(Pearl and Verma 1991), and *faithfulness* (Spirtes et al. 1993), which assist in the elucidation of causal diagrams from sparse statistical associations (see Chapter 2). The same conception has evidently been shared by authors who aspired to connect associational criteria with confounding.

The advent of structural model analysis, assisted by graphical methods, offers a mathematical framework in which considerations of confounding can be formulated and managed more effectively. Using this framework, this chapter explicates the criterion of stable unbiasedness and shows that this criterion (i) has implicitly been the target of many investigations in epidemiology and biostatistics, and (ii) can be given operational statistical tests similar to those invoked in testing collapsibility. We further show (Section 6.5.3) that the structural framework overcomes basic cognitive and methodological barriers that have made confounding one of the most confused topics in the literature. It is therefore natural to predict that this framework will become the primary mathematical basis for future studies of confounding.

Acknowledgment

Sections 6.2–6.3 began as a commentary on Sander Greenland's 1997 manuscript entitled "Causation, Confounding, and Collapsibility." Greenland's paper was motivated by considerations similar to those exposed in this chapter, and it was based on a counterfactual–exchangeability approach that he and James Robins introduced to epidemiology in the mid-1980s. I have since joined Sander and Jamie as co-author on "Confounding and Collapsibility in Causal Inference" (Greenland et al. 1999b). However, space limitations and other constraints did not permit the ideas presented in this chapter to be fully expressed in our joint paper.

Technical discussions with James Robin and Sander Greenland were extremely valuable. Sander, in particular, gave many constructive comments on two early drafts and helped to keep them comprehensible to epidemiologists. Jan Koster called my attention to the connection between Stone's and Robins's criteria of no-confounding and caught several oversights in an earlier draft. Other helpful discussants were Michelle Pearl, Bill Shipley, Rolf Steyer, Stephen Stigler, and David Trichler.

CHAPTER SEVEN

The Logic of Structure-Based Counterfactuals

And the Lord said,
"If I find in the city of Sodom fifty good men,
I will pardon the whole place for their sake."
Genesis 18:26

Preface

This chapter provides a formal analysis of structure-based *counterfactuals,* a concept introduced briefly in Chapter 1 that will occupy the rest of our discussion in this book. Through this analysis, we will obtain sharper mathematical definitions of other concepts that were introduced in earlier chapters, including causal models, action, causal effects, causal relevance, error terms, and exogeneity.

After casting the concepts of causal model and counterfactuals in abstract mathematical terms, we will demonstrate by examples how counterfactual questions can be answered from both deterministic and probabilistic causal models (Section 7.1). In Section 7.2.1, we will argue that policy analysis is an exercise in counterfactual reasoning and demonstrate this thesis in a simple example taken from econometrics. This will set the stage for our discussion in Section 7.2.2, where we explicate the empirical content of counterfactuals in terms of policy predictions. Section 7.2.3 discusses the role of counterfactuals in the interpretation and generation of causal explanations. Section 7.2 concludes with discussions of how causal relationships emerge from actions and mechanisms (Section 7.2.4) and how causal directionality can be induced from a set of symmetric equations (Section 7.2.5).

In Section 7.3 we develop an axiomatic characterization of counterfactual and causal relevance relationships as they emerge from the structural model semantics. Section 7.3.1 will identify a set of properties, or axioms, that allow us to derive new counterfactual relations from assumptions, and Section 7.3.2 demonstrates the use of these axioms in algebraic derivation of causal effects. Section 7.3.3 introduces axioms for the relationship of causal relevance and, using their similarity to the axioms of graphs, describes the use of graphs for verifying relevance relationships.

The axiomatic characterization developed in Section 7.3 enables us to compare structural models with other approaches to causality and counterfactuals, most notably those based on Lewis's closest-world semantics (Sections 7.4.1–7.4.4). The formal equivalence of the structural approach and the Neyman–Rubin potential-outcome framework is discussed in Section 7.4.4. Finally, we revisit the topic of exogeneity and extend our discussion of Section 5.4.3 with counterfactual definitions of exogenous and instrumental variables in Section 7.4.5.

The final part of this chapter (Section 7.5) compares the structural account of causality with that based on probabilistic relationships. We elaborate our preference toward the structural account and highlight the difficulties that the probabilistic account is currently facing.

7.1 STRUCTURAL MODEL SEMANTICS

How do scientists predict the outcome of one experiment from the results of other experiments run under totally different conditions? Such predictions require us to envision what the world would be like under various hypothetical changes and so invoke *counterfactual* inference. Though basic to scientific thought, counterfactual inference cannot easily be formalized in the standard languages of logic, algebraic equations, or probability. The formalization of counterfactual inference requires a language within which the invariant relationships in the world are distinguished from transitory relationships that represent one's beliefs about the world, and such distinction is not supported by standard algebras, including the algebra of equations, Boolean algebra, and probability calculus. Structural models offer such distinction, and this section presents a structural model semantics of counterfactuals as defined in Balke and Pearl (1995), Galles and Pearl (1997, 1998), and Halpern (1998).[1] Related approaches have been proposed in Simon and Rescher (1966), Robins (1986), and Ortiz (1999).

We start with a deterministic definition of a causal model, which consists (as we have discussed in earlier chapters) of functional relationships among variables of interest, each relationship representing an autonomous mechanism. Causal and counterfactual relationships are defined in this model in terms of response to local modifications of those mechanisms. Probabilistic relationships emerge naturally by assigning probabilities to background conditions. After demonstrating, by examples, how this model facilitates the computation of counterfactuals in both deterministic and probabilistic contexts (Section 7.1.2), we then present a general method of computing probabilities of counterfactual expressions using causal diagrams (Section 7.1.3).

7.1.1 Definitions: Causal Models, Actions, and Counterfactuals

A "model," in the common use of the word, is an idealized representation of reality that highlights some aspects and ignores others. In logical systems, however, a model is a mathematical object that assigns truth values to sentences in a given language, where each sentence represents some aspect of reality. Truth tables, for example, are models in propositional logic; they assign a truth value to any Boolean expression, which may represent an event or a set of conditions in the domain of interest. A joint probability function, as another example, is a model in probability logic; it assigns a truth value to any sentence of the form $P(A \mid B) < p$, where A and B are Boolean expressions representing events. A *causal model,* naturally, should encode the truth values of sentences

[1] Similar models, called "neuron diagrams" (Lewis 1986, p. 200; Hall 1998) are used informally by philosophers to illustrate chains of causal processes.

that deal with causal relationships; these include action sentences (e.g., "A will be true if we do B"), counterfactuals (e.g., "A would have been different were it not for B"), and plain causal utterances (e.g., "A may cause B" or "B occurred because of A"). Such sentences cannot be interpreted in standard propositional logic or probability calculus because they deal with changes that occur in the external world rather than with changes in our beliefs about a static world. Causal models encode and distinguish information about external changes through an explicit representation of the mechanisms that are altered in such changes.

Definition 7.1.1 (Causal Model)
A causal model *is a triple*

$$M = \langle U, V, F \rangle,$$

where:

(i) U *is a set of* background *variables, (also called* exogenous[2]*), that are determined by factors outside the model;*

(ii) V *is a set* $\{V_1, V_2, \ldots, V_n\}$ *of variables, called* endogenous, *that are determined by variables in the model – that is, variables in* $U \cup V$; *and*

(iii) F *is a set of functions* $\{f_1, f_2, \ldots, f_n\}$ *such that each* f_i *is a mapping from (the respective domains of)* $U \cup (V \setminus V_i)$ *to* V_i *and such that the entire set* F *forms a mapping from* U *to* V. *In other words, each* f_i *tells us the value of* V_i *given the values of all other variables in* $U \cup V$, *and the entire set* F *has a unique solution* $V(u)$.[3] *Symbolically, the set of equations* F *can be represented by writing*

$$v_i = f_i(pa_i, u_i), \quad i = 1, \ldots, n,$$

where pa_i *is any realization of the unique minimal set of variables* PA_i *in* $V \setminus V_i$ *(connoting* parents*) sufficient for representing* f_i. *Likewise,* $U_i \subseteq U$ *stands for the unique minimal set of variables in* U *sufficient for representing* f_i.[4]

Every causal model M can be associated with a directed graph, $G(M)$, in which each node corresponds to a variable and the directed edges point from members of PA_i and U_i toward V_i. We call such a graph the *causal diagram* associated with M. This graph merely identifies the endogenous and background variables that have direct influence on each V_i; it does not specify the functional form of f_i. The convention of confining the parent set PA_i to variables in V stems from the fact that the background variables are often unobservable. In general, however, we can extend the parent sets to include observed variables in U.

[2] We will try to refrain from using the term "exogenous" in referring to background conditions, because this term has acquired more refined technical connotations (see Sections 5.4.3 and 7.4). The term "predetermined" is used in the econometric literature.

[3] Uniqueness is ensured in recursive (i.e. acyclic) systems. Halpern (1998) allows multiple solutions in nonrecursive systems.

[4] A set of variables X is *sufficient* for representing a function $y = f(x, z)$ if f is trivial in Z – that is, if for every x, z, z' we have $f(x, z) = f(x, z')$.

Definition 7.1.2 (Submodel)
Let M be a causal model, X a set of variables in V, and x a particular realization of X. A submodel M_x *of M is the causal model*

$$M_x = \langle U, V, F_x \rangle,$$

where

$$F_x = \{f_i : V_i \notin X\} \cup \{X = x\}. \tag{7.1}$$

In words, F_x is formed by deleting from F all functions f_i corresponding to members of set X and replacing them with the set of constant functions $X = x$.

Submodels are useful for representing the effect of local actions and hypothetical changes, including those implied by counterfactual antecedents. If we interpret each function f_i in F as an independent physical mechanism and define the action $do(X = x)$ as the minimal change in M required to make $X = x$ hold true under any u, then M_x represents the model that results from such a minimal change, since it differs from M by only those mechanisms that directly determine the variables in X. The transformation from M to M_x modifies the algebraic content of F, which is the reason for the name "modifiable structural equations" used in Galles and Pearl (1998).[5]

Definition 7.1.3 (Effect of Action)
Let M be a causal model, X a set of variables in V, and x a particular realization of X. The effect of action $do(X = x)$ *on M is given by the submodel* M_x.

Definition 7.1.4 (Potential Response)
Let X and Y be two subsets of variables in V. The potential response *of Y to action* $do(X = x)$, *denoted* $Y_x(u)$, *is the solution for Y of the set of equations* F_x.[6]

We will confine our attention to actions in the form of $do(X = x)$. Conditional actions of the form "$do(X = x)$ if $Z = z$" can be formalized using the replacement of equations by functions of Z, rather than by constants (Section 4.2). We will not consider disjunctive actions of the form "$do(X = x$ or $Z = z)$," since these complicate the probabilistic treatment of counterfactuals.

Definition 7.1.5 (Counterfactual)
Let X and Y be two subsets of variables in V. The counterfactual sentence "The value that Y would have obtained, had X been x" is interpreted as denoting the potential response $Y_x(u)$.

[5] Structural modifications date back to Marschak (1950) and Simon (1953). An explicit translation of interventions into "wiping out" equations from the model was first proposed by Strotz and Wold (1960) and later used in Fisher (1970), Sobel (1990), Spirtes et al. (1993), and Pearl (1995a). A similar notion of submodel was introduced by Fine (1985), though not specifically for representing actions and counterfactuals.

[6] If Y is a set of variables $Y = (Y_1, Y_2, \ldots)$, then $Y_x(u)$ stands for a vector of functions $(Y_{1_x}(u), Y_{2_x}(u), \ldots)$.

Definition 7.1.5 thus interprets the counterfactual phrase "had X been x" in terms of a hypothetical modification of the equations in the model; it simulates an external action (or spontaneous change) that modifies the actual course of history and enforces the condition "$X = x$" with minimal change of mechanisms. This is a crucial step in the semantics of counterfactuals (Balke and Pearl 1994b), as it permits x to differ from the current value of $X(u)$ without creating logical contradiction; it also suppresses abductive inferences (or backtracking) from the counterfactual antecedent $X = x$.[7] In Chapter 3 (Section 3.6.3) we used the notation $Y(x, u)$ to denote the subjunctive conditional "the value that Y would obtain in unit u, had X been x" (as used in the Neyman–Rubin potential-outcome model). Throughout the rest of this book we will use the notation $Y_x(u)$ to denote counterfactuals tied specifically to the structural model interpretation of Definition 7.1.5 (paralleling (3.51)); $Y(x, u)$ will be reserved for generic subjunctive conditionals, uncommitted to any specific semantics.

Definition 7.1.5 endows the atomic mechanisms $\{f_i\}$ themselves with interventional–counterfactual interpretation, because $v_i = f_i(pa_i, u_i)$ is the value of V_i in the submodel $M_{v \setminus v_i}$. In other words, $f_i(pa_i, u_i)$ stands for the potential response of V_i when we hold constant *all* other variables in V.

This formulation generalizes naturally to probabilistic systems as follows.

Definition 7.1.6 (Probabilistic Causal Model)
A probabilistic causal model *is a pair*

$$\langle M, P(u) \rangle,$$

where M is a causal model and $P(u)$ *is a probability function defined over the domain of U.*

The function $P(u)$, together with the fact that each endogenous variable is a function of U, defines a probability distribution over the endogenous variables. That is, for every set of variables $Y \subseteq V$, we have

$$P(y) \triangleq P(Y = y) = \sum_{\{u \mid Y(u) = y\}} P(u). \tag{7.2}$$

The probability of counterfactual statements is defined in the same manner, through the function $Y_x(u)$ induced by the submodel M_x:

$$P(Y_x = y) = \sum_{\{u \mid Y_x(u) = y\}} P(u). \tag{7.3}$$

Likewise, a causal model defines a joint distribution on counterfactual statements. That is, $P(Y_x = y, Z_w = z)$ is defined for any (not necessarily disjoint) sets of variables Y, X, Z, and W. In particular, $P(Y_x = y, X = x')$ and $P(Y_x = y, Y_{x'} = y')$ are well-defined for $x \neq x'$ and are given by

[7] Simon and Rescher (1966, p. 339) did not include this step in their account of counterfactuals and noted that backward inferences triggered by the antecedents can lead to ambiguous interpretations.

$$P(Y_x = y,\ X = x') = \sum_{\{u \mid Y_x(u)=y\ \&\ X(u)=x'\}} P(u) \tag{7.4}$$

and

$$P(Y_x = y,\ Y_{x'} = y') = \sum_{\{u \mid Y_x(u)=y\ \&\ Y_{x'}(u)=y'\}} P(u). \tag{7.5}$$

If x and x' are incompatible then Y_x and $Y_{x'}$ cannot be measured simultaneously, and it may seem meaningless to attribute probability to the joint statement "Y would be y if $X = x$ *and* Y would be y' if $X = x'$." Such concerns have been a source of recent objections to treating counterfactuals as jointly distributed random variables (Dawid 1997). The definition of Y_x and $Y_{x'}$ in terms of two distinct submodels, driven by a standard probability space over U, explains away these objections (see Section 7.2.2) and further illustrates that joint probabilities of counterfactuals can be encoded rather parsimoniously using $P(u)$ and F.

Of particular interest to us would be probabilities of counterfactuals that are conditional on actual observations. For example, the probability that event $X = x$ "was the cause" of event $Y = y$ may be interpreted as the probability that Y would not be equal to y had X not been x, given that $X = x$ and $Y = y$ have in fact occurred (see Chapter 9 for an in-depth discussion of the probabilities of causation). Such probabilities are well-defined in the model just described; they require the evaluation of expressions of the form $P(Y_{x'} = y' \mid X = x,\ Y = y)$ with x' and y' incompatible with x and y, respectively. Equation (7.4) allows the evaluation of this quantity as follows:

$$\begin{aligned} P(Y_{x'} = y' \mid X = x,\ Y = y) &= \frac{P(Y_{x'} = y',\ X = x,\ Y = y)}{P(X = x,\ Y = y)} \\ &= \sum_u P(Y_{x'}(u) = y') P(u \mid x, y). \end{aligned} \tag{7.6}$$

In other words, we first update $P(u)$ to obtain $P(u \mid x, y)$ and then use the updated distribution $P(u \mid x, y)$ to compute the expectation of the index function $Y_{x'}(u) = y'$.

This substantiates the three-step procedure introduced in Section 1.4, which we now summarize in a theorem.

Theorem 7.1.7
Given model $\langle M, P(u)\rangle$, the conditional probability $P(B_A \mid e)$ of a counterfactual sentence "If it were A then B," given evidence e, can be evaluated using the following three steps.

1. ***Abduction** – Update $P(u)$ by the evidence e to obtain $P(u \mid e)$.*
2. ***Action** – Modify M by the action $do(A)$, where A is the antecedent of the counterfactual, to obtain the submodel M_A.*
3. ***Prediction** – Use the modified model $\langle M_A, P(u \mid e)\rangle$ to compute the probability of B, the consequence of the counterfactual.*

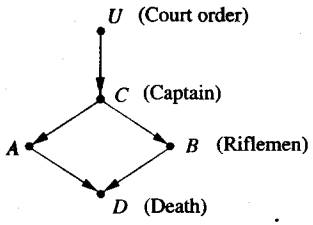

Figure 7.1 Causal relationships in the example of the two-man firing squad.

To complete this section, we introduce two additional objects that will prove useful in subsequent discussions: *worlds*[8] and *theories*.

Definition 7.1.8 (Worlds and Theories)
A causal world w *is a pair $\langle M, u\rangle$, where M is a causal model and u is a particular realization of the background variables U. A causal* theory *is a set of causal worlds.*

A world w can be viewed as a degenerate probabilistic model for which $P(u) = 1$. Causal theories will be used to characterize partial specifications of causal models, for example, models sharing the same causal diagram or models in which the functions f_i are linear with undetermined coefficients.

7.1.2 Evaluating Counterfactuals: Deterministic Analysis

In Section 1.4.1 we presented several examples demonstrating the interpretation of actions and counterfactuals in structural models. We now apply the definitions of Section 7.1.1 to demonstrate how counterfactual queries, both deterministic and probabilistic, can be answered formally using structural model semantics.

Example 1: The Firing Squad

Consider a two-man firing squad as depicted in Figure 7.1, where A, B, C, D, and U stand for the following propositions:

U = court orders the execution;
C = captain gives a signal;
A = rifleman A shoots;
B = rifleman B shoots;
D = prisoner dies.

Assume that the court's decision is unknown, that both riflemen are accurate, alert, and law-abiding, and that the prisoner is not likely to die from fright or other extraneous causes. We wish to construct a formal representation of the story, so that the following sentences can be evaluated mechanically.

[8] Adnan Darwiche called my attention to the importance of this object.

S1 *Prediction* – If rifleman A did not shoot then the prisoner is alive:

$$\neg A \implies \neg D.$$

S2 *Abduction* – If the prisoner is alive, then the captain did not signal:

$$\neg D \implies \neg C.$$

S3 *Transduction* – If rifleman A shot, then B shot as well:

$$A \implies B.$$

S4 *Action* – If the captain gave no signal and rifleman A decides to shoot, then the prisoner will die and B will not shoot.

$$\neg C \implies D_A \ \& \ \neg B_A.$$

S5 *Counterfactual* – If the prisoner is dead, then the prisoner would be dead even if rifleman A had not shot:

$$D \implies D_{\neg A}.$$

Evaluating Standard Sentences

To prove the first three sentences we need not invoke causal models; these sentences involve standard logical connectives and thus can be handled using standard logical deduction. The story can be captured in any convenient logical theory (a set of propositional sentences), for example,

$$T_1\text{: } U \iff C,\ C \iff A,\ C \iff B,\ A \vee B \iff D$$

or

$$T_2\text{: } U \iff C \iff A \iff B \iff D,$$

where each theory admits the two logical models

$$m_1\text{: } \{U, C, A, B, D\} \quad \text{and} \quad m_2\text{: } \{\neg U, \neg C, \neg A, \neg B, \neg D\}.$$

In words, any theory T that represents our story should imply that either all five propositions are true or all are false; models m_1 and m_2 present these two possibilities explicitly. The validity of S1–S3 can easily be verified, either by derivation from T or by noting that the antecedent and consequent in each sentence are both part of the same model.

Two remarks are worth making before we go on to analyze sentences S4 and S5. First, the two-way implications in T_1 and T_2 are necessary for supporting abduction; if we were to use one-way implications (e.g. $C \implies A$) then we would not be able to conclude C from A. In standard logic, this symmetry removes all distinctions between the tasks of prediction (reasoning forward in time), abduction (reasoning from evidence to explanation), and transduction (reasoning from evidence to explanation and then from explanation to predictions). Using two-way implication, these three modes of reasoning differ only in the interpretations they attach to antecedents and consequents of conditional sentences – not in their methods of inference. In nonstandard logics (e.g., logic programming), where the implication sign dictates the direction of inference and even contraposition is not licensed, metalogical inference machinery must be invoked to perform abduction (Eshghi and Kowalski 1989).

Second, the feature that renders S1-S3 manageable in standard logic is that they all deal with *epistemic* inference – that is, inference from beliefs to beliefs about a static world. Sentence S2, for example, can be explicated to state: If we find that the prisoner is alive then we have the license to believe that the captain did not give the signal. The material implication sign ($\implies$) in logic does not extend beyond this narrow meaning, to be contrasted next with the counterfactual implication.

Evaluating Action Sentences

Sentence S4 invokes a deliberate action, "rifleman A decides to shoot." From our discussion of actions (see e.g. Chapter 4 or Definition 7.1.3), any such action must violate some premises, or mechanisms, in the initial theory of the story. To formally identify what remains invariant under the action, we must incorporate causal relationships into the theory; logical relationships alone are not sufficient. The causal model corresponding to our story is as follows.

Model M

	(U)
$C = U$	(C)
$A = C$	(A)
$B = C$	(B)
$D = A \vee B$	(D)

Here we use equality rather than implication in order to (i) permit two-way inference and (ii) stress that, unlike logical sentences, each equation represents an autonomous mechanism (an "integrity constraint" in the language of databases) – it remains invariant unless specifically violated. We further use parenthetical symbols next to each equation in order to identify explicitly the dependent variable (on the l.h.s.) in the equation, thus representing the causal asymmetry associated with the arrows in Figure 7.1.

To evaluate S4, we follow Definition 7.1.3 and form the submodel M_A, in which the equation $A = C$ is replaced by A (simulating the decision of of rifleman A to shoot regardless of signals).

Model M_A

	(U)
$C = U$	(C)
A	(A)
$B = C$	(B)
$D = A \vee B$	(D)
Facts: $\neg C$	
Conclusions: $A, D, \neg B, \neg U, \neg C$	

We see that, given $\neg C$, we can easily deduce D and $\neg B$ and thus confirm the validity of S4.

It is important to note that "problematic" sentences like S4, whose antecedent violates one of the basic premises in the story (i.e., that both riflemen are law-abiding) are handled naturally in the same deterministic setting in which the story is told. Traditional

logicians and probabilists tend to reject sentences like S4 as contradictory and insist on reformulating the problem probabilistically so as to tolerate exceptions to the law $A = C$.[9] Such reformulations are unnecessary; the structural approach permits us to process commonplace causal statements in their natural deterministic habitat without first immersing them in nondeterministic decor. In this framework, all laws are understood to represent "defeasible" default expressions – subject to breakdown by deliberate intervention. The basic laws of physics remain immutable, of course, but their applicability to any given scenario is subject to modification by agents' actions or external intervention.

Evaluating Counterfactuals

We are now ready to evaluate the counterfactual sentence S5. Following Definition 7.1.5, the counterfactual $D_{\neg A}$ stands for the value of D in submodel $M_{\neg A}$. This value is ambiguous because it depends on the value of U, which is not specified in $M_{\neg A}$. The observation D removes this ambiguity; upon finding the prisoner dead we can infer that the court has given the order (U) and, consequently, if rifleman A had refrained from shooting then rifleman B would have shot and killed the prisoner, thus confirming $D_{\neg A}$.

Formally, we can derive $D_{\neg A}$ by using the steps of Theorem 7.1.7 (though no probabilities are involved). We first add the fact D to the original model M and evaluate U; then we form the submodel $M_{\neg A}$ and reevaluate the truth of D in $M_{\neg A}$, using the value of U found in the first step. These steps are explicated as follows.

Step 1

Model M

	(U)
$C = U$	(C)
$A = C$	(A)
$B = C$	(B)
$D = A \vee B$	(D)

Facts: D

Conclusions: U, A, B, C, D

Step 2

Model $M_{\neg A}$

	(U)
$C = U$	(C)
$\neg A$	(A)
$B = C$	(B)
$D = A \vee B$	(D)

Facts: U

Conclusions: $U, \neg A, C, B, D$

[9] This problem, I speculate, was one of the primary forces for the emergence of probabilistic causality in the 1960s (see Section 7.5 for review).

Note that it is only the value of U, the background variable, that is carried over from step 1 to step 2; all other propositions must be reevaluated subject to the new modification of the model. This reflects the understanding that background factors U are not affected by either the variables or the mechanisms in the model $\{f_i\}$; hence, the counterfactual consequent (in our case, D) must be evaluated under the same background conditions as those prevailing in the actual world. In fact, the background variables are the main carriers of information from the actual world to the hypothetical world; they serve as the "guardians of invariance" (or persistence) in the dynamic process that transforms the former into the latter (an observation by David Heckerman, personal communication).

Note also that this two-step procedure of evaluating counterfactuals can be combined into one. If we use an asterisk to distinguish postmodification from premodification variables, then we can combine M and M_x into one logical theory and prove the validity of S5 by purely logical deduction in the combined theory. To illustrate, we write S5 as $D \implies D^*_{\neg A^*}$ (read: If D is true in the actual world, then D would also be true in the hypothetical world created by the modification $\neg A^*$) and prove the validity of D^* in the combined theory as follows.

Combined Theory

		(U)
$C^* = U$	$C = U$	(C)
$\neg A^*$	$A = C$	(A)
$B^* = C^*$	$B = C$	(B)
$D^* = A^* \vee B^*$	$D = A \vee B$	(D)

Facts: D

Conclusions: $U, A, B, C, D, \neg A^*, C^*, B^*, D^*$

Note that U need not be "starred," reflecting the assumption that background conditions remain unaltered.

It is worth reflecting at this point on the difference between S4 and S5. The two appear to be syntactically identical, as both involve a fact implying a counterfactual, and yet we labeled S4 an "action" sentence and S5 a "counterfactual" sentence. The difference lies in the relationship between the given fact and the antecedent of the counterfactual (i.e., the "action" part). In S4, the fact given ($\neg C$) is not affected by the antecedent (A); in S5, the fact given (D) is potentially affected by the antecedent ($\neg A$). The difference between these two situations is fundamental, as can be seen from their methods of evaluation. In evaluating S4, we knew in advance that C would not be affected by the model modification $do(A)$; therefore, we were able to add C directly to the modified model M_A. In evaluating S5, on the other hand, we were contemplating a possible reversal, from D to $\neg D$, attributable to the modification $do(\neg A)$. As a result, we first had to add fact D to the preaction model M, summarize its impact via U, and reevaluate D once the modification $do(\neg A)$ takes place. Thus, although the causal effect of actions can be expressed syntactically as a counterfactual sentence, this need to route the impact of known facts through U makes counterfactuals a different species than actions (see Section 1.4).

We should also emphasize that most counterfactual utterances in natural language presume, often implicitly, knowledge of facts that are affected by the antecedent. For

example, when we say that "B would be different were it not for A," we imply knowledge of what the actual value of B is and that B is susceptible to A. It is this sort of relationship that gives counterfactuals their unique character – distinct from action sentences – and, as we saw in Section 1.4, it is this sort of sentence that would require a more detailed specification for its evaluation: some knowledge of the functional mechanisms $f_i(pa_i, u_i)$ would be necessary.

7.1.3 Evaluating Counterfactuals: Probabilistic Analysis

To demonstrate the probabilistic evaluation of counterfactuals (equations (7.3)–(7.5)), let us modify the firing-squad story slightly, assuming that:

1. there is a probability $P(U) = p$ that the court has ordered the execution;
2. rifleman A has a probability q of pulling the trigger out of nervousness; and
3. rifleman A's nervousness is independent of U.

With these assumptions, we wish to compute the quantity $P(\neg D_{\neg A} \mid D)$ – namely, the probability that the prisoner would be alive if A had not shot, given that the prisoner is in fact dead.

Intuitively, we can figure out the answer by noting that $\neg D_{\neg A}$ is true if and only if the court has not issued an order. Thus, our task amounts to that of computing $P(\neg U \mid D)$, which evaluates to $q(1-p)/[1-(1-q)(1-p)]$. However, our aim is to demonstrate a general and formal method of deriving such probabilities, based on (7.4), that makes little use of intuition.

The probabilistic causal model (Definition 7.1.6) associated with the new story contains two background variables, U and W, where W stands for rifleman A's nervousness. This model is given as follows.

Model $\langle M, P(u, w)\rangle$

$$(U, W) \sim P(u, w)$$

$$\begin{aligned} C &= U && (C)\\ A &= C \vee W && (A)\\ B &= C && (B)\\ D &= A \vee B && (D) \end{aligned}$$

In this model, the background variables are distributed as

$$P(u, w) = \begin{cases} pq & \text{if } u = 1,\ w = 1,\\ p(1-q) & \text{if } u = 1,\ w = 0,\\ (1-p)q & \text{if } u = 0,\ w = 1,\\ (1-p)(1-q) & \text{if } u = 0,\ w = 0. \end{cases} \tag{7.7}$$

Following Theorem 7.1.7, our first step (abduction) is to compute the posterior probability $P(u, w \mid D)$, accounting for the fact that the prisoner is found dead. This is easily evaluated to:

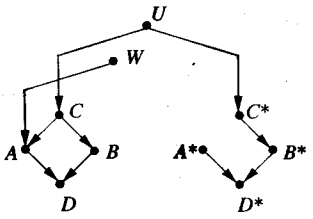

Figure 7.2 Twin network representation of the firing squad.

$$P(u, w \mid D) = \begin{cases} \frac{P(u,w)}{1-(1-p)(1-q)} & \text{if } u = 1 \text{ or } w = 1,\\ 0 & \text{if } u = 0 \text{ and } w = 0. \end{cases} \tag{7.8}$$

The second step (action) is to form the submodel $M_{\neg A}$ while retaining the posterior probability of (7.8).

Model $\langle M_{\neg A}, P(u, w \mid D)\rangle$

$$(U, W) \sim P(u, w \mid D)$$

$$\begin{aligned} C &= U && (C)\\ \neg A & && (A)\\ B &= C && (B)\\ D &= A \vee B && (D) \end{aligned}$$

The last step (prediction) is to compute $P(\neg D)$ in this probabilistic model. Noting that $\neg D = \neg U$, the result (as expected) is

$$P(\neg D_{\neg A} \mid D) = P(\neg U \mid D) = \frac{q(1-p)}{1-(1-q)(1-p)}.$$

7.1.4 The Twin Network Method

A major practical difficulty in the procedure just described is the need to compute, store, and use the posterior distribution $P(u \mid e)$, where u stand for the set of all background variables in the model. As illustrated in the preceding example, even when we start with Markovian model in which the background variables are mutually independent, conditioning on e normally destroys this independence and so makes it necessary to carry over a full description of the joint distribution of U, conditional on e. Such description may be prohibitively large if encoded in the form of a table, as we have done in (7.8).

A graphical method of overcoming this difficulty is described in Balke and Pearl (1994b); it uses two networks, one to represent the actual world and one to represent the hypothetical world. Figure 7.2 illustrates this construction for the firing-squad story analyzed.

The two networks are identical in structure, save for the arrows entering A^*, which have been deleted to mirror the equation deleted from $M_{\neg A}$. Like Siamese twins, the two networks share the background variables (in our case, U and W), since those remain invariant under modification. The endogenous variables are replicated and labeled distinctly, because they may obtain different values in the hypothetical versus the actual

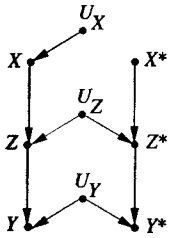

Figure 7.3 Twin network representation of the counterfactual Y_x in the model $X \rightarrow Z \rightarrow Y$.

world. The task of computing $P(\neg D)$ in the model $\langle M_{\neg A}, P(u, v \mid D)\rangle$ thus reduces to that of computing $P(\neg D^* \mid D)$ in the twin network shown, setting A^* to false.

In general, if we wish to compute the counterfactual probability $P(Y_x = y \mid z)$, where X, Y, and Z are arbitrary sets of variables (not necessarily disjoint), Theorem 7.1.7 instructs us to compute $P(y)$ in the submodel $\langle M_x, P(u \mid z)\rangle$, which reduces to computing an ordinary conditional probability $P(y^* \mid z)$ in an augmented Bayesian network. Such computation can be performed by standard evidence propagation techniques. The advantages of delegating this computation to inference in a Bayesian network are that the distribution $P(u \mid z)$ need not be explicated, conditional independencies can be exploited, and local computation methods can be employed (such as those summarized in Section 1.2.4).

The twin network representation also offers a useful way of testing independencies among counterfactual quantities. To illustrate, suppose that we have a chainlike causal diagram, $X \rightarrow Z \rightarrow Y$, and that we wish to test whether Y_x is independent of X given Z (i.e., $Y_x \perp\!\!\!\perp X \mid Z$). The twin network associated with this chain is shown in Figure 7.3. To test whether $Y_x \perp\!\!\!\perp X \mid Z$ holds in the original model, we test whether Z d-separates X from Y^* in the twin network. As can be easily seen (via Definition 1.2.3), conditioning on Z renders the path between X and Y^* d-connected through the collider at Z and hence $Y_x \perp\!\!\!\perp X \mid Z$ does not hold in the model. This conclusion is not easily discernible from the chain model itself or from the equations in that model. In the same fashion, we can see that whenever we condition on either Y or on $\{Y, Z\}$, we form a connection between Y^* and X; hence, Y_x and X are not independent conditional on those variables. The connection is disrupted, however, if we do not condition on either Y or Z, in which case $Y_x \perp\!\!\!\perp X$.

The twin network reveals an interesting interpretation of counterfactuals of the form Z_{pa_Z}, where Z is any variable and PA_Z stands for the set of Z's parents. Consider the question of whether Z_x is independent of some given set of variables in the model of Figure 7.3. The answer to this question depends on whether Z^* is d-separated from that set of variables. However, any variable that is d-separated from Z^* would also be d-separated from U_Z, so the node representing U_Z can serve as a proxy for representing the counterfactual variable Z_x. This is not a coincidence, considering that Z is governed by the equation $z = f_Z(x, u_Z)$. By definition, the distribution of Z_x is equal to the distribution of Z under the condition where X is held fixed at x. Under such condition, Z may vary only if U_Z varies. Therefore, if U_Z obeys a certain independence relationship then Z_x (more generally, Z_{pa_Z}) must obey that relationship as well. We thus obtain a simple graphical

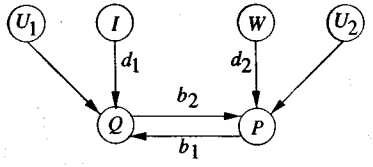

Figure 7.4 Causal diagram illustrating the relationship between price (P) and demand (Q).

representation for any counterfactual variable of the form Z_{pa_Z}. Using this representation, we can easily verify from Figure 7.3 that $(Y^* \perp\!\!\!\perp X \mid \{Z, U_Z, Y\})_G$ and $(Y^* \perp\!\!\!\perp X \mid \{U_Y, U_Z, Y\})_G$ both hold in the twin network and therefore

$$Y_x \perp\!\!\!\perp X \mid \{Z, Z_x, Y\} \quad \text{and} \quad Y_x \perp\!\!\!\perp X \mid \{Y_z, Z_x, Y\}$$

must hold in the model. The verification of such independencies is important for deciding the identification of plans, because these independencies permit us to reduce counterfactual probabilities to ordinary probabilistic expression on observed variables (see Section 7.3.2).

7.2 APPLICATIONS AND INTERPRETATION OF STRUCTURAL MODELS

7.2.1 Policy Analysis in Linear Econometric Models: An Example

In Section 1.4 we illustrated the nature of structural equations modeling using the canonical economic problem of demand and price equilibrium (see Figure 7.4). In this chapter, we use this problem to answer policy-related questions.

To recall, this example consists of the two equations

$$q = b_1 p + d_1 i + u_1, \tag{7.9}$$

$$p = b_2 q + d_2 w + u_2, \tag{7.10}$$

where q is the quantity of household demand for a product A, p is the unit price of product A, i is household income, w is the wage rate for producing product A, and u_1 and u_2 represent error terms – unmodeled factors that affect quantity and price, respectively (Goldberger 1992).

This system of equations constitutes a causal model (Definition 7.1.1) if we define $V = \{Q, P\}$ and $U = \{U_1, U_2, I, W\}$ and assume that each equation represents an autonomous process in the sense of Definition 7.1.3. It is normally assumed that I and W are observed, while U_1 and U_2 are unobservable and independent in I and W. Since the error terms U_1 and U_2 are unobserved, a complete specification of the model must include the distribution of these errors, which is usually taken to be Gaussian with the covariance matrix $\Sigma_{ij} = \text{cov}(u_i, u_j)$. It is well known in economics (dating back to Wright 1928) that the assumptions of linearity, normality, and the independence of $\{I, W\}$ and $\{U_1, U_2\}$ permit consistent estimation of all model parameters, including the covariance matrix Σ_{ij}. However, the focus of this book is not the estimation of parameters but rather their

utilization in policy predictions. Accordingly, we will demonstrate how to evaluate the following three queries.

1. What is the expected value of the demand Q if the price is *controlled at* $P = p_0$?
2. What is the expected value of the demand Q if the price is *reported to be* $P = p_0$?
3. Given that the current price is $P = p_0$, what would be the expected value of the demand Q if we *were to control* the price at $P = p_1$?

The reader should recognize these queries as representing (respectively) actions, predictions, and counterfactuals – our three-level hierarchy. The second query, representing prediction, is standard in the literature and can be answered directly from the covariance matrix without reference to causality, structure, or invariance. The first and third queries rest on the structural properties of the equations and, as expected, are not treated in the standard literature of structural equations.[10]

In order to answer the first query, we replace (7.10) with $p = p_0$, leaving

$$q = b_1 p + d_1 i + u_1, \tag{7.11}$$

$$p = p_0, \tag{7.12}$$

with the statistics of U_1 and I unaltered. The controlled demand is then $q = b_1 p_0 + d_1 i + u_1$, and its expected value (conditional on $I = i$) is given by

$$E[Q \mid do(P = p_0), i] = b_1 p_0 + d_1 i + E(U_1 \mid i). \tag{7.13}$$

Since U_1 is independent of I, the last term evaluates to

$$E(U_1 \mid i) = E(U_1) = E(Q) - b_1 E(P) - d_1 E(I)$$

and, substituted into (7.13), yields

$$E[Q \mid do(P = p_0), i] = E(Q) + b_1(p_0 - E(P)) + d_1(i - E(I)).$$

The answer to the second query is obtained by conditioning (7.9) on the current observation $\{P = p_0, I = i, W = w\}$ and taking the expectation,

$$E(Q \mid p_0, i, w) = b_1 p_0 + d_1 i + E(U_1 \mid p_0, i, w). \tag{7.14}$$

The computation of $E[U_1 \mid p_0, i, w]$ is a standard procedure once Σ_{ij} is given (Whittaker 1990, p. 163). Note that, although U_1 was assumed to be independent of I and W, this independence no longer holds once $P = p_0$ is observed. Note also that (7.9) and (7.10)

[10] I have presented this example to well over a hundred econometrics students and faculty across the United States. Respondents had no problem answering question 2, one person was able to solve question 1, and none managed to answer question 3. Chapter 5 (Section 5.1) suggests an explanation.

both participate in the solution and that the observed value p_0 will affect the expected demand Q (through $E(U_1 \mid p_0, i, w)$) even when $b_1 = 0$, which is not the case in query 1.

The third query requires the expectation of the counterfactual quantity $Q_{P=p_1}$, conditional on the current observations $\{P = p_0, I = i, W = w\}$. According to Definition 7.1.5, $Q_{P=p_1}$ is governed by the submodel

$$q = b_1 p + d_1 i + u_1, \tag{7.15}$$

$$p = p_1; \tag{7.16}$$

the density of u_1 should be conditioned on the observations $\{P = p_0, I = i, W = w\}$. We therefore obtain

$$E(Q_{p=p_1} \mid p_0, i, w) = b_1 p_1 + d_1 i + E(U_1 \mid p_0, i, w). \tag{7.17}$$

The expected value $E(U_1 \mid p_0, i, w)$ is the same as in the solution to the second query; the latter differs only in the term $b_1 p_1$. A general matrix method for evaluating counterfactual queries in linear Gaussian models is described in Balke and Pearl (1995).

At this point, it is worth emphasizing that the problem of computing counterfactual expectations is not an academic exercise; it represents in fact the typical case in almost every decision-making situation. Whenever we undertake to predict the effect of policy, two considerations apply. First, the policy variables (e.g., price and interest rates in economics, pressure and temperature in process control) are rarely exogenous. Policy variables are endogenous when we observe a system under operation; they become exogenous in the planning phase, when we contemplate actions and changes. Second, policies are rarely evaluated in the abstract; rather, they are brought into focus by certain eventualities that demand remedial correction. In troubleshooting, for example, we observe undesirable effects e that are influenced by other conditions $X = x$ and wish to predict whether an action that brings about a change in X would remedy the situation. The information provided by e is extremely valuable, and it must be processed (using abduction) before we can predict the effect of any action. This step of abduction endows practical queries about actions with a counterfactual character, as we have seen in the evaluation of the third query (7.17).

The current price p_0 reflects economic conditions (e.g. Q) that prevail at the time of decision, and these conditions are presumed to be changeable by the policies considered. Thus, the price P represents an endogenous decision variable (as shown in Figure 7.4) that becomes exogenous in deliberation, as dictated by the submodel $M_{P=p_1}$. The hypothetical mood of query 3 translates into a practical problem of policy analysis: "Given that the current price is $P = p_0$, find the expected value of the demand (Q) if we change the price *today* to $P = p_1$." The reasons for using hypothetical phrases in practical decision-making situations are discussed in the next section.

7.2.2 The Empirical Content of Counterfactuals

The word "counterfactual" is a misnomer, since it connotes a statement that stands contrary to facts or, at the very least, a statement that escapes empirical verification. Counterfactuals are in neither category; they are fundamental to scientific thought and carry as clear an empirical message as any scientific law.

Consider Ohm's law, $V = IR$. The empirical content of this law can be encoded in two alternative forms.

1. *Predictive form:* If at time t_0 we measure current I_0 and voltage V_0 then, ceteris paribus, at any future times $t > t_0$, if the current flow is $I(t)$ then the voltage will be
$$V(t) = \frac{V_0}{I_0} I(t).$$
2. *Counterfactual form:* If at time t_0 we measure current I_0 and voltage V_0 then, had the current flow at time t_0 been I' instead of I_0, the voltage would have been
$$V' = \frac{V_0 I'}{I_0}.$$

On the surface, it seems that the predictive form makes meaningful and testable empirical claims whereas the counterfactual form merely speculates about events that have not (and could not have) occurred, since it is impossible to apply two different currents into the same resistor at the same time. However, if we interpret the counterfactual form to be neither more nor less than a conversational shorthand of the predictive form, the empirical content of the former shines through clearly. Both enable us to make an infinite number of predictions from just one measurement (I_0, V_0), and both derive their validity from a scientific law that ascribes a time-invariant property (the ratio V/I) to any object that conducts electricity.

But if counterfactual statements are merely a roundabout way of stating sets of predictions, why do we resort to such convoluted modes of expression instead of using the predictive mode directly? One obvious answer is that we often use counterfactuals to convey not the predictions themselves but rather the logical ramifications of those predictions. For example, the intent of saying: "if A were not to have shot, then the prisoner would still be alive" may be merely to convey the factual information that B did not shoot. The counterfactual mood, in this case, serves to supplement the fact conveyed with logical justification based on a general law. The less obvious answer rests with the ceteris paribus (all else held equal) qualification that accompanies the predictive claim, which is not entirely free of ambiguities. What should be held constant when we change the current in a resistor – the temperature? the laboratory equipment? the time of day? Certainly not the reading on the voltmeter!

Such matters must be carefully specified when we pronounce predictive claims and take them seriously. Many of these specifications are implicit (and hence superfluous) when we use counterfactual expressions, especially when we agree on the underlying causal model. For example, we do not need to specify under what temperature and pressure the predictions should hold true; these are implied by the statement "had the current flow at time t_0 been I', instead of I_0." In other words, we are referring to precisely those conditions that prevailed in our laboratory at time t_0. The statement also implies that we do not really mean for anyone to hold the reading on the voltmeter constant; variables should run their natural course, and the only change we should envision is in the mechanism that (according to our causal model) is currently determining the current.

To summarize, a counterfactual statement might well be interpreted as conveying a set of predictions under a well-defined set of conditions – those prevailing in the factual part of the statement. For these predictions to be valid, two components must remain invariant: the laws (or mechanisms) and the boundary conditions. Cast in the language of structural models, the laws correspond to the equations $\{f_i\}$ and the boundary conditions correspond to the state of the background variables U. Thus, a precondition for the validity of the predictive interpretation of a counterfactual statement is the assumption that U will not change when our predictive claim is to be applied or tested.

This is best illustrated by using a betting example. We must bet heads or tails on the outcome of a fair coin toss; we win a dollar if we guess correctly and lose one if we don't. Suppose we bet heads and win a dollar, without glancing at the outcome of the coin. Consider the counterfactual "Had I bet differently I would have lost a dollar." The predictive interpretation of this sentence translates into the implausible claim: "If my next bet is tails, I will lose a dollar." For this claim to be valid, two invariants must be assumed: the payoff policy and the outcome of the coin. Whereas the former is a plausible assumption in a betting context, the latter would be realized in only rare circumstances. It is for this reason that the predictive utility of the statement "Had I bet differently I would have lost a dollar" is rather low, and some would even regard it as hindsighted nonsense. It is the persistence across time of U and $f(x, u)$ that endows counterfactual expressions with predictive power; absent this persistence, the counterfactual loses its obvious predictive utility.

However, there is an element of utility in counterfactuals that does not translate immediately to predictive payoff and thus may serve to explain the ubiquity of counterfactuals in human discourse. I am thinking of explanatory value. Suppose, in the betting story, coins were tossed afresh for every bet. Is there no value whatsoever to the statement "Had I bet differently I would have lost a dollar?" I believe there is; it tells us that we are not dealing here with a whimsical bookie but instead with one who at least glances at the bet, compares it to some standard, and decides a win or a loss using a consistent policy. This information may not be very useful to us as players, but it may be useful to, say, state inspectors who come every so often to calibrate the gambling machines and so ensure the state's take of the profit. More significantly, it may be useful to us players, too, if we venture to cheat slightly – say, by manipulating the trajectory of the coin, or by installing a tiny transmitter to tell us which way the coin landed. For such cheating to work, we should know the payoff policy $y = f(x, u)$, and the statement "Had I bet differently I would have lost a dollar" reveals important aspects of that policy.

Is it far-fetched to argue for the merit of counterfactuals by hypothesizing unlikely situations where players cheat and rules are broken? I suggest that such unlikely operations are precisely the norm for gauging the explanatory value of sentences. It is the nature of any causal explanation that its utility be proven not over standard situations but rather over novel settings that require innovative manipulations of the standards. The utility of understanding how television works comes not from turning the knobs correctly but from the ability to repair a TV set when it breaks down. Recall that every causal model advertises not one but rather a host of submodels, each created by violating some laws. The autonomy of the mechanisms in a causal model thus stands for an open invitation to

remove or replace those mechanisms, and it is only natural that the explanatory value of sentences be judged by how well they predict the ramifications of such replacements.

Counterfactuals with Intrinsic Nondeterminism

Recapping our discussion, we see that counterfactuals may earn predictive value under two conditions: (1) when the unobserved uncertainty-producing variables (U) remain constant (until our next prediction or action); or (2) when the uncertainty-producing variables offer the potential of being observed sometime in the future (before our next prediction or action). In both cases, we also need to ensure that the outcome-producing mechanism $f(x, u)$ persists unaltered.

These conclusions raise interesting questions regarding the use of counterfactuals in microscopic phenomena, as none of these conditions holds for the type of uncertainty that we encounter in quantum theory. Heisenberg's die is rolled afresh billions of times each second, and our measurement of U will never be fine enough to remove all uncertainty from the response equation $y = f(x, u)$. Thus, when we include quantum-level processes in our analysis we face a dilemma: either dismiss all talk of counterfactuals (a strategy recommended by some researchers, including Dawid 1997) or continue to use counterfactuals but limit their usage to situations where they assume empirical meaning. This amounts to keeping in our analysis only those U that satisfy conditions (1) and (2) of the previous paragraph. Instead of hypothesizing U that completely remove all uncertainties, we admit only those U that are either (1) persistent or (2) potentially observable.

Naturally, coarsening the granularity of the background variables has its price: the mechanism equations $v_i = f_i(pa_i, u_i)$ lose their deterministic character and hence should be made stochastic. Instead of constructing causal models from a set of deterministic equations $\{f_i\}$, we should consider models made up of stochastic functions $\{f_i^*\}$, where each f_i^* is a mapping from $V \cup U$ to some intrinsic probability distribution $P^*(v_i)$ over the states of V_i. This option leads to a causal Bayesian network (Section 1.3) in which the conditional probabilities $P^*(v_i \mid pa_i, u_i)$ represent intrinsic nondeterminism (sometimes called "objective chance"; Skyrms 1980) and in which the root nodes represent background variables U that are either persistent or potentially observable. In this representation, counterfactual probabilities $P(Y_x = y \mid e)$ can still be evaluated using the three steps (abduction, action, and prediction) of Theorem 7.1.7. In the abduction phase, we condition the prior probability $P(u)$ of the root nodes on the evidence available, e, and so obtain $P(u \mid e)$. In the action phase, we delete the arrows entering variables in set $\boldsymbol{X}$ and instantiate their values to $X = x$. Finally, in the prediction phase, we compute the probability of $Y = y$ resulting from the updated manipulated network.

This evaluation can, of course, be implemented in ordinary causal Bayesian networks (i.e., not only in ones that represent intrinsic nondeterminism), but in that case the results computed would not represent the probability of the counterfactual $Y_x = y$. Such evaluation amounts to assuming that units are homogeneous, with each possessing the stochastic properties of the population – namely, $P(v_i \mid pa_i, u) = P(v_i \mid pa_i)$. Such an assumption may be adequate in quantum-level phenomena, where units stands for specific experimental conditions, but it will not be adequate in macroscopic phenomena, where units may differ appreciably from each other. In the example of Chapter 1 (Section 1.4.4, Figure 1.6), the stochastic attribution amounts to assuming that no individual

is affected by the drug (as dictated by model 1) while ignoring the possibility that some individuals may, in fact, be more sensitive to the drug than others (as in model 2).

7.2.3 Causal Explanations, Utterances, and Their Interpretation

It is a commonplace wisdom that explanation improves understanding and that he who understands more can reason and learn more effectively. It is also generally accepted that the notion of explanation cannot be divorced from that of causation; for example, a symptom may explain our *belief* in a disease, but it does not explain the disease itself. However, the precise relationship between causes and explanations is still a topic of much discussion (Cartwright 1989; Woodward 1997). Having a formal theory of causality and counterfactuals in both deterministic and probabilistic settings casts new light on the question of what constitutes an adequate explanation, and it opens new possibilities for automatic generation of explanations by machine.

A natural starting point for generating explanations would be to use a causal Bayesian network (Section 1.3) in which the events to be explained (explanadum) consist of some combination e of instantiated nodes in the network, and where the task is to find an instantiation c of a subset of e's ancestors (i.e. causes) that maximizes some measure of "explanatory power," namely, the degree to which c explains e. However, the proper choice of this measure is unsettled. Many philosophers and statisticians argue for the likelihood ratio $L = \frac{P(e|c)}{P(e|c')}$ as the proper measure of the degree to which c is a better explanation of e than c'. In Pearl (1988b, chap. 5) and Peng and Reggia (1986), the best explanation is found by maximizing the posterior probability $P(c \mid e)$. Both measures have their faults and have been criticized by several researchers, including Pearl (1988b), Shimony (1991, 1993), Suermondt and Cooper (1993), and Chajewska and Halpern (1997). To remedy these faults, more intricate combinations of the probabilistic parameters $[P(e \mid c), P(e \mid c'), P(c), P(c')]$ have been suggested, none of which seems to capture well the meaning people attach to the word "explanation."

The problem with probabilistic measures is that they cannot capture the strength of a *causal* connection between c and e; any proposition h whatsoever can, with a small stretch of imagination, be thought of as having some influence on e, however feeble. This would then qualify h as an ancestor of e in the causal network and would permit h to compete and win against genuine explanations by virtue of h having strong spurious association with e.

To rid ourselves of this difficulty, we must go beyond probabilistic measures and concentrate instead on causal parameters, such as causal effects $P(y \mid do(x))$ and counterfactual probabilities $P(Y_{x'} = y' \mid x, y)$, as the basis for defining explanatory power. Here x and x' range over the set of alternative explanations, and Y is the set of response variables observed to take on the value y. The expression $P(Y_{x'} = y' \mid x, y)$ is read as: the probability that Y would take on a different value, y', had X been x' (instead of the actual values x). (Note that $P(y \mid do(x)) \triangleq P(Y_x = y)$.) The developments of computational models for evaluating causal effects and counterfactual probabilities now make it possible to combine these parameters with standard probabilistic parameters and so synthesize a more faithful measure of explanatory power that may guide the selection and generation of adequate explanations.

These possibilities trigger an important basic question: Is "explanation" a concept based on *general* causes (e.g., "Drinking hemlock causes death") or *singular* causes (e.g., "Socrates' drinking hemlock caused his death")? Causal effect expressions $P(y \mid do(x))$ belong to the first category whereas counterfactual expressions $P(Y_{x'} = y' \mid x, y)$ belong to the second, since conditioning on x and y narrows down world scenarios to those compatible with the most specific information at hand: $X = x$ and $Y = y$.

The classification of causal statements into general and singular categories has been the subject of intensive research in philosophy (see e.g. Good 1961; Kvart 1986; Cartwright 1989; Eells 1991; see also discussions in Sections 7.5.4 and 10.1.1). This research has attracted little attention in cognitive science and artificial intelligence, partly because it has not entailed practical inferential procedures and partly because it is based on problematic probabilistic semantics (see Section 7.5 for discussion of probabilistic causality). In the context of machine-generated explanations, this classification assumes both cognitive and computational significance. We discussed in Chapter 1 (Section 1.4) the sharp demarcation line between two types of causal queries, those that are answerable from the pair $\langle P(M), G(M) \rangle$ (the probability and diagram, respectively, associated with model M) and those that require additional information in the form of functional specification. Generic causal statements (e.g., $P(y \mid do(x))$) often fall into the first category (as in Chapter 3) whereas counterfactual expressions (e.g., $P(Y_{x'} = y \mid x, y)$) fall into the second, thus demanding more detailed specifications and higher computational resources.

The proper classification of explanation into a general or singular category depends on whether the cause c attains its explanatory power relative to its effect e by virtue of c's general *tendency* to produce e (as compared with the weaker tendencies of c's alternatives) or by virtue of c being *necessary* for triggering a specific chain of events leading to e in the specific situation at hand (as characterized by e and perhaps other facts and observations). Formally, the difference hinges on whether, in evaluating explanatory powers of various hypotheses, we should condition our beliefs on the events c and e that actually occurred.

Formal analysis of these alternatives is given in Chapters 9 and 10, where we discuss the necessary and sufficient aspects of causation as well as the notion of single-event causation. In the balance of this section we will be concerned with the interpretation and generation of explanatory utterances, taking the necessary aspect as a norm.

The following list, taken largely from Galles and Pearl (1997), provides examples of utterances used in explanatory discourse and their associated semantics within the modifiable structural model approach described in Section 7.1.1.

- "X is a cause of Y" if there exist two values x and x' of X and a value u of U such that $Y_x(u) \neq Y_{x'}(u)$.
- "X is a cause of Y in the context $Z = z$" if there exist two values x and x' of X and a value u of U such that $Y_{xz}(u) \neq Y_{x'z}(u)$.
- "X is a direct cause of Y" if there exist two values x and x' of X and a value u of U such that $Y_{xr}(u) \neq Y_{x'r}(u)$, where r is some realization of $V \setminus \{X, Y\}$.
- "X is an indirect cause of Y" if X is a cause of Y and X is not a direct cause of Y.

- "Event $X = x$ always causes $Y = y$" if:
  (i) $Y_x(u) = y$ for all u; and
  (ii) there exists a value u' of U such that $Y_{x'}(u') \neq y$ for some $x' \neq x$.
- "Event $X = x$ may have caused $Y = y$" if:
  (i) $X = x$ and $Y = y$ are true; and
  (ii) there exists a value u of U such that $X(u) = x$, $Y(u) = y$, and $Y_{x'}(u) \neq y$ for some $x' \neq x$.
- "The unobserved event $X = x$ is a likely cause of $Y = y$" if:
  (i) $Y = y$ is true; and
  (ii) $P(Y_x = y, Y_{x'} \neq y \mid Y = y)$ is high for all $x' \neq x$.
- "Event $Y = y$ occurred despite $X = x$" if:
  (i) $X = x$ and $Y = y$ are true; and
  (ii) $P(Y_x = y)$ is low.

The preceding list demonstrates the flexibility of modifiable structural models in formalizing nuances of causal expressions. Additional nuances (invoking such notions as enabling, preventing, sustaining, producing, etc.) will be analyzed in Chapters 9 and 10. Related expressions include: "Event A explains the occurrence of event B"; "A would explain B if C were the case"; "B occurred despite A because C was true." The ability to interpret and generate such explanatory sentences, or to select the expression most appropriate for the context, is one of the most intriguing challenges of research in man–machine conversation.

7.2.4 From Mechanisms to Actions to Causation

The structural model semantics described in Section 7.1.1 suggests solutions to two problems in cognitive science and artificial intelligence: the representation of actions and the role of causal ordering. We will discuss these problems in turn, since the second builds on the first.

Action, Mechanisms, and Surgeries

Whether we take the probabilistic paradigm that actions are transformations from probability distributions to probability distributions or the deterministic paradigm that actions are transformations from states to states, such transformations could in principle be infinitely complex. Yet in practice, people teach each other rather quickly the normal results of actions in the world, and people predict the consequences of most actions without much trouble. How?

Structural models answer this question by assuming that the actions we normally invoke in common reasoning can be represented as *local surgeries*. The world consists of a huge number of autonomous and invariant linkages or mechanisms, each corresponding to a physical process that constrains the behavior of a relatively small group of variables. If we understand how the linkages interact with each other (usually, they simply share variables), then we should also be able to understand what the effect of any given action would be: simply respecify those few mechanisms that are perturbed by the action; then let the mechanisms in the modified assembly interact with one another and see what state

will evolve at equilibrium. If the specification is complete (i.e., if M and U are given), then a single state will evolve. If the specification is probabilistic (i.e., if $P(u)$ is given), then a new probability distribution will emerge; if the specification is partial (i.e., if some f_i are not given), then a new, partial theory will be created. In all three cases we should be able to answer queries about postaction states of affair, albeit with decreasing level of precision.

The ingredient that makes this scheme operational is the *locality* of actions. Standing alone, locality is a vague concept because what is local in one space may not be local in another. A speck of dust, for example, appears extremely diffused in the frequency (or Fourier) representation; conversely, a pure musical tone requires a long stretch of time to be appreciated. Structural semantics emphasizes that actions are local in the space of mechanisms and not in the space of variables or sentences or time slots. For example, tipping the leftmost object in an array of domino tiles does not appear to be "local" in physical space, yet it is quite local in the mechanism domain: only one mechanism is perturbed, the gravitational restoring force that normally keeps that tile in a stable erect position; all other mechanisms remain unaltered, as specified, obedient to the usual equations of physics. Locality makes it easy to specify this action without enumerating all its ramifications. The listener, assuming she shares our understanding of domino physics, can figure out for herself the ramifications of this action, or any action of the type: "tip the ith domino tile to the right." By representing the domain in the form of an assembly of stable mechanisms, we have in fact created an oracle capable of answering queries about the effects of a huge set of actions and action combinations – without us having to explicate those effects.

Laws versus Facts

This surgical procedure sounds trivial when expressed in the context of structural equation models. However, it has encountered great difficulties when attempts were made to implement such schemes in classical logic. In order to implement surgical procedures in mechanism space, we need a language in which some sentences are given different status than others. Sentences describing mechanisms should be treated differently than those describing other facts of life (e.g., observations, assumptions, and conclusions), because the former are presumed to be stable whereas the latter are transitory. Indeed, the equations describing how the domino tiles interact with one another remain unaltered even though the states of the tiles themselves are free to vary with circumstances.

Admitting the need for this distinction has been a difficult transition in the logical approach to actions and causality, perhaps because much of the power of classical logic stems from its representational uniformity and syntactic invariance, where no sentence commands special status. Probabilists were much less reluctant to embrace the distinction between laws and facts, because this distinction has already been programmed into probability language by Reverend Bayes in 1763: Facts are expressed as ordinary propositions and hence can obtain probability values and can be conditioned on; laws, on the other hand, are expressed as conditional probability sentences (e.g., $P(\text{accident} \mid \text{careless driving}) = \text{high}$) and hence should not be assigned probabilities and cannot be conditioned on. It is because of this tradition that probabilists have always attributed nonpropositional character to conditional sentences (e.g., birds fly), refused to allow nested

conditionals (Levi 1988), and insisted on interpreting one's confidence in a conditional sentence as a conditional probability judgment (Adams 1975; see also Lewis 1976). Remarkably, these constraints, which some philosophers view as limitations, are precisely the safeguards that have kept probabilists from confusing laws and facts, protecting them from some of the traps that have ensnared logical approaches.[11]

Mechanisms and Causal Relationships

From our discussion thus far, it may seem that one can construct an effective representation for computing the ramification of actions without appealing to any notion of causation. This is indeed feasible in many areas of physics and engineering. For instance, if we have a large electric circuit consisting of resistors and voltage sources, and if we are interested in computing the effect of changing one resistor in the circuit, then the notion of causality hardly enters the computation. We simply insert the modified value of the resistor into Ohm's and Kirchhoff's equations and proceed to solve the set of (symmetric) equations for the variables needed. This computation can be performed effectively without committing to any directional causal relationship between the currents and voltages.

To understand the role of causality, we should note that (unlike our electrical circuit example) most mechanisms do not have names in common everyday language. We say: "raise taxes," or "make him laugh," or "press the button" – in general, $do(q)$, where q is a proposition, not a mechanism. It would be meaningless to say "increase this current" or "if this current were higher ..." in the electrical circuit example, because there are many ways of (minimally) increasing that current, each with different ramifications. Evidently, common-sense knowledge is not as entangled as a resistor network. In the STRIPS language (Fikes and Nilsson 1971), to give another example, an action is not characterized by the name of the mechanisms it modifies but rather by the action's immediate effects (the ADD and DELETE lists), and these effects are expressed as ordinary propositions. Indeed, if our knowledge is organized causally then this specification is sufficient, because each variable is governed by one and only one mechanism (see Definition 7.1.1). Thus, we should be able to figure out for ourselves which mechanism it is that must be perturbed in realizing the effect specified, and this should enable us to predict the rest of the scenario.

This linguistic abbreviation defines a new relation among events, a relation we normally call "causation": Event A causes B if the perturbation needed for realizing A entails the realization of B.[12] Causal abbreviations of this sort are used very effectively for specifying domain knowledge. Complex descriptions of what relationships are stable and how mechanisms interact with one another are rarely communicated explicitly in terms of mechanisms. Instead, they are communicated in terms of cause–effect relationships

[11] The distinction between laws and facts was proposed by Poole (1985) and Geffner (1992) as a fundamental principle for nonmonotonic reasoning. In database theory, laws are expressed by special sentences called *integrity constraints* (Reiter 1987). The distinction seems to be gaining broader support as a necessary requirement for formulating actions in artificial intelligence (Sandewall 1994; Lin 1995).

[12] The word "needed" connotes minimality and can be translated as: "... if every minimal perturbation realizing A entails B." The necessity and sufficiency aspects of this entailment relationship are formalized in Chapter 9 (Section 9.2).

between events or variables. We say, for example: "If tile i is tipped to the right, it causes tile $i+1$ to tip to the right as well"; we do not communicate such knowledge in terms of the tendencies of each domino tile to maintain its physical shape, to respond to gravitational pull, and to obey Newtonian mechanics.

7.2.5 Simon's Causal Ordering

Our ability to talk directly in terms of one event causing another, (rather than an action altering a mechanism and the alteration, in turn, producing the effect) is computationally very useful, but at the same time it requires that the assembly of mechanisms in our domain satisfy certain conditions that accommodate causal directionality. Indeed, the formal definition of causal models given in Section 7.1.1 assumes that each equation is designated a distinct privileged variable, situated on its left-hand side, that is considered "dependent" or "output." In general, however, a mechanism may be specified as a functional constraint

$$G_k(x_1, \ldots, x_l; u_1, \ldots, u_m) = 0$$

without identifying any "dependent" variable.

Simon (1953) devised a procedure for deciding whether a collection of such symmetric G functions dictates a unique way of selecting an endogenous dependent variable for each mechanisms (excluding the background variables, since they are determined outside the system). Simon asked: When can we order the variables $(V_1, V_2, \ldots, V_n)$ in such a way that we can solve for each V_i without solving for any of V_i's successors? Such an ordering, if it exists, dictates the direction we attribute to causation. This criterion might at first sound artificial, since the order of solving equations is a matter of computational convenience whereas causal directionality is an objective attribute of physical reality. (For discussion of this issue see De Kleer and Brown 1986; Iwasaki and Simon 1986; Druzdzel and Simon 1993.) To justify the criterion, let us rephrase Simon's question in terms of actions and mechanisms. Assume that each mechanism (i.e. equation) can be modified independently of the others, and let A_k be the set of actions capable of modifying equation G_k (while leaving other equations unaltered). Imagine that we have chosen an action a_k from A_k and that we have modified G_k in such a way that the set of solutions $(V_1(u), V_2(u), \ldots, V_n(u))$ to the entire system of equations differs from what it was prior to the action. If X is the set of endogenous variables constrained by G_k, then we can ask which members of X would change by the modification. If only one member of X changes, say X_k, and if the identity of that distinct member remains the same for all choices of a_k and u, then we designate X_k as the dependent variable in G_k.

Formally, this property means that changes in a_k induce a *functional mapping* from the domain of X_k to the domain of $\{V \setminus X_k\}$; all changes in the system (generated by a_k) can be attributed to changes in X_k. It would make sense, in such a case, to designate X_k as a "representative" of the mechanism G_k, and we would be justified in replacing the sentence "action a_k caused event $Y = y$" with "event $X_k = x_k$ caused $Y = y$" (where Y is any variable in the system). The invariance of X_k to the choice of a_k is the basis for treating an action as a modality $do(X_k = x_k)$ (Definition 7.1.3). It provides a license for characterizing an action by its immediate consequence(s), independent of the instrument

that actually brought about those consequences, and it defines in fact the notion of "local action" or "local surgery."

It can be shown (Nayak 1994) that the uniqueness of X_k can be determined by a simple criterion that involves purely topological properties of the equation set (i.e., how variables are grouped into equations). The criterion is that one should be able to form a one-to-one correspondence between equations and variables and that the correspondence be unique. This can be decided by solving the "matching problem" (Serrano and Gossard 1987) between equations and variables. If the matching is unique, then the choice of dependent variable in each equation is unique and the directionality induced by that choice defines a directed acyclic graph (DAG). In Figure 7.1, for example, the directionality of the arrows need not be specified externally; they can be determined mechanically from the set of symmetrical constraints (i.e., logical propositions)

$$S = \{G_1(C, U), G_2(A, C), G_3(B, C), G_4(A, B, D)\} \tag{7.18}$$

that characterizes the problem. The reader can easily verify that the selection of a privileged variable from each equation is unique and hence that the causal directionality of the arrows shown in Figure 7.1 is inevitable.

Thus, we see that causal directionality, according to Simon, emerges from two assumptions: (1) the partition of variables into background (U) and endogenous (V) sets; and (2) the overall configuration of mechanisms in the model. Accordingly, a variable designated as "dependent" in a given mechanism may well be labeled "independent" when that same mechanism is embedded in a different model. Indeed, the engine causes the wheels to turn when the train goes uphill but changes role in going downhill.

Of course, if we have no way of determining the background variables, then several causal orderings may ensue. In (7.18), for example, if we were not given the information that U is a background variable, then either one of $\{U, A, B, C\}$ could be chosen as background, and each such choice would induce a different ordering on the remaining variables. (Some would conflict with common-sense knowledge, e.g., that the captain's signal influences the court's decision.) However, the directionality of $A \rightarrow D \leftarrow B$ would be maintained in all those orderings. The question of whether there exists a partition $\{U, V\}$ of the variables that would yield a causal ordering in a system of symmetric constraints can also be solved (in polynomial time) by topological means (Dechter and Pearl 1991).

Simon's ordering criterion fails when we are unable to solve the equations one at a time and so must solve a block of k equations simultaneously. In such a case, all the k variables determined by the block would be mutually unordered, though their relationships with other blocks may still be ordered. This occurs, for example, in the economic model of Figure 7.4, where (7.9) and (7.10) need to be solved simultaneously for P and Q and hence the correspondence between equations and variables is not unique; either Q or P could be designated as "independent" in either of the two equations. Indeed, the information needed for classifying (7.9) as the "demand" equation (and, respectively, (7.10) as the "price" equation) comes not from the way variables are assigned to equations but rather from subject-matter considerations. Our understanding that household income directly affects household demand (and not prices) plays a major role in this classification.

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.

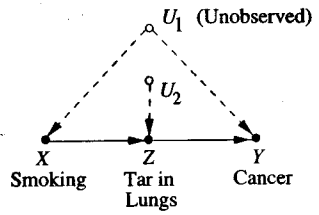

Figure 7.5 Causal diagram illustrating the effect of smoking on lung cancer.

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

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.

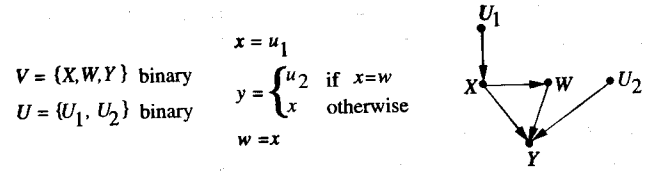

Figure 7.6 Example of a causal model that requires the examination of submodels to determine causal relevance.

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).

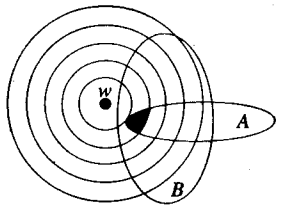

Figure 7.7 Graphical representation of Lewis's closest-world semantics. Each circular region corresponds to a set of worlds that are equally similar to w. The shaded region represents the set of closest A-worlds; since all these worlds satisfy B, the counterfactual sentence $A \mathrel{\Box\!\!\rightarrow}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

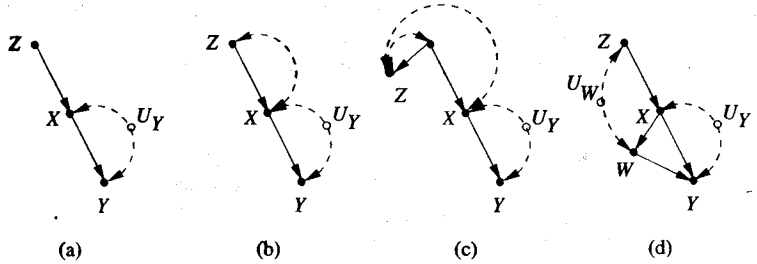

Figure 7.8 Z is a proper instrumental variable in the (linear) models of (a), (b), and (c), since it satisfies $Z \perp\!\!\!\perp U_Y$. It is not an instrument in (d) because it is correlated with U_W, which influences Y.

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

alone – that is, without the added assumptions of stability or minimality. By imposing the constraint that an effect never precede its cause, the symmetry is broken and causal inference can commence.

The reliance on temporal information has its price, as it excludes a priori the analysis of cases in which the temporal order is not well-defined, either because processes overlap in time or because they (appear to) occur instantaneously. For example, one must give up the prospect of determining (by uncontrolled methods) whether sustained physical exercise contributes to low cholesterol levels or if, conversely, low cholesterol levels enhance the urge to engage in physical exercise. Likewise, the philosophical theory of probabilistic causality would not attempt to distinguish between the claims "tall flag poles cause long shadows" and "long shadows cause tall flag poles" – where, for all practical purposes, the putative cause and effect occur simultaneously.

We have seen in Chapter 2 that some determination of causal directionality can be made from atemporal statistical information, if fortified with the assumptions of minimality or stability. These assumptions, however, implicitly reflect generic properties of physical processes – invariance and autonomy (see Section 2.9.1) – that constitute the basis of the structural approach to causality.

7.5.2 The Perils of Circularity

Despite the reliance on temporal precedence, the criteria that philosophers have devised for identifying causal relations suffer from glaring circularity: In order to determine whether an event C is a cause of event E, one must know in advance how other factors are causally related to C and E. Such circularity emerges from the need to define the "background context" under which a causal relation is evaluated, since the intuitive idea that causes should increase the probability of their effects must be qualified by the condition that other things are assumed equal. For example, "studying arithmetic" increases the probability of passing a science test, but only if we keep student age constant; otherwise, studying arithmetic may actually lower the probability of passing the test because it is indicative of young age. Thus, it seems natural to offer the following.

Definition 7.5.1

An event C is causally relevant *to E if there is at least one condition F in some background context K such that* $P(E \mid C, F) > P(E \mid \neg C, F)$.[30]

But what kind of conditions should we include in the background context? On the one hand, insisting on a complete description of the physical environment would reduce probabilistic causality to deterministic physics (barring quantum-level considerations). On the other hand, ignoring background factors altogether – or describing them too coarsely – would introduce spurious correlations and other confounding effects. A natural compromise is to require that the background context itself be "causally relevant" to the variables

[30] The reader can interpret K to be a set of variables and F a particular truth-value assignment to those variables.

in question, but this very move is the source of circularity in the definition of probabilistic causality.

The problem of choosing an appropriate set of background factors is similar to the problem of finding an appropriate adjustment for confounding, as discussed in several previous chapters in connection with Simpson's paradox (e.g., Sections 3.3, 5.1.3, and 6.1). We have seen (e.g., in Section 6.1) that the criterion for choosing an appropriate set of covariates for adjustment cannot be based on probabilistic relationships alone but must rely on causal information. In particular, we must make sure that factors listed as background are not affected by C; otherwise, no C would ever qualify as a cause of E, because we can always find factors F that are intermediaries between C and E and that screen off E from C.[31] Here we see the emergence of circularity: In order to determine the causal role of C relative to E (e.g., the effect of the drug on recovery), we must first determine the causal role of every factor F (e.g., gender) relative to C and E.

Factors affecting both C and E can be rescued from circularity by conditioning on *all* factors preceding C but, unfortunately, other factors that cannot be identified through temporal ordering alone must also be weighed. Consider the betting example used in Section 7.1.2. I must bet heads or tails on the outcome of a fair coin toss; I win if I guess correctly and lose if I don't. Naturally, once the coin is tossed (and while the outcome is still unknown), the bet is deemed causally relevant to winning, even though the probability of winning is the same whether I bet heads or tails. In order to reveal the causal relevance of the bet (C), we must include the outcome of the coin (F) in the background context even though F does not meet the common-cause criterion – it does not affect my bet (C) nor is it causally relevant to winning (E) (unless we first declare the bet *is* relevant to winning). Worse yet, we cannot justify including F in the background context by virtue of its occurring earlier than C because whether the coin is tossed before or after my bet is totally irrelevant to the problem at hand. We conclude that temporal precedence alone is insufficient for identifying the background context, and we must refine the definition of the background context to include what Eells (1991) called "interacting causes" – namely, (simplified) factors F that (i) are not affected causally by C and (ii) jointly with C (or $\neg C$) increase the probability of E.

Because of the circularity inherent in all definitions of causal relevance, probabilistic causality cannot be regarded as a program for extracting causal relations from temporal–probabilistic information; rather, it should be viewed as a program for validating whether a proposed set of causal relationships is consistent with the available temporal–probabilistic information. More formally, suppose someone gives us a probability distribution P and a temporal order O on a (complete) set of variables V. Furthermore, any pair of variable sets (say, X and Y) in V is annotated by a symbol R or I, where R stands for "causally relevant" and I for "causally irrelevant." Probabilistic causality deals with testing whether the proposed R and I labels are consistent with the pair $\langle P, O \rangle$ and with the restriction that causes should both precede and increase the probability of their effect.

[31] We say that F "screens off" E from C if C and E are conditionally independent, given both F and $\neg F$.

Currently, the most advanced consistency test is the one based on Eells's (1991) criterion of relevance, which may be translated as follows.

Consistency Test
For each pair of variables labeled $R(X, Y)$, test whether

(i) X precedes Y in O, and

(ii) there exist x, x', y such that $P(y \mid x, z) > P(y \mid x', z)$ for some z in Z, where Z is a set of variables in the background context K such that $I(X, Z)$ and $R(Z, Y)$.

This now raises additional questions.

(a) Is there a consistent label for every pair $\langle P, O \rangle$?

(b) When is the label unique?

(c) Is there a procedure for finding a consistent label when it exists?

Although some insights into these questions are provided by graphical methods (Pearl 1996), the point is that, owing to circularity, the mission of probabilistic causality has been altered: from discovery to consistency testing.

It should also be remarked that the basic program of defining causality in terms of conditionalization, even if it turns out to be successful, is at odds with the natural conception of causation as an oracle for interventions. This program first confounds the causal relation $P(E \mid do(C))$ with epistemic conditionalization $P(E \mid C)$ and then removes spurious correlations through steps of remedial conditionalization, yielding $P(E \mid C, F)$. The structural account, in contrast, defines causation directly in terms of Nature's invariants (i.e., submodel M_x in Definition 7.1.2); see the discussion following Theorem 3.2.2.

7.5.3 The Closed-World Assumption

By far the most critical and least defensible paradigm underlying probabilistic causality rests on the assumption that one is in the possession of a probability function on all variables relevant to a given domain. This assumption absolves the analyst from worrying about unmeasured spurious causes that might (physically) affect several variables in the analysis and still remain obscure to the analyst. It is well known that the presence of such "confounders" may reverse or negate any causal conclusion that might be drawn from probabilities. For example, observers might conclude that "bad air" is the cause of malaria if they are not aware of the role of mosquitoes, or that falling barometers are the cause of rain, or that speeding to work is the cause of being late to work, and so on. Because they are unmeasured (or even unsuspected), the confounding factors in such examples cannot be neutralized by conditioning or by "holding them fixed." Thus, taking seriously Hume's program of extracting causal information from raw data entails coping with the problem that the validity of any such information is predicated on the untestable assumption that all relevant factors have been accounted for.

This raises the question of how people ever acquire causal information from the environment and, more specifically, how children extract causal information from experience.

The proponents of probabilistic causality who attempt to explain this phenomenon through statistical theories of learning cannot ignore the fact that the child never operates in a closed, isolated environment. Unnoticed external conditions govern the operation of every learning environment, and these conditions often have the potential to confound cause and effect in unexpected and clandestine ways.

Fortunately, that children do not grow in closed, sterile environments does have its advantages. Aside from passive observations, a child possesses two valuable sources of causal information that are not available to the ordinary statistician: manipulative experimentation and linguistic advice. Manipulation subjugates the putative causal event to the sole influence of a known mechanism, thus overruling the influence of uncontrolled factors that might also produce the putative effect. "The beauty of independent manipulation is, of course, that other factors can be kept constant without their being identified" (Cheng 1992). The independence is accomplished by subjecting the object of interest to the whims of one's volition in order to ensure that the manipulation is not influenced by any environmental factor likely to produce the putative effect. Thus, for example, a child can infer that shaking a toy can produce a rattling sound because it is the child's hand, governed solely by the child's volition, that brings about the shaking of the toy and the subsequent rattling sound. The whimsical nature of free manipulation replaces the statistical notion of randomized experimentation and serves to filter sounds produced by the child's actions from those produced by uncontrolled environmental factors.

But manipulative experimentation cannot explain all of the causal knowledge that humans acquire and possess, simply because most variables in our environment are not subject to direct manipulation. The second valuable source of causal knowledge is linguistic advice: explicit causal sentences about the workings of things which we obtain from parents, friends, teachers, and books and which encode the manipulative experience of past generations. As obvious and uninteresting as this source of causal information may appear, it probably accounts for the bulk of our causal knowledge, and understanding how this transference of knowledge works is far from trivial. In order to comprehend and absorb causal sentences such as "The glass broke because you pushed it," the child must already possess a causal schema within which such inputs make sense. To further infer that pushing the glass will make someone angry at you and not at your brother, even though he was responsible for all previous breakage, requires a truly sophisticated inferential machinery. In most children, this machinery is probably innate.

Note, however, that linguistic input is by and large qualitative; we rarely hear parents explaining to children that placing the glass at the edge of the table increases the probability of breakage by a factor of 2.85. The probabilistic approach to causality embeds such qualitative input in an artificial numerical frame, whereas the structural approach to causality (Section 7.1) builds directly on the qualitative knowledge that we obtain and transmit linguistically.

7.5.4 Singular versus General Causes

In Section 7.2.3 we saw that the distinction between general causes (e.g., "Drinking hemlock causes death") and singular causes (e.g., "Socrates' drinking hemlock caused his death") plays an important role in understanding the nature of explanations. We have

also remarked that the notion of singular causation (also known as "token" or "single-event" causation) has not reached an adequate state of conceptualization or formalization in the probabilistic account of causation. In this section we elaborate the nature of these difficulties and conclude that they stem from basic deficiencies in the probabilistic account.

In Chapter 1 (Figure 1.6) we demonstrated that the evaluation of singular causal claims requires knowledge in the form of counterfactual or functional relationships and that such knowledge cannot be extracted from bare statistical data even when obtained under controlled experimentation. This limitation was attributed in Section 7.2.2 to the temporal persistence (or invariance) of information that is needed to sustain counterfactual statements – persistence that is washed out (by averaging) in statistical statements even when enriched with temporal and causally relevant information. The manifestations of this basic limitation have taken an interesting slant in the literature of probabilistic causation and have led to intensive debates regarding the relationships between singular and generic statements (see e.g. Good 1961; Cartwright 1989; Eells 1991; Hausman 1998).

According to one of the basic tenets of probabilistic causality, a cause should raise the probability of the effect. It is often the case, however, that we judge an event x to be the cause of y when the conditional probability $P(y \mid x)$ is *lower* than $P(y \mid x')$. For example, a vaccine (x) usually decreases the probability of the disease (y) and yet we often say (and can medically verify) that the vaccine itself caused the disease in a given person u. Such reversals would not be problematic to students of structural models, who can interpret the singular statement as saying that "had person u not taken the vaccine (x') then u would still be healthy (y')." The probability of this counterfactual statement $P(Y_{x'} = y' \mid x, y)$ can be high while the conditional probability $P(y \mid x)$ is low, with both probabilities evaluated formally from the same structural model (Section 9.2 provides precise relationships between the two quantities). However, this reversal is traumatic to students of probabilistic causation, who mistrust counterfactuals for various reasons – partly because counterfactuals carry an aura of determinism (Kvart 1986, pp. 256–63) and partly because counterfactuals are perceived as resting on shaky formal foundation "for which we have only the beginnings of a semantics (via the device of measures over possible worlds)" (Cartwright 1983, p. 34).

In order to reconcile the notion of probability increase with that of singular causation, probabilists claim that, if we look hard enough at any given scenario in which x is judged to be a cause of y, then we will always be able to find a subpopulation $Z = z$ in which x raises the probability of y – namely,

$$P(y \mid x, z) > P(y \mid x', z). \tag{7.47}$$

In the vaccine example, we might identify the desired subpopulation as consisting of individuals who are adversely susceptible to the vaccine; by definition, the vaccine would no doubt raise the probability of the disease in that subpopulation. Oddly, only few philosophers have noticed that factors such as being "adversely susceptible" are defined counterfactually and that, in permitting conditionalization on such factors, one opens a clandestine back door for sneaking determinism and counterfactual information back into the analysis.

Perhaps a less obvious appearance of counterfactuals surfaces in Hesslow's example of the birth-control pill (Hesslow 1976), discussed in Section 4.5.1. Suppose we find that

Mrs. Jones is not pregnant and ask whether taking a birth-control pill was the cause of her suffering from thrombosis. The population of pregnant women turns out to be too coarse for answering this question unequivocally. If Mrs. Jones belongs to the class of women who would have become pregnant *but for* the pill, then the pill might actually have lowered the probability of thrombosis in her case by preventing her pregnancy. If, on the other hand, she belongs to the class of women who would *not* have become pregnant regardless of the pill, then her taking the pill has surely increased the chance of thrombosis. This example is illuminating because the two classes of test populations do not have established names in the English language (unlike "susceptibility" of the vaccine example) and must be defined explicitly in counterfactual vocabulary. Whether a woman belongs to the former or latter class depends on many social and circumstantial contingencies, which are usually unknown and are not likely to define an invariant attribute of a given person. Still, we recognize the need to consider the two classes separately in evaluating whether the pill was the cause of Mrs. Jones's thrombosis.

Thus we see that there is no escape from counterfactuals when we deal with token-level causation. Probabilists' insistence on counterfactual-free syntax in defining token causal claims has led to subpopulations delineated by none other but counterfactual expressions: "adversely susceptible" in the vaccine example and "would not have become pregnant" in the case of Mrs. Jones.[32]

Probabilists can argue, of course, that there is no need to refine the subclasses $Z = z$ down to deterministic extremes, since one can stop the refinement as soon as one finds a subclass that increases the probability of y, as required in (7.47). This argument borders on the tautological, unless it is accompanied with formal procedures for identifying the test subpopulation $Z = z$ and for computing the quantities in (7.47) from some reasonable model of human knowledge, however hypothetical. Unfortunately, the probabilistic causality literature is silent on questions of procedures and representation.[33]

In particular, probabilists face a tough dilemma in explaining how people search for that rescuing subpopulation z so swiftly and consistently and how the majority of people end up with the same answer when asked whether it was x that caused y. For example (due to Debra Rosen, quoted in Suppes 1970), a tree limb (x) that fortuitously deflects a golf ball is immediately and consistently perceived as "the cause" for the ball finally ending up in the hole, though such collisions generally lower one's chances of reaching the hole (y). Clearly, if there is a subpopulation z that satisfies (7.47) in such examples (and I doubt it ever enters anyone's mind), it must have at least two features.

(1) It must contain events that occur both before and after x. For example, both the angle at which the ball hit the limb and the texture of the grass on which the ball bounced after hitting the limb should be part of z.

[32] Cartwright (1989, chap. 3) recognized the insufficiency of observable partitions (e.g. pregnancy) for sustaining the thesis of increased probability, but she did not emphasize the inevitable counterfactual nature of the finer partitions that sustain that thesis. Not incidentally, Cartwright was a strong advocate of excluding counterfactuals from causal analysis (Cartwright 1983, pp. 34–5).

[33] Even Eells (1991, chap. 6) and Shafer (1996a), who endeavored to uncover discriminating patterns of increasing probabilities in the actual trajectory of the world leading to y, did not specify what information is needed either to select the appropriate trajectory or to compute the probabilities associated with a given trajectory.

(2) It must depend on x and y. For, surely, a different conditioning set z' would be necessary in (7.47) if we were to test whether the limb caused an alternative consequence y' – say, that the ball stopped two yards short of the hole.

And this brings us to a major methodological inconsistency in the probabilistic approach to causation: If ignorance of x and y leads to the wrong z and if awareness of x and y leads to the correct selection of z, then there must be some process by which people incorporate the occurrence of x and y into their awareness. What could that process be? According to the norms of probabilistic epistemology, evidence is incorporated into one's corpus of knowledge by means of conditionalization. How, then, can we justify excluding from z the very evidence that led to its selection – namely, the occurrence of x and y?

Inspection of (7.47) shows that the exclusion of x and y from z is compelled on syntactic grounds, since it would render $P(y \mid x', z)$ undefined and make $P(y \mid x, z) = 1$. Indeed, in the syntax of probability calculus we cannot ask what the probability of event y *would* be, given that y has in fact occurred – the answer is (trivially) 1. The best we can do is detach ourselves momentarily from the actual world, pretend that we are ignorant of the occurrence of y, and ask for the probability of y under such a state of ignorance. This corresponds precisely to the three steps (abduction, action, and prediction) that govern the evaluation of $P(Y_{x'} = y' \mid x, y)$ (see Theorem 7.1.7), which attains a high value (in our example) and correctly qualifies the tree limb (x) as the cause of making the hole (y). As we see, the desired quantity *can* be expressed and evaluated by ordinary conditionalization on x and y, without explicitly invoking any subpopulation z.[34]

Ironically, by denying counterfactual conditionals, probabilists deprived themselves of using standard conditionals – the very conditionals they were trying to preserve – and were forced to accommodate simple evidential information in roundabout ways. This syntactic barrier that probabilists erected around causation has created an artificial tension between singular and generic causes, but the tension disappears in the structural account. In Section 10.1.1 we show that, by accommodating both standard and counterfactual conditionals (i.e. Y_x), singular and generic causes no longer stand in need of separate analyses. The two types of causes differ merely in the level of scenario-specific information that is brought to bear on a problem, that is, in the specificity of the evidence e that enters the quantity $P(Y_x = y \mid e)$.

7.5.5 Summary

Cartwright (1983, p. 34) listed several reasons for pursuing the probabilistic versus the counterfactual approach to causation:

> [the counterfactual approach] requires us to evaluate the probability of counterfactuals for which we have only the beginnings of a semantics (via the device of measures over possible worlds) and no methodology, much less an account of why the methodology is suited to the semantics. How do we test claims about probabilities of counterfactuals? We have no answer, much less an answer that fits with our nascent semantics. It would be preferable to have a measure of effectiveness that requires only probabilities over events that can be tested in the actual world in the standard ways.

[34] The desired subpopulation z is equal to the set of all u that are mapped into $X(u) = x$, $Y(u) = y$, and $Y_{x'}(u) = y'$.

Examining the progress of the probabilistic approach in the past two decades, it seems clear that Cartwright's aspirations have materialized not in the framework she advocated but rather in the competing framework of counterfactuals, as embodied in structural models. Full characterization of "effectiveness" ("causal effects" in our vocabulary) in terms of "events that can be tested" emerged from Simon's (1953) and Strotz and Wold's (1960) conception of modifiable structural models and led to the back-door criterion (Theorem 3.3.2) and to the more general Theorem 4.3.1, of which the probabilistic criteria (as in (3.13)) are but crude special cases. The interpretation of singular causation in terms of the counterfactual probability $P(Y_{x'} \neq y \mid x, y)$ has enlisted the support of meaningful formal semantics (Section 7.1) and effective evaluation methodology (Theorem 7.1.7 and Sections 7.1.3–7.2.1), while the probabilistic criterion of (7.47) lingers in vagueness and procedureless debates. The original dream of rendering causal claims testable was given up in the probabilistic framework as soon as unmeasured entities (e.g., state of the world, background context, causal relevance, susceptibility) were allowed to infiltrate the analysis, and methodologies for answering questions of testability have moved over to the structural–counterfactual framework (see Chapter 9).

The ideal of remaining compatible with the teachings of nondeterministic physics seems to be the only viable aspect remaining in the program of probabilistic causation, and this section questions whether maintaining this ideal justifies the sacrifices. It further suggests that the basic agenda of the probabilistic causality program is due for a serious reassessment. If the program is an exercise in epistemology, then the word "probabilistic" is oxymoronic – human perception of causality has remained quasi-deterministic, and these fallible humans are still the main consumers of causal talk. If the program is an exercise in modern physics, then the word "causality" is nonessential – quantum-level causality follows its own rules and intuitions, and another name (perhaps "qua-sality") might be more befitting. However, regarding artificial intelligence and cognitive science, I would venture to predict that robots programmed to emulate the quasi-deterministic macroscopic approximations of Laplace and Einstein would far outperform those built on the correct but counterintuitive theories of Born, Heisenberg, and Bohr.

Acknowledgment

Sections of this chapter are based on the doctoral research of Alex Balke and David Galles. This research has benefitted significantly from the input of Joseph Halpern.