figure 7.7 Graphical representation of Lewis's closest-world semantics. Each circular region corresponds to a set of worlds that are equally similar to w. The shaded region represents the set of closest A-worlds; since all these worlds satisfy B, the counterfactual sentence $A \mathrel{\Box\!\!\rightarrow} B$ is declared true in w.

Structure versus Similarity

According to Lewis's account (1973b), the evaluation of counterfactuals involves the notion of *similarity*: one orders possible worlds by some measure of similarity, and the counterfactual $A \mathrel{\Box\!\!\rightarrow} B$ (read: "B if it were A") is declared true in a world w just in case B is true in all the closest A-worlds to w (see Figure 7.7).[19]

This semantics still leaves questions of representation unsettled. What choice of similarity measure would make counterfactual reasoning compatible with ordinary conceptions of cause and effect? What mental representation of worlds ordering would render the computation of counterfactuals manageable and practical (in both man and machine)?

In his initial proposal, Lewis was careful to keep the formalism as general as possible; save for the requirement that every world be closest to itself, he did not impose any structure on the similarity measure. However, simple observations tell us that similarity measures cannot be arbitrary. The very fact that people communicate with counterfactuals already suggests that they share a similarity measure, that this measure is encoded parsimoniously in the mind, and hence that it must be highly structured. Kit Fine (1975) further demonstrated that similarity of appearance is inadequate. Fine considers the counterfactual "Had Nixon pressed the button, a nuclear war would have started," which is generally accepted as true. Clearly, a world in which the button happened to be disconnected is many times more similar to our world, as we know it, than the one yielding a nuclear blast. Thus we see not only that similarity measures could not be arbitrary but also that they must respect our conception of causal laws.[20] Lewis (1979) subsequently set up an intricate system of weights and priorities among various aspects of similarity – size of "miracles" (violations of laws), matching of facts, temporal precedence, and so forth – in attempting to bring similarity closer to causal intuition. But these priorities are rather post hoc and still yield counterintuitive inferences (J. Woodward, personal communication).

Such difficulties do not enter the structural account. In contrast with Lewis's theory, counterfactuals are not based on an abstract notion of similarity among hypothetical worlds; instead, they rest directly on the mechanisms (or "laws," to be fancy) that produce those worlds and on the invariant properties of those mechanisms. Lewis's elusive "miracles" are replaced by principled minisurgeries, $do(X = x)$, which represent the minimal change (to a model) necessary for establishing the antecedent $X = x$ (for all u).

[19] Related possible-world semantics were introduced in artificial intelligence to represent actions and database updates (Ginsberg 1986; Ginsberg and Smith 1987; Winslett 1988; Katsuno and Mendelzon 1991).

[20] In this respect, Lewis's reduction of causes to counterfactuals is somewhat circular.

Thus, similarities and priorities – if they are ever needed – may be read into the $do(\cdot)$ operator as an afterthought (see discussion following (3.11) and Goldszmidt and Pearl 1992), but they are not basic to the analysis.

The structural account answers the mental representation question by offering a parsimonious encoding of knowledge from which causes, counterfactuals, and probabilities of counterfactuals can be derived by effective algorithms. However, this parsimony is acquired at the expense of generality; limiting the counterfactual antecedent to conjunction of elementary propositions prevents us from analyzing disjunctive hypotheticals such as "if Bizet and Verdi were compatriots."

7.4.2 Axiomatic Comparison

If our assessment of interworld distances comes from causal knowledge, the question arises of whether that knowledge does not impose its own structure on distances, a structure that is not captured in Lewis's logic. Phrased differently: By agreeing to measure closeness of worlds on the basis of causal relations, do we restrict the set of counterfactual statements we regard as valid? The question is not merely theoretical. For example, Gibbard and Harper (1976) characterized decision-making conditionals (i.e., sentences of the form "If we do A, then B") using Lewis's general framework, whereas our $do(\cdot)$ operator is based directly on causal mechanisms; whether the two formalisms are identical is uncertain.[21]

We now show that the two formalisms are identical for recursive systems; in other words, composition and effectiveness hold with respect to Lewis's closest-world framework whenever recursiveness does. We begin by providing a version of Lewis's logic for counterfactual sentences (from Lewis 1973c).

Rules

(1) If A and $A \implies B$ are theorems, then so is B.

(2) If $(B_1 \ \& \ \cdots) \implies C$ is a theorem, then so is $((A \boxrightarrow B_1)\cdots) \implies (A \boxrightarrow C)$.

Axioms

(1) All truth-functional tautologies.

(2) $A \boxrightarrow A$.

(3) $(A \boxrightarrow B) \ \& \ (B \boxrightarrow A) \implies (A \boxrightarrow C) \equiv (B \boxrightarrow C)$.

(4) $((A \vee B) \boxrightarrow A) \vee ((A \vee B) \boxrightarrow B) \vee$
$(((A \vee B) \boxrightarrow C) \equiv (A \boxrightarrow C) \ \& \ (B \boxrightarrow C))$.

(5) $A \boxrightarrow B \implies A \implies B$.

(6) $A \ \& \ B \implies A \boxrightarrow B$.

[21] Ginsberg and Smith (1987) and Winslett (1988) have also advanced theories of actions based on closest-world semantics; they have not imposed any special structure for the distance measure to reflect causal considerations.

The statement $A \boxrightarrow B$ stands for "In all closest worlds where A holds, B holds as well." To relate Lewis's axioms to those of causal models, we must translate his syntax. We will equate Lewis's world with an instantiation of all the variables, including those in U, in a causal model. Values assigned to subsets of variables in a causal model will stand for Lewis's propositions (e.g., A and B in the stated rules and axioms). Thus, let A stand for the conjunction $X_1 = x_1, \ldots, X_n = x_n$, and let B stand for the conjunction $Y_1 = y_1, \ldots, Y_m = y_m$. Then

$$\begin{aligned} A \boxrightarrow B \equiv\ & Y_{1_{x_1,\ldots,x_n}}(u) = y_1 \\ & \ \& \ Y_{2_{x_1,\ldots,x_n}}(u) = y_2 \\ & \ \vdots \\ & \ \& \ Y_{m_{x_1,\ldots,x_n}}(u) = y_m. \end{aligned} \tag{7.41}$$

Conversely, we need to translate causal statements such as $Y_x(u) = y$ into Lewis's notation. Let A stand for the proposition $X = x$ and B for the proposition $Y = y$. Then

$$Y_x(u) = y \equiv A \boxrightarrow B. \tag{7.42}$$

Axioms (1)–(6) follow from the closest-world interpretation without imposing any restrictions on the distance measured, except for the requirement that each world w be no further from itself than any other world $w' \neq w$. Because structural semantics defines an obvious distance measure among worlds, $d(w, w')$, given by the minimal number of local interventions needed for transforming w into w', all of Lewis's axioms should hold in causal models and must follow logically from effectiveness, composition, and (for nonrecursive systems) reversibility. This will be shown explicitly first. However, to guarantee that structural semantics does not introduce new constraints we need to show the converse: that the three axioms of structural semantics follow from Lewis's axioms. This will be shown second.

To show that Axioms (1)–(6) hold in structural semantics, we examine each axiom in turn.

(1) This axiom is trivially true.

(2) This axiom is the same as effectiveness: If we force a set of variables X to have the value x, then the resulting value of X is x. That is, $X_x(u) = x$.

(3) This axiom is a weaker form of reversibility, which is relevant only for nonrecursive causal models.

(4) Because actions in structural models are restricted to conjunctions of literals, this axiom is irrelevant.

(5) This axiom follows from composition.

(6) This axiom follows from composition.

To show that composition and effectiveness follow from Lewis's axioms, we note that composition is a consequence of axiom (5) and rule (1) in Lewis's formalism, while effectiveness is the same as Lewis's axiom (2).

In sum, for recursive models, the causal model framework does not add any restrictions to counterfactual statements beyond those imposed by Lewis's framework; the very general concept of closest worlds is sufficient. Put another way, the assumption of recursiveness is so strong that it already embodies all other restrictions imposed by structural semantics. When we consider nonrecursive systems, however, we see that reversibility is not enforced by Lewis's framework. Lewis's axiom (3) is similar to but not as strong as reversibility; that is, even though $Y = y$ may hold in all closest w-worlds and $W = w$ in all closest y-worlds, $Y = y$ still may not hold in the actual world. Nonetheless, we can safely conclude that, in adopting the causal interpretation of counterfactuals (together with the representational and algorithmic machinery of modifiable structural equation models), we are not introducing any restrictions on the set of counterfactual statements that are valid relative to recursive systems.

7.4.3 Imaging versus Conditioning

If action is a transformation from one probability function to another, one may ask whether every such transformation corresponds to an action, or if there are some constraints that are peculiar to those transformations that originate from actions. Lewis's (1976) formulation of counterfactuals indeed identifies such constraints: the transformation must be an *imaging* operator.

Whereas Bayes conditioning $P(s \mid e)$ transfers the entire probability mass from states excluded by e to the remaining states (in proportion to their current $P(s)$), imaging works differently; each excluded state s transfers its mass individually to a select set of states $S^*(s)$ that are considered "closest" to s. Indeed, we saw in (3.11) that the transformation defined by the action $do(X_i = x_i')$ can be interpreted in terms of such a mass-transfer process; each excluded state (i.e., one in which $X_i \neq x_i'$) transferred its mass to a select set of nonexcluded states that shared the same value of pa_i. This simple characterization of the set $S^*(s)$ of closest states is valid for Markovian models, but imaging generally permits the selection of any such set.

The reason why imaging is a more adequate representation of transformations associated with actions can be seen through a representation theorem due to Gardenfors (1988, thm. 5.2, p. 113; strangely, the connection to actions never appears in Gardenfors's analysis). Gardenfors's theorem states that a probability update operator $P(s) \rightarrow P_A(s)$ is an imaging operator if and only if it preserves mixtures; that is,

$$[\alpha P(s) + (1-\alpha)P'(s)]_A = \alpha P_A(s) + (1-\alpha)P_A'(s) \tag{7.43}$$

for all constants $1 > \alpha > 0$, all propositions A, and all probability functions P and P'. In other words, the update of any mixture is the mixture of the updates.[22]

This property, called *homomorphy,* is what permits us to specify actions in terms of transition probabilities, as is usually done in stochastic control and Markov decision processes. Denoting by $P_A(s \mid s')$ the probability resulting from acting A on a known state s', the homomorphism (7.43) dictates that

[22] Property (7.43) is reflected in the (U8) postulate of Katsuno and Mendelzon (1991): $(K_1 \vee K_2)o\mu = (K_1 o\mu) \vee (K_2 o\mu)$, where o is an update operator, similar to our $do(\cdot)$ operator.

$$P_A(s) = \sum_{s'} P_A(s \mid s')P(s'); \tag{7.44}$$

this means that, whenever s' is not known with certainty, $P_A(s)$ is given by a weighted sum of $P_A(s \mid s')$ over s', with the weight being the current probability function $P(s')$.

This characterization, however, is too permissive; although it requires any action-based transformation to be describable in terms of transition probabilities, it also accepts any transition probability specification, howsoever whimsical, as a descriptor of some action. The valuable information that actions are defined as *local* surgeries is ignored in this characterization. For example, the transition probability associated with the atomic action $A_i = do(X_i = x_i)$ originates from the deletion of just one mechanism in the assembly. Hence, the transition probabilities associated with the set of atomic actions would normally constrain one another. Such constraints emerge from the axioms of effectiveness, composition, and reversibility when probabilities are assigned to the states of U (Galles and Pearl 1997).

7.4.4 Relations to the Neyman–Rubin Framework

A Language in Search of a Model

The notation $Y_x(u)$ that we used for denoting counterfactual quantities is borrowed from the potential-outcome framework of Neyman (1923) and Rubin (1974), briefly introduced in Section 3.6.3, which was devised for statistical analysis of treatment effects.[23] In that framework, $Y_x(u)$ (often written $Y(x, u)$) stands for the outcome of experimental unit u (e.g., an individual or an agricultural lot) under a hypothetical experimental condition $X = x$. In contrast to the structural modeling, however, the variable $Y_x(u)$ in the potential-outcome framework is not a derived quantity but is taken as a primitive – that is, as an undefined symbol that represents the English phrase "the value that Y would assume in u, had X been x." Researchers pursuing the potential-outcome framework (e.g. Robins 1987; Manski 1995; Angrist et al. 1996) have used this interpretation as a guide for expressing subject-matter information and for devising plausible relationships between counterfactual and observed variables, including Robins's consistency rule $X = x \implies Y_x = Y$ (equation (7.20)). However, the potential-outcome framework does not provide a mathematical model from which such relationships could be derived or on the basis of which the question of completeness could be decided – that is, whether the relationships at hand are sufficient for managing all valid inferences.

The structural equation model formulated in Section 7.1 provides a formal semantics for the potential-outcome framework, since each such model assigns coherent truth values to the counterfactual quantities used in potential-outcome studies. From the structural perspective, the quantity $Y_x(u)$ is not a primitive but rather is derived mathematically from a set of equations F that is modified by the operator $do(X = x)$ (see Definition 7.1.4). Subject-matter information is expressed directly through the variables participating in those equations, without committing to their precise functional form. The variable

[23] A related (if not identical) framework that has been used in economics is the *switching regression.* For a brief review of such models, see Heckman (1996; see also Heckman and Honoré 1990 and Manski 1995). Winship and Morgan (1999) provided an excellent overview of the two schools.

U represents any set of background factors relevant to the analysis, not necessarily the identity of a specific individual in the population.

Using this semantics, in Section 7.3 we established an axiomatic characterization of the potential-response function $Y_x(u)$ and its relationships with the observed variables $X(u)$ and $Y(u)$. These basic axioms include or imply restrictions such as Robins's consistency rule (equation (7.20)), which were taken as given by potential-outcome researchers.

The completeness result further assures us that derivations involving counterfactual relationships in recursive models may safely be managed with two axioms only, effectiveness and composition. All truths implied by structural equation semantics are also derivable using these two axioms. Likewise – in constructing hypothetical contingency tables for recursive models (see Section 6.5.3) – we are guaranteed that, once a table satisfies effectiveness and composition, there exists at least one causal model that would generate that table. In essence, this establishes the formal equivalence of structural equation modeling, which is popular in economics and the social sciences (Goldberger 1991), and the potential-outcome framework as used in statistics (Rubin 1974; Holland 1986; Robins 1986).[24] In nonrecursive models, however, this is not the case. Attempts to evaluate counterfactual statements using only composition and effectiveness may fail to certify some valid conclusions (i.e., true in all causal models) whose validity can only be recognized through the use of reversibility.

Graphical versus Counterfactual Analysis

This formal equivalence between the structural and potential-outcome frameworks covers issues of semantics and expressiveness but does not imply equivalence in conceptualization or practical usefulness. Structural equations and their associated graphs are particularly useful as means of expressing assumptions about cause–effect relationships. Such assumptions rest on prior experiential knowledge, which – as suggested by ample evidence – is encoded in the human mind in terms of interconnected assemblies of autonomous mechanisms. These mechanisms are thus the building blocks from which judgments about counterfactuals are derived. Structural equations $\{f_i\}$ and their graphical abstraction $G(M)$ provide direct mappings for these mechanisms and therefore constitute a natural language for articulating or verifying causal knowledge or assumptions. The major weakness of the potential-outcome framework lies in the requirement that assumptions be articulated as conditional independence relationships involving counterfactual variables. For example, an assumption such as the one expressed in (7.30) is not easily comprehended even by skilled investigators, yet its structural image $U_1 \perp\!\!\!\perp U_2$ evokes an immediate process-based interpretation.[25]

[24] This equivalence was anticipated in Holland (1988), Pratt and Schlaifer (1988), Pearl (1995a), and Robins (1995). Note, though, that counterfactual claims and the equation deletion part of our model (Definition 7.1.3) are not made explicit in the standard literature on structural equation modeling.

[25] These views are diametrically opposite to those expressed by Angrist et al. (1996), who stated: "Typically the researcher does not have a firm idea what these disturbances really represent, and therefore it is difficult to draw realistic conclusions or communicate results based on their properties." I have found that researchers who are knowledgeable in their respective subjects have a very clear idea what these disturbances really represent, and those who don't would certainly not be able to make realistic judgments about counterfactual dependencies.

A happy symbiosis between graphs and counterfactual notation was demonstrated in Section 7.3.2. In that example, assumptions were expressed in graphical form, then translated into counterfactual notation (using the rules of (7.25) and (7.26)), and finally submitted to algebraic derivation. Such symbiosis offers a more effective method of analysis than methods that insist on expressing assumptions directly as counterfactuals. Additional examples will be demonstrated in Chapter 9, where we analyze probability of causation. Note that, in the derivation of Section 7.3.2, the graph continued to assist the procedure by displaying independence relationships that are not easily derived by algebraic means alone. For example, it is hardly straightforward to show that the assumptions of (7.27)–(7.30) imply the conditional independence $(Y_z \perp\!\!\!\perp Z_x \mid \{Z, X\})$ but do not imply the conditional independence $(Y_z \perp\!\!\!\perp Z_x \mid Z)$. Such implications can, however, easily be tested in the graph of Figure 7.5 or in the twin network construction of Section 7.1.3 (see Figure 7.3).

The most compelling reason for molding causal assumptions in the language of graphs is that such assumptions are needed before the data are gathered, at a stage when the model's parameters are still "free" (i.e., still to be determined from the data). The usual temptation is to mold those assumptions in the language of statistical independence, which carries an aura of testability and hence of scientific legitimacy. (Chapter 6 exemplifies the difficulties associated with such temptations.) However, conditions of statistical independence – regardless of whether they relate to V variables, U variables, or counterfactuals – are generally sensitive to the values of the model's parameters, which are not available at the model construction phase. The substantive knowledge available at the modeling phase cannot support such assumptions unless they are *stable,* that is, insensitive to the values of the parameters involved. The implications of graphical models, which rest solely on the interconnections among mechanisms, satisfy this stability requirement and can therefore be ascertained from generic substantive knowledge *before* data are collected. For example, the assertion $(X \perp\!\!\!\perp Y \mid Z, U_1)$, which is implied by the graph of Figure 7.5, remains valid for any substitution of functions in $\{f_i\}$ and for any assignment of prior probabilities to U_1 and U_2.

These considerations apply not only to the formulation of causal assumptions but also to the language in which causal concepts are defined and communicated. Many concepts in the social and medical sciences are defined in terms of relationships among unobserved U variables, also known as "errors" or "disturbance terms." We have seen in Chapter 5 (Section 5.4.3) that key econometric notions such as exogeneity and instrumental variables have traditionally been defined in terms of absence of correlation between certain observed variables and certain error terms. Naturally, such definitions attract criticism from strict empiricists, who regard unobservables as metaphysical or definitional (Richard 1980; Engle et al. 1983; Holland 1988), and also (more recently) from potential-outcome analysts, who regard the use of structural models as an unwarranted commitment to a particular functional form (Angrist et al. 1996). This new criticism will be considered in the following section.

7.4.5 Exogeneity Revisited: Counterfactual and Graphical Definitions

The analysis of this chapter provides a counterfactual interpretation of the error terms in structural equation models, supplementing the operational definition of (5.25). We have

seen that the meaning of the error term u_Y in the equation $Y = f_Y(pa_Y, u_Y)$ is captured by the counterfactual variable Y_{pa_Y}. In other words, the variable U_Y can be interpreted as a modifier of the functional mapping from PA_Y to Y. The statistics of such modifications is observable when pa_Y is held fixed. This translation into counterfactual notation may facilitate algebraic manipulations of U_Y without committing to the functional form of f_Y. However, from the viewpoint of model specification, the error terms should be still viewed as (summaries of) omitted factors.

Armed with this interpretation, we can obtain graphical and counterfactual definitions of causal concepts that were originally given error-based definitions. Examples of such concepts are causal influence, exogeneity, and instrumental variables (Section 5.4.3). In clarifying the relationships among error-based, counterfactual, and graphical definitions of these concepts, we should first note that these three modes of description can be organized in a simple hierarchy. Since graph separation implies independence but independence does not imply graph separation (Theorem 1.2.4), definitions based on graph separation should imply those based on error-term independence. Likewise, since for any two variables X and Y the independence relation $U_X \perp\!\!\!\perp U_Y$ implies the counterfactual independence $X_{pa_X} \perp\!\!\!\perp Y_{pa_Y}$ (but not the other way around), it follows that definitions based on error independence should imply those based on counterfactual independence. Overall, we have the following hierarchy:

$$\text{graphical criteria} \Longrightarrow \text{error-based criteria} \Longrightarrow \text{counterfactual criteria.}$$

The concept of exogeneity may serve to illustrate this hierarchy. The pragmatic definition of exogeneity is best formulated in counterfactual or interventional terms as follows.

Exogeneity (Counterfactual Criterion)
A variable X is exogenous relative to Y if and only if the effect of X on Y is identical to the conditional probability of Y given X – that is, if

$$P(Y_x = y) = P(y \mid x) \tag{7.45}$$

or, equivalently,

$$P(Y = y \mid do(x)) = P(y \mid x); \tag{7.46}$$

this in turn is equivalent to the independence condition $Y_x \perp\!\!\!\perp X$, named "weak ignorability" in Rosenbaum and Rubin (1983).[26]

This definition is pragmatic in that it highlights the reasons economists should be concerned with exogeneity by explicating the policy-analytic benefits of discovering that a variable is exogenous. However, this definition fails to guide an investigator toward

[26] We focus the discussion in this section on the causal component of exogeneity, which the econometric literature has unfortunately renamed "superexogeneity" (see Section 5.4.3). We also postpone discussion of "strong ignorability," defined as the joint independence $\{Y_x, Y_{x'}\} \perp\!\!\!\perp X$, to Chapter 9 (Definition 9.2.3).

verifying, from substantive knowledge of the domain, whether this independence condition holds in any given system, especially when many equations are involved. To facilitate such judgments, economists (e.g. Koopmans 1950; Orcutt 1952) have adopted the error-based criterion of Definition 5.4.6.

Exogeneity (Error-Based Criterion)
A variable X is exogenous in M relative to Y if X is independent of all error terms that have an influence on Y that is not mediated by X.[27]

This definition is more transparent to human judgment because the reference to error terms tends to focus attention on specific factors, potentially affecting Y, with which scientists are familiar. Still, to judge whether such factors are statistically independent is a difficult mental task unless the independencies considered are dictated by topological considerations that assure their stability. Indeed, the most popular conception of exogeneity is encapsulated in the notion of "common cause"; this may be stated formally as follows.

Exogeneity (Graphical Criterion)
A variable X is exogenous relative to Y if X and Y have no common ancestor in $G(M)$ or, equivalently, if all back-door paths between X and Y are blocked (by colliding arrows).[28]

It is easy to show that the graphical condition implies the error-based condition, which in turn implies the counterfactual (or pragmatic) condition of (7.46). The converse implications do not hold. For example, Figure 6.4 illustrates a case where the graphical criterion fails and both the error-based and counterfactual criteria classify X as exogenous. We argued in Section 6.4 that this type of exogeneity (there called "no confounding") is unstable or incidental, and we have raised the question of whether such cases were meant to be embraced by the definition. If we exclude unstable cases from consideration, then our three-level hierarchy collapses and all three definitions coincide.

Instrumental Variables: Three Definitions

A three-level hierarchy similarly characterizes the notion of instrumental variables (Bowden and Turkington 1984; Pearl 1995c; Angrist et al. 1996), illustrated in Figure 5.9. The traditional definition qualifies a variable Z as *instrumental* (relative to the pair (X, Y)) if (i) Z is independent of all error terms that have an influence on Y that is not mediated by X and (ii) Z is not independent of X.

[27] Independence relative to *all* errors is sometimes required in the literature (e.g. Dhrymes 1970, p. 169), but this is obviously too strong.

[28] As in Chapter 6 (note 19), the expression "common ancestors" should exclude nodes that have no other connection to Y except through X and should include latent nodes for every pair of dependent errors. Generalization to conditional exogeneity relative to observed covariates is straightforward in all three definitions.

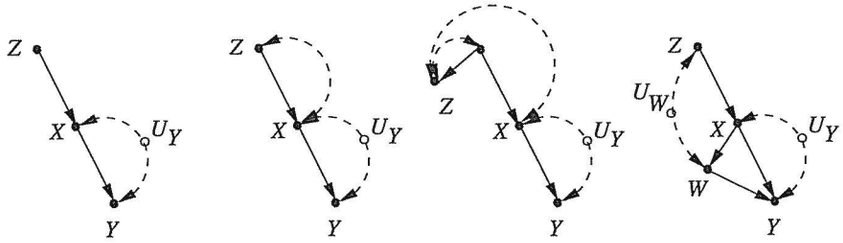

Figure 7.8 Z is a proper instrumental variable in the (linear) models of (a), (b), and (c), since it satisfies $Z \perp\!\!\!\perp U_Y$. It is not an instrument in (d) because it is correlated with U_W, which influences Y.

The counterfactual definition[29] replaces condition (i) with (i′): Z is independent of Y_x. The graphical definition replaces condition (i) with (i″): every unblocked path connecting Z and Y must contain an arrow pointing into X (alternatively, $(Z \perp\!\!\!\perp Y)_{G_{\overline{X}}}$). Figure 7.8 illustrates this definition through examples.

When a set S of covariates is measured, these definitions generalize as follows.

Definition 7.4.1 (Instrument)
A variable Z is an instrument *relative to the total effect of X on Y if there exists a set of measurements $S = s$, unaffected by X, such that either of the following criteria holds.*

1. *Counterfactual criterion*:
   (i) $Z \perp\!\!\!\perp Y_x \mid S = s$;
   (ii) $Z \not\perp\!\!\!\perp X \mid S = s$.
2. *Graphical criterion*:
   (i) $(Z \perp\!\!\!\perp Y \mid S)_{G_{\overline{X}}}$;
   (ii) $(Z \not\perp\!\!\!\perp X \mid S)_G$.

In concluding this section, I should reemphasize that it is because graphical definitions are insensitive to the values of the model's parameters that graphical vocabulary guides and expresses so well our intuition about causal effects, exogeneity, instruments, confounding, and even (I speculate) more technical notions such as randomness and statistical independence.

[29] There is, in fact, no agreed-upon generalization of instrumental variables to nonlinear systems. The definition here, taken from Galles and Pearl (1998), follows by translating the error-based definition into counterfactual vocabulary. Angrist et al. (1996), who expressly rejected all reference to graphs or error terms, assumed two unnecessary restrictions: that Z be ignorable (i.e. randomized; this is violated in Figures 7.8(b) and (c)) and that Z affect X (violated in Figure 7.8(c)). Similar assumptions were made by Heckman and Vytlacil (1999), who used both counterfactuals and structural equation models.

7.5 STRUCTURAL VERSUS PROBABILISTIC CAUSALITY

Probabilistic causality is a branch of philosophy that attempts to explicate causal relationships in terms of probabilistic relationships. This attempt is motivated by several ideas and expectations. First and foremost, probabilistic causality promises a solution to the centuries-old puzzle of causal discovery – that is, how humans discover genuine causal relationships from bare empirical observations, free of any causal preconceptions. Given the Humean dictum that all knowledge originates with human experience and the (less compelling but fashionable) assumption that human experience is encoded in the form of a probability function, it is natural to expect that causal knowledge be reducible to a set of relationships in some probability distribution that is defined over the variables of interest. Second, in contrast to deterministic accounts of causation, probabilistic causality offers substantial cognitive economy. Physical states and physical laws need not be specified in minute detail because instead they can be summarized in the form of probabilistic relationships among macro states so as to match the granularity of natural discourse. Third, probabilistic causality is equipped to deal with the modern (i.e. quantum-theoretical) conception of uncertainty, according to which determinism is merely an epistemic fiction and nondeterminism is the fundamental feature of physical reality.

The formal program of probabilistic causality owes its inception to Reichenbach (1956) and Good (1961), and it has subsequently been pursued by Suppes (1970), Skyrms (1980), Spohn (1980), Otte (1981), Salmon (1984), Cartwright (1989), and Eells (1991). The current state of this program is rather disappointing, considering its original aspirations. Salmon has abandoned the effort altogether, concluding that "causal relations are not appropriately analyzable in terms of statistical relevance relations" (1984, p. 185); instead, he has proposed an analysis in which "causal processes" are the basic building blocks. More recent accounts by Cartwright and Eells have resolved some of the difficulties encountered by Salmon, but at the price of either complicating the theory beyond recognition or compromising its original goals. The following is a brief account of the major achievements, difficulties, and compromises of probabilistic causality as portrayed in Cartwright (1989) and Eells (1991).

7.5.1 The Reliance on Temporal Ordering

Standard probabilistic accounts of causality assume that, in addition to a probability function P, we are also given the temporal order of the variables in the analysis. This is understandable, considering that causality is an asymmetric relation whereas statistical relevance is symmetric. Lacking temporal information, it would be impossible to decide which of two dependent variables is the cause and which the effect, since every joint distribution $P(x, y)$ induced by a model in which X is a cause of Y can also be induced by a model in which Y is the cause of X. Thus, any method of inferring that X is a cause of Y must also infer, by symmetry, that Y is a cause of X. In Chapter 2 we demonstrated that, indeed, at least three variables are needed for determining the directionality of arrows in a DAG and, more serious yet, no arrow can be oriented from probability information