# An Axiomatic Characterization of Causal Counterfactuals

**David Galles and Judea Pearl**
Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles, CA 90024
*galles@cs.ucla.edu* *judea@cs.ucla.edu*

**Abstract**

This paper studies the causal interpretation of counterfactual sentences using a modifiable structural equation model. It is shown that two properties of counterfactuals, namely, composition and effectiveness, are sound and complete relative to this interpretation, when recursive (i.e., feedback-less) models are considered. Composition and effectiveness also hold in Lewis's closest-world semantics, which implies that for recursive models the causal interpretation imposes no restrictions beyond those embodied in Lewis's framework. A third property, called reversibility, holds in nonrecursive causal models but not in Lewis's closest-world semantics, which implies that Lewis's axioms do not capture some properties of systems with feedback. Causal inferences based on counterfactual analysis are exemplified and compared to those based on graphical models.

Keywords: Causality, counterfactuals, interventions, structural equations, policy analysis, graphical models.

## 1 Introduction

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, namely to 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 changes occurring in the world are distinguished from changes of 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.

Lewis (1973b) has proposed a logic of counterfactuals based on the notion of *closest worlds*: A sentence of the form "If $A$ were the case, then $B$ would be the case" is true in a world $w$ just in case $B$ is true in the closest world to $w$ in which $A$ is true. This framework presupposes the existence of a measure of distance between worlds that can be used to identify the closest $A$-world to $w$, for any world $w$ and any sentence $A$ in the language of discourse. Lewis is careful to keep his formalism as general as possible, and, save for the obvious requirement that every world be closest to itself, he does not impose any specific structure on the distance measure. However, the fact that people communicate with counterfactuals suggests that they share a distance measure that is encoded parsimoniously in the mind. What mental representation is used for encoding those inter-world distances?

Lewis himself provides a clue, the closest worlds that he envisions are *causal* in nature. For instance, when Lewis considers as an example a hypothetical world in which kangaroos have no tails, he argues that not just the state of the tail, but also the tracks that the animal made, the animal's balance, and a variety of other factors would also be different. Thus, Lewis appeals to our common knowledge of cause and effect in laying out which factors are expected to change in the hypothetical world, and which factors are expected to be unaltered.

If our assessment of inter-world distances comes from causal knowledge, the question arises 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 closest 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 (1981) characterize decision-making conditionals, namely, sentences of the form "If we do $A$, then $B$," using Lewis's general framework, while Pearl (1994, 1995) constructs a calculus of action based directly on causal semantics, and whether the two formalisms are identical is uncertain.[1]

Another application occurs in statistics. Rubin (1974), Holland (1986), and Robins (1986) all used counterfactual variables to analyze the effectiveness of treatments in clinical studies. Starting with the primitive notion of potential response $Y(x, u)$ (read: the value that the response variable $Y$ would have attained in patient $u$ had the treatment been $x$), which is treated as a random variable, they ask under what conditions one can infer the average treatment effect $E_u[Y(x, u)]$ from clinical data (Typical conditions require that the treatment assignment be randomized or semi-randomized.) Although the logic underlying this type of analysis has not been stated formally, statisticians use the closest-world framework as a guiding paradigm and have adopted certain rules of inference that plausibly follow from this framework. For example, among their most commonly employed rules is the implication (called *consistency* [Robins, 1987]):

$$X = x \Longrightarrow Y(x, u) = Y(u) \tag{1}$$

which states that the potential response of patient $u$ to a hypothetical treatment $x$, $Y(x, u)$, must coincide with the patient's observed response, $Y(u)$, whenever the actual treatment $X$ happened to be $x$. This rule, as we shall see, is a special case of an axiom of counterfactuals called composition (see Eq. (17)), an axiom that follows from the requirement that the actual world be closer to itself than any world that differs from the actual world.

[1] Winslett (1988) and Ginsberg and Smith (1987) have also advanced theories of actions based on closest-world semantics, while Katsuno and Mendelzon (1991) have used this semantics to characterize belief updating. None of these assumes a special structure for the distance measure, to reflect causal considerations.

The question remains, however, whether inference rules beyond Lewis's axioms are necessary for statisticians to fully and accurately capture the causal structure of the clinical test environment and the causal character of the counterfactuals considered in such an environment. This paper analyzes this question for both recursive and nonrecursive causal models, namely, models of systems without and with feedback. We first show that Lewis's axioms, together with the assumption of recursiveness, are sound and complete with respect to the causal interpretation of counterfactual, that is, causality per se imposes no restrictions beyond those embodied in the closest-world framework together with recursiveness. When we consider nonrecursive systems, however, Lewis's axioms are not complete. We show that a property called *reversibility* holds in nonrecursive causal models yet it does not follow from Lewis's axioms. Thus, Lewis's framework misses some properties of causality in the general case of feedback systems in equilibrium.

Section 2 gives a brief overview of causal models employing modifiable structural equations and illustrates their use in the interpretation of causal and counterfactual utterances. Section 3 defines the properties of composition, effectiveness, and reversibility, and shows that composition and effectiveness are sound and complete for recursive causal models, where reversibility holds trivially. Section 4 compares causal models to Lewis's framework, and finds that composition and effectiveness are sound in that formalism as well. Section 5 illustrates the derivation of probabilistic answers to counterfactual queries using only composition and effectiveness as rules of inferences. Section 6 concludes with remarks on the role of counterfactual calculus vis a vis structural equations and graphs.

# 2 Causal Models

## 2.1 Definitions

A causal model is a mathematical object that provides an interpretation (and effective computation) of every causal query about the domain. Following [Pearl, 1995a], we adopt here a construct named *modifiable structural equations*, that generalizes most causal models used in engineering, biology, and economics.

**Definition 1** *(causal model) A* causal model *is a triple*

$$M = \ < U, V, F >$$

*where*

**(i)** *$U$ is a set of variables, called* exogenous*, 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.*

**(iii)** *$F$ is a set of functions $\{f_1, f_2, \ldots, f_n\}$ where each $f_i$ is a mapping from $U \cup (V \setminus V_i)$ to $V_i$ such that $F$ defines a mapping from $U$ to $V$. (i.e., $F$ has a unique solution for each state $u$ in the domain of $U$). Symbolically, $F$ can be represented by writing*

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

*where $pa_i$ is any realization of the (unique) set of variables $PA_i$ in $V/V_i$ (connoting parents) that renders $f_i$ nontrivial.*

Every causal model $M$ can be associated with a directed graph, $G(M)$, in which each node corresponds to a variable in $V$ and the directed edges point from members of $PA_i$ toward $V_i$. We call such a graph, the *causal graph* associated with $M$. This graph merely identifies the endogenous variables $PA_i$ that have direct influence on each $V_i$ but it does not specify the functional form of $f_i$.

**Definition 2** *(submodel) Let $M$ be a causal model, $X$ be a set of variables in $V$, and $x$ be a particular realization of $X$. A* submodel *$M_x$ of $M$ is the causal model*

$$M_x = \quad < U, V, F_x >$$

*where*

$$F_x = \{f_i : V_i \notin X\} \cup \{X = x\} \tag{2}$$

In words, $F_x$ is formed by deleting from $F$ all functions $f_i$ corresponding to members of $X$ and replacing them with the set of functions $X = x$. Implicit in the definition of submodels is the assumption that $F_x$ possesses a unique solution for every $u$.

Submodels are useful for representing the effect of local actions and changes. 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 choosing the name *modifiable structural equations.*

**Definition 3** *(effect of action) Let $M$ be a causal model, $X$ be a set of variables in $V$, and $x$ be a particular realization of $X$. The* effect of action *$do(X = x)$ on $M$ is given by the submodel $M_x$.*[2]

**Definition 4** *(potential response) Let $Y$ be a variable in $V$, and let $X$ be a subset of $V$. The* potential response *of $Y$ to action $do(X = x)$, denoted $Y_x(u)$, is the solution for $Y$ of the set of equations $F_x$.*

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, rather than their deletion [Pearl, 1994]. We will not consider disjunctive actions, of the form "$do(X = x$ or $X = x')$", as these complicate the probabilistic treatment of counterfactuals.

[2]Readers that are disturbed by the impracticality of some local actions (e.g., creating a world where kangaroos have no tails) are invited to replace the word "action" with the word "modification" (see [Leamer, 1985]). The advantages of using hypothetical external interventions to convey the notion of "local change" are emphasized in [Pearl, 1995a, p. 706].

**Definition 5** *(counterfactual) Let $Y$ be a variable in $V$, and let $X$ a subset of $V$. The counterfactual sentence "The value that $Y$ would have obtained, had $X$ been $x$" is interpreted as denoting the potential response $Y_x(u)$.*[3]

Two special cases are worth noting. First, if $Y = V_i$ and $X = V \setminus Y$, then $Y_x(u) = f_i(pa_i, u)$ where $pa_i$ is the projection of $X = x$ on $PA_i$. Thus, each function $f_i$ in $M$ may be given a counterfactual interpretation; it specifies the potential response of $V_i$ to a hypothetical manipulation of all other variables in $V$. Second, if $Y$ is included in $X$ and $X = x \Longrightarrow Y = y$, then $Y_x(u) = y$. This means that the potential response of a manipulated variable coincides with the values set by the manipulation.

The formulation above shares many features with that of Simon and Rescher (1966). Both are based on an assembly of stable physical mechanisms, represented as a set of equations, and both assume a one-to-one correspondence between equations and variables. Simon and Rescher, however, do not treat counterfactual antecedents as actions and, therefore, they encounter difficulties handling counterfactuals whose antecedents involve endogenous variables. Our formulation overcomes these difficulties by explicitly representing actions and counterfactuals in terms of equation-deletion operators[4] and, furthermore, our formulation generalizes naturally to probabilistic systems, as is seen below.

**Definition 6** *(probabilistic causal model) A* probabilistic causal model *is a pair*

$$< M, P(u) >$$

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

$P(u)$, 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

$$P(y) = \sum_{\{u \ | \ Y(u)=y\}} P(u) \tag{3}$$

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

$$P(Y_x = y) = \sum_{\{u \ | \ Y_x(u)=y\}} P(u) \tag{4}$$

Likewise a causal model defines a joint distribution on all counterfactual statements, i.e., $P(Y_x = y, Z_w = z)$ is defined for any sets of variables $Y, X, Z, W$, not necessarily disjoint. 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

$$P(Y_x = y, X = x') = \sum_{\{u | Y_x(u)=y \ \& \ X(u)=x'\}} P(u) \tag{5}$$

[3] The connection between counterfactuals and local actions is made in [Lewis, 1973a] and is further elaborated in [Balke and Pearl, 1994] and [Heckerman and Shachter, 1995].

[4] 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) and Sobel (1990).

and

$$P(Y_x = y, Y_{x'} = y') = \sum_{\{u \mid Y_x(u)=y \ \& \ Y_{x'}(u)=y'\}} P(u). \tag{6}$$

When $x$ and $x'$ are incompatible, $Y_x$ 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 objections to treating counterfactuals as jointly distributed random variables [Dawid, 1997]. The definition of $Y_x$ in terms of submodels diffuses such objections and further illustrates that joint probabilities of counterfactuals can be encoded rather parsimoniously using $P(u)$ and $F$.

## 2.2 Examples

Next we demonstrate the generality of the modifiable structural equation model using two familiar applications: evidential reasoning and policy analysis. Additional applications involving the formalization of causal relevance and the interpretation of causal utterances can be found in [Galles and Pearl, 1997b].

### 2.2.1 Sprinkler Example

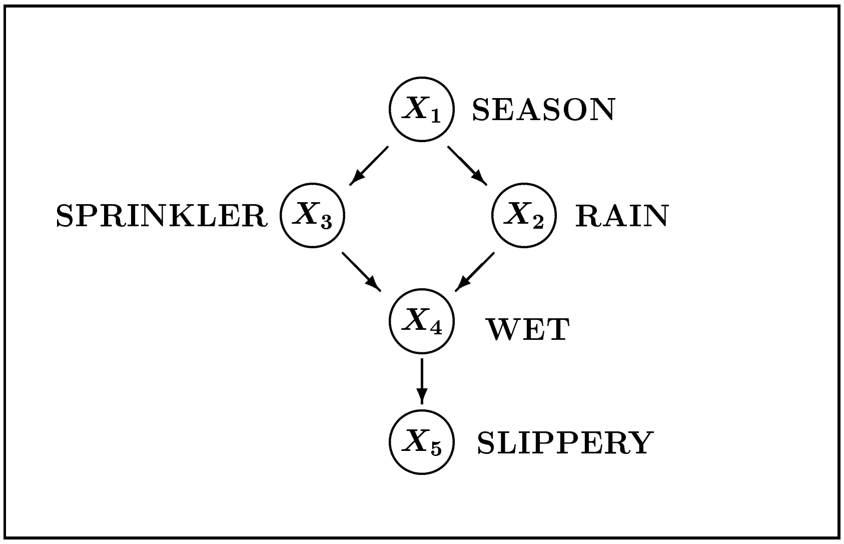


Figure 1: Causal graph illustrating causal relationships among five variables.

Figure 1 is a simple yet typical causal graph used in common sense reasoning. It describes the causal relationships among the season of the year ($X_1$), whether rain falls ($X_2$) during the season, whether the sprinkler is on ($X_3$) during the season, whether the pavement is wet ($X_4$), and whether the pavement is slippery ($X_5$). All variables in this graph except the root variable $X_1$ take a value of either "True" or "False." $X_1$ takes one of four values: "Spring," "Summer," "Fall," or "Winter." Here, the absence of a direct link between, for example, $X_1$ and $X_5$, captures our understanding that the influence of the season on the slipperiness of the pavement is mediated by other conditions (e.g., the wetness of the pavement). The corresponding model consists of five functions, each representing an autonomous mechanism:

$$\begin{aligned} x_1 &= u_1 \\ x_2 &= f_2(x_1, u_2) \\ x_3 &= f_3(x_1, u_3) \\ x_4 &= f_4(x_3, x_2, u_4) \\ x_5 &= f_5(x_4, u_5) \end{aligned} \tag{7}$$

The disturbances $U_1, \ldots, U_5$ are not shown explicitly in Figure 1 but are understood to govern the uncertainties associated with the causal relationships. The causal graph coincides with the Bayesian network [Pearl, 1988] associated with $P(x_1, \ldots, x_5)$ whenever the disturbances are assumed to be independent, $U_i \perp\!\!\!\perp U \setminus U_i$. When some disturbances are judged to be dependent, it is customary to encode such dependencies by augmenting the graph with double-headed arrows, as shown in Figure 3, Section 5.

A typical specification of the functions $\{f_1, \ldots, f_5\}$ and the disturbance terms is given by the Boolean model below:

$$\begin{aligned} x_2 &= [(X_1 = \text{Winter}) \vee (X_1 = \text{Fall}) \vee ab_2] \wedge \neg ab'_2 \\ x_3 &= [(X_1 = \text{Summer}) \vee (X_1 = \text{Spring}) \vee ab_3] \wedge \neg ab'_3 \\ x_4 &= (x_2 \vee x_3 \vee ab_4) \wedge \neg ab'_4 \\ x_5 &= (x_4 \vee ab_5) \wedge \neg ab'_5 \end{aligned} \tag{8}$$

where $x_i$ stands for $X_i = true$, and $ab_i$ and $ab'_i$ stand, respectively, for triggering and inhibiting abnormalities. For example, $ab_4$ stands for (unspecified) events that might cause the pavement to get wet ($x_4$) when the sprinkler is off ($\neg x_2$) and it does not rain ($\neg x_3$) (e.g., pouring a pail of water on the pavement), while $\neg ab'_4$ stands for (unspecified) events that will keep the pavement dry in spite of the rain ($x_3$), the sprinkler ($x_2$), and $ab_4$ (e.g., covering the pavement with a plastic sheet).

To represent the action "turning the sprinkler ON," or $do(X_3 = \text{ON})$, we replace the equation $x_3 = f_3(x_1, u_3)$ in the model of Eq. (7) with the equation $x_3 = \text{ON}$. The resulting submodel, $M_{X_3=\text{ON}}$, contains all the information needed for computing the effect of the action on the other variables. Thus, the operation $do(X_3 = \text{ON})$ stands in marked contrast to that of *finding* the sprinkler ON; the latter involves making the substitution *without* removing the equation for $X_3$, and therefore may potentially influence (the belief in) every variable in the network. In contrast, the only variables affected by the action $do(X_3 = \text{ON})$ are $X_4$ and $X_5$, that is, the descendants of the manipulated variable $X_3$. This mirrors indeed the difference between *seeing* and *doing*: after observing that the sprinkler is ON, we wish to infer that the season is dry, that it probably did not rain, and so on; no such inferences should be drawn in evaluating the effects of the contemplated action "turning the sprinkler ON."

This distinction obtains a vivid symbolic representation in cases where the $U_i$'s are assumed independent, because the joint distribution of the endogenous variables then admits the product decomposition:

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2, x_3)P(x_5|x_4) \tag{9}$$

Similarly, the joint distribution associated with the submodel $M_x$ representing the action $do(X_3 = \text{ON})$ is obtained from the product above by deleting the factor $P(x_3|x_1)$ and substituting $X_3 = \text{ON}$:

$$P(x_1, x_2, x_4, x_5|do(X_3 = \text{ON})) = P(x_1)\ P(x_2|x_1)\ P(x_4|x_2, X_3 = \text{ON})\ P(x_5|x_4) \tag{10}$$

The difference between the action $do(X_3 = \text{ON})$ and the observation $X_3 = \text{ON}$ is thus seen from the corresponding distributions. The former is represented by Eq. (10), while the latter by *conditioning* Eq. (9) on the observation; i.e.,

$$P(x_1, x_2, x_4, x_5|X_3 = \text{ON}) = \frac{P(x_1)\ P(x_2|x_1)\ P(x_3 = \text{ON}|x_1)P(x_4|x_2, X_3 = \text{ON})P(x_5|x_4)}{P(X_3 = \text{ON})}$$

Note that the conditional probabilities on the r.h.s. of Eq. (10) are the same as those in Eq. (9), and can therefore be estimated from pre-action observations. However, the pre-action distribution $P$ is not sufficient for evaluating *conditional* counterfactuals whenever the conditions given are affected by the counterfactual antecedent. For example, the probability that "the pavement would *continue* to be slippery once we turn the sprinkler off," tacitly presuming that currently the pavement *is* slippery, cannot be evaluated from the conditional probabilities $P(x_i|pa_i)$ alone; the functional forms of the $f_i$'s (Eq. 7) are necessary for evaluating such queries [Balke and Pearl 1994; Pearl 1996].

### 2.2.2 Policy Analysis in Linear Econometric Models

Causal models are often used to predict the behavior of systems in dynamic equilibrium. In the economic literature, for example, we find the system of equations

$$q = b_1 p + d_1 i + u_1 \tag{11}$$
$$p = b_2 q + d_2 w + u_2 \tag{12}$$

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, namely, unmodeled factors that affect quantity and price, respectively [Goldberger, 1992].

This system of equations constitutes a causal model (Definition 1) if we define $V = \{Q, P\}$, $U = \{U_1, U_2, I, W\}$ and assume that each equation represents an autonomous process in the sense of Definition 3. The causal graph of this model is shown in Figure 2. It is normally assumed that $I$ and $W$ are known, 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, the model must be augmented with the distribution of these errors, which is usually taken to be a Gaussian distribution with the covariance matrix $\Sigma_{ij} = cov(u_i, u_j)$.

We can use this model to answer queries such as:

1. Find the expected value of the demand $(Q)$ if the price is controlled at $P = p_0$.

2. Find the expected value of the demand $(Q)$ if the price is reported to be $P = p_0$.

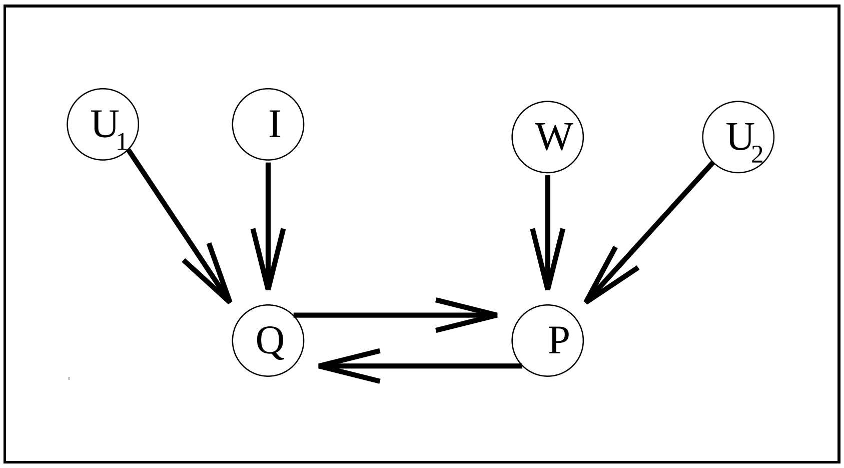


Figure 2: Causal graph illustrating the relationship between supply and demand

3. Given that the current price is $P = p_0$, find the expected value of the demand ($Q$) had the price been controlled at $P = p_1$.

To find the answer to the first query, we replace Eq. (12) with $p = p_0$, leaving

$$q = b_1 p + d_1 i + u_1 \tag{13}$$
$$p = p_0 \tag{14}$$

The demand is then $q = b_1 p_0 + d_1 i + u_1$, and the expected value of $Q$ can be obtained from $i$ and the expectation of $U_1$, giving

$$E[Q|do(P = p_0)] = E[Q] + b_1(p - E[P]) + d_1(i - E[I]).$$

The answer to the second query is given by conditioning Eq. (11) on the current observation $\{P = p_0, I = i, W = w\}$ and taking the expectation,

$$E[Q|p_0, i, w] = b_i p_0 + d_1 i + E[U_1|p_0, i, w]. \tag{15}$$

The computation of $E[U_1|p_0, i, w]$ is a standard procedure once $\Sigma_{ij}$ is given [Meditch, 1969]. Note that, although $U_1$ was assumed independent of $I$ and $W$, this independence no longer holds once $P = p_0$ is observed. Note also that Eqs. (11) and (12) both participate in the solution and that the observed value $p_0$ will affect the expected demand $Q$ (through $E[U_1|p_0, i, w]$) even when $b_1 = 0$, which is not the case in query 1.

The third query requires the conditional expectation of the counterfactual quantity $Q_{p=p_1}$, given the current observations $\{P = p_0, I = i, W = w\}$, namely,

$$E[Q_{p=p_1}|p_0, i, w] = b_1 p_1 + d_1 i + E[U_1|p_0, i, w] \tag{16}$$

The expected value $E[U_1|p_0, i, w]$ is the same in the solutions to the second and third queries; the latter differs only in the term $b_1 p_1$. A general method for solving such counterfactual queries is described in [Balke and Pearl, 1995].[5]

[5]Readers concerned with teaching of policy analysis would be interested to note that the second author has

# 3 Composition, Effectiveness, and Reversibility

We now present three properties of counterfactuals — composition, effectiveness, and reversibility — which hold in all causal models.

**Property 1** *(composition) For any two singleton variables $Y$ and $W$, and any set of variables $X$ in a causal model, we have*

$$W_x(u) = w \Longrightarrow Y_{xw}(u) = Y_x(u) \tag{17}$$

Composition states that if we force a variable ($W$) to a value that it would have had without our intervention, then the intervention will have no effect on other variables in the system.

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** *(null action) $Y_\emptyset(u) \doteq Y(u)$.*

**Corollary 1** (Consistency) *For any variables $Y$ and $X$ in a causal model, we have*

$$X(u) = x \Longrightarrow Y(u) = Y_x(u) \tag{18}$$

**Proof:**
Eq. 18 follows directly from Composition. Substituting $X$ for $W$ and $\emptyset$ for $X$ in Eq. (17), we obtain $X_\emptyset(u) = x \Longrightarrow Y_\emptyset(u) = Y_x(u)$. Null Action allows us to drop the $\emptyset$, leaving $X(u) = x \Longrightarrow Y(u) = Y_x(u)$. □

The implication in Eq. (18) was called *Consistency* by Robins (1987).[6]

**Property 2** *(effectiveness) For all 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 $X$,*

$$(Y_{xw}(u) = y) \;\&\; (W_{xy}(u) = w) \Longrightarrow Y_x(u) = y \tag{19}$$

presented this example to well over a hundred econometrics students and faculty across the US. Respondents had no problem answering question 2, one person was able to solve question 1, and none managed to answer question 3. Pearl (1997a) suggests an explanation.

[6] Consistency and composition are informally used in economics [Manski, 1990] and statistics within the potential-response framework [Rubin, 1974]. To the best of our knowledge, Robins (1987) was the first to state consistency formally and to use it to derive other properties of counterfactuals. Composition was brought to our attention by Jamie Robins (personal communication, February 1995), a weak version of it is mentioned explicitly in [Holland, 1986, p. 968].

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 $(Y_{xw}(u) = y)$ & $(W_x(u) = w) \Longrightarrow Y_x(u) = y$, which is another form of composition, or to $(Y_x(u) = y)$ & $(W_{xy}(u) = w) \Longrightarrow Y_x(u) = y$, which is trivially true. In nonrecursive systems, reversibility is a property of causal loops. If forcing $X$ to a value $x$ results in a value $y$ for $Y$, and forcing $Y$ to the value $y$ results in $X$ achieving the value $x$, then $X$ and $Y$ will have the values $x$ and $y$, respectively, without any intervention.

In nonrecursive systems, the properties of composition, effectiveness, and reversibility are independent – none is a consequence of the other two. This is shown in the Appendix by constructing a truth table for counterfactual statements such that when any two properties hold, the third does not. In recursive systems, reversibility holds trivially, and the independence of composition and effectiveness is easily shown.

## 3.1 Soundness of Composition, Effectiveness, and Reversibility

**Theorem 1** *Composition holds in all causal models.*

**Proof:**
Since $Y_x(u)$ has a unique solution, forming $M_x$ and substituting out all other variables would yield a unique solution for $Y$, regardless of the order of substitution. So we will form $M_x$ and examine the structural equation for $Y$ in $M_x$, $Y_x = f_Y(x, z, w, u)$, where $Z$ stands for the rest of the parent set of $Y$. To solve for $Z$, we substitute out all variables except $X$,$Y$, and $W$. In other words, we substitute out all variables in $M_x$, without substituting into $X$, $W$, and $Y$, and express $Z$ as a function of $x, w$, and $u$. We then plug this solution into $f_Y$ to get $Y_x = f_Y(x, w, Z(x, w, u), u)$, which we can write as $Y_x = f(x, w, u)$. At this point, we can solve for $W$ by substituting out all variables in $M_X$ other than $X$, which leaves $Y_x = f(x, W(u, x), u)$. We can now see that if $w = W_x(u)$, then $Y_x(u) = Y_{xw}(u)$. □

This proof is still valid in cases where $X = \emptyset$.

**Theorem 2** *Effectiveness holds in all causal models.*

**Proof:**
This theorem follows from Definition 5, where $Y_x(u)$ is interpreted as the unique solution for $Y$ of a set of equations under $X = x$. □

**Theorem 3** *Reversibility holds in all causal models.*

**Proof:**
Reversibility follows from the assumption that the solution for $V$ in every submodel is unique. Since $Y_x(u)$ has a unique solution, forming $M_x$ and substituting out all other variables would yield a unique solution for $Y$, regardless of the order of substitution. So, we will form $M_x$ and examine the structural equation for $Y$ in $M_x$, which might in general be a function of $X, W, U$, and additional variables: $Y_x = f_Y(x, w, z, u)$, where $Z$ stands for parents of $Y$ not contained in $X \cup W \cup U$. We now solve for $Z$ by substituting out all variables except $X$, $Y$, and $W$. That is, we substitute out all variables in $M_x$, avoiding substitutions into $X$, $W$ and $Y$, and express $Z$ as a function of $x, w$, and $u$. We then plug this solution into $f_Y$ to get

$Y_x = f_Y(x, w, Z(x, w, u), u)$, which we can write as $Y_x = f(x, w, u)$. We now consider what would happen if we solved for $Y$ in $M_{xw}$. Since we avoided substituting anything into $W$ when we solved for $Y$ in $M_x$, we will get the same result as before, namely, $Y_{xw} = f(x, w, u)$. In the same way, we can show that $W_x = g(x, y, u)$ and $W_{xy} = g(x, y, u)$. So, solving for $y = Y_x(u)$, $w = W_x(u)$ is the same as solving for $y = f(x, w, u)$ and $w = g(x, y, u)$, which is the same as solving for $y = Y_{xw}(u)$, $w = W_{xy}(u)$. Thus, any solution $y$ to $y = Y_{xw}(u), w = W_{xy}(u)$ would also be a solution to $y = Y_x(u)$. □

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 knowing this outcome does not guarantee cooperation from the start ($Y(u) = y, W(u) = w$). Irreversibility, in such systems, is a product of using a state description that is too coarse, one where all of the factors that determine the ultimate state of the system are not included in $U$. In a tit-for-tat system, the state description should include factors such as the previous actions of the players, and reversibility is restored once the missing factors are included.

## 3.2 Completeness of Composition and Effectiveness

**Definition 8** (causal ordering) *A causal ordering $X_1 \ldots X_n$ of a set of variables is an ordering such that for any two variables $X = X_i$ and $Y = X_k$, $i < k$, we have $X_{yz}(u) = X_z(u)$, where $Z$ is any set of variables not including $X$ or $Y$.*

Clearly, for every recursive model we can find an ordering that satisfies the condition of Definition 8. In fact, every ordering consistent with the arrows of the causal graph $G(M)$ will satisfy this condition. A system in which the variables are indexed along a specific causal ordering will be called a *causally ordered system.*

**Theorem 4** *Composition, together with effectiveness, are complete for causally ordered systems, relative to conjunctions of counterfactual statements.*

A formal proof of completeness requires the explication of two properties, definiteness and uniqueness,[7] which are implied by the definition of causal models (Definition 1).

**Property 4** (definiteness) *For any variable $X$ and set of variables $Y$,*

$$\exists x \in X \ s.t. \ \ X_y(u) = x \tag{20}$$

**Property 5** (uniqueness) *For every variable $X$ and set of variables $Y$,*

$$X_y(u) = x \;\&\; X_y(u) = x' \Longrightarrow x = x' \tag{21}$$

[7] These two properties, definiteness and uniqueness, were kept implicit in the completeness proof originally reported in [Galles and Pearl, 1997a]; the benefit to explicating them formally was brought to our attention by [Halpern, 1998].

**Definition 9** (statement) *By a counterfactual statement, or* statement *for short, we denote a sentence of the form* $Y_x(u) = y$ *for a specific variable* $Y \in V$*, a specific realization* $x$ *of a set of variables* $X \subseteq V$*, and a specific* $u$ *in the domain of* $U$*.*

**Definition 10** (semantic entailment) *Given a set* $S$ *of counterfactual statements, let* $M_S$ *be the set of models of* $S$*, namely, the set* $\{m_1, \ldots, m_n\}$ *of all causal models such that all statements in* $S$ *hold for each* $m_i$*. A counterfactual statement* $\sigma$ *is* semantically entailed *by* $S$*, written* $S \models \sigma$*, if* $\sigma$ *holds in each* $m_i \in M_S$*.*

**Definition 11** (syntactic entailment) *Given a set* $A$ *of axioms, a set of counterfactual statements* $S$ syntactically entails *a counterfactual statement* $\sigma$*, written* $S \vdash_A \sigma$*, if* $\sigma$ *can be derived from* $S$ *using repeated applications of axioms from* $A$ *together with the rules of logic.*

Denote by $CO$ the set of $n(n-1)/2$ statements $X_{yz}(u) = X_z(u) \forall X, Y \in V$ such that $X$ precedes $Y$ in the causal ordering. Define $A_C$ to be the set {composition, effectiveness, definiteness, uniqueness, $CO$}. We want to show that all statements that are semantically entailed by $S$ are also syntactically entailed by $S$, namely, that

$$S \models \sigma \implies S \vdash_{A_C} \sigma$$

[Note that] the axiom of definiteness require the use of disjunction, which is not part of a simple counterfactual statement as specified in Definition 9. Thus, [by limiting our target sentences to conjunctions of counterfactual statements (Definition 9), the language relative which we need to establish completeness is weaker than the one used for expressing axioms $A_C$.]

To establish completeness, it is enough to show that every set of statements $S$ that is consistent with $A_C$ has a model. To see that this condition is sufficient to prove the completeness of $A_C$, assume that there is some set $S$ and statement $p : X_z(u) = x$ such that in every model consistent with $S$, $p$ holds, and $p$ is not derivable from $S$ using $A_C$. Since $p$ is not derivable from $S$, there must be some other statement $p' : X_z(u) = x', x \neq x'$, such that $S \cup \{p'\}$ is consistent with $A_C$. Since in every model consistent with $S$, $X_z(u) = x$ holds, no model is consistent with $S \cup \{p'\}$. Thus, if $A_C$ is not complete, then there must exist some set $S'$ that is consistent with $A_C$, and has no model. Looking at the contrapositive, if every set of statements $S$ that is consistent with $A_C$ has a model, then $A_C$ is complete.

We now show that for any set of statements $S$, if $S$ is consistent under $A_C$ then $S$ has a model. We will use the concept of a maximally consistent set, which is a standard technique used to prove completeness in modal logic [Fagin *et al.*, 1995]. Consider a maximally consistent set $S^*$. That is, a superset of $S$ that is consistent with $A_C$ such that any superset of $S^*$ is not consistent with $A_C$. We will show that there is a causal model $M$ which satisfies every statement in $S^*$, and thus satisfies every statement in $S$.[8]

**Proof** (by induction)**:** We prove that, for any maximally consistent set $S^*$, there exists a causal model $M$ which satisfies every statement in $S^*$, by induction on the number of variables $|V|$ in $S^*$.

[8]We thank Joseph Halpern for calling our attention to this technique which simplifies appreciably the completeness proof originally reported in [Galles and Pearl, 1997a]. Halpern (1997) further shows that composition and effectiveness are complete in recursive models for which the causal order is not specified and, furthermore, the target language can be extended to disjunctions and negation of counterfactual statements.

Base Case:
If $|V| = 1$, then the statements $X(u)$ in $S^*$ determine the function for $X$, and effectiveness ensures that $X_x(u) = x$ for all $x \in X$.
Inductive Case:
Consider the variables $V$ that are in $S^*$. Let $Y \in V$ be the last element in the causal ordering. Consider the set $S'^*$, which is $S^*$ with all statements of the form $Y_z(u) = y$ and $X_{yz}(u) = x$ removed. By the inductive hypothesis, there is a model $M'$ such that every element of $S'^*$ is satisfied.

We now extend $M'$ to $M$, such that every element in $S^*$ is satisfied in $M$. For each variable $X \in M'$ and each value $y$ of $Y$, $f_{X\,M}(x_1, \ldots, x_k, y, u) = f_{X\,M'}(x_1, \ldots, x_k, u)$. We define $f_Y$ as follows: for each statement $(Y_z(u) = y) \in S^*$ such that $|Z| = |V| - 1$ and $Y \notin Z$, $f_Y(z, u) = y$. Definiteness ensures that $f_Y$ will be completely determined.

Since $M'$ satisfied all elements of $S'^*$, and given the causal ordering such that $X_{yz}(u) = X_z(u)$ for all $X_{yz}(u), X_z(u)$ in $S^*$, $M$ satisfies all statements of the form form $X_z(u)$ in $S^*$.

We now show that $M$ satisfies every element of $S^*$ of the form $Y_z(u) = y$. We show this by induction on the size of $|V| - |Z|$.
Base Cases:

(i) $Y \in Z$. By effectiveness, $Y_z(u) = y$ is in $M$.

(ii) $|V| - |Z| = 1$. By construction of $f_Y$, $Y_z(u) = y \Longrightarrow Y = y$ is in $M_z$.

Inductive Case:
$|V| - |Z| = k$. Consider $Y_{zx}(u) = y'$, where $x = X_z(u)$. Above, we proved that $X_z(u)$ is satisfied in $M$, and by the inductive hypothesis, $Y_{zx}(u) = y'$ is satisfied in $M$. Thus, by composition, $Y_z(u) = y'$ is satisfied in $M$ and, also by composition, $y = y'$. Thus, $Y_z(u) = y$ is satisfied in $M$. □

# 4 Comparison of Causal Models with Lewis's Closest-World Formalism

We now show that for recursive systems, composition and effectiveness are sound and complete within Lewis's closest-world framework [Lewis, 1973b]. We begin by providing a version of Lewis's logic for counterfactual sentences (from [Lewis, 1981]).

Rules

(1) If $A$ and $A \Longrightarrow B$ are theorems, so is $B$.
(2) If $(B_1 \,\&\, \ldots) \Longrightarrow C$ is a theorem, so is $((A \,\Box\!\!\rightarrow B_1) \ldots) \Longrightarrow (A \,\Box\!\!\rightarrow C)$.

Axioms

(1) All truth-functional tautologies.
(2) $A \,\Box\!\!\rightarrow A$.
(3) $(A \,\Box\!\!\rightarrow B) \,\&\, (B \,\Box\!\!\rightarrow A) \Longrightarrow (A \,\Box\!\!\rightarrow C) \equiv (B \,\Box\!\!\rightarrow C)$.
(4) $((A \vee B) \,\Box\!\!\rightarrow A) \vee ((A \vee B) \,\Box\!\!\rightarrow B) \vee (((A \vee B) \,\Box\!\!\rightarrow C) \equiv (A \,\Box\!\!\rightarrow C) \,\&\, (B \,\Box\!\!\rightarrow C))$.
(5) $A \,\Box\!\!\rightarrow B \Longrightarrow A \Longrightarrow B$.
(6) $A \,\&\, B \Longrightarrow A \,\Box\!\!\rightarrow B$.

The statement $A \mathrel{\Box\!\!\rightarrow} B$ stands for "In all closest worlds where $A$ holds, $B$ holds as well." Lewis does not put any restrictions on the distance measured, except for the obvious requirement that world $w$ be no further from itself than any other world $w' \neq w$. In essence, causal models define an obvious distance measure among worlds, $d(w, w')$, given by the minimal number of local interventions needed for transforming $w$ into $w'$. As such, all of Lewis's axioms are true for causal models and follow from effectiveness, composition, and (for nonrecursive systems) reversibility.

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 of subsets of variables in causal models will stand for Lewis's propositions, (e.g., $A$ and $B$ in the statements above). Thus, in a causal model, the meaning of the Lewis statement $A \mathrel{\Box\!\!\rightarrow} B$ is "If we force a set of variables to have the values $A$, a second set of variables will have the values $B$." Let $A$ stand for a set of values $x_1, \ldots, x_n$ of the variables $X_1, \ldots, X_n$, and let $B$ stand for a set of values $y_1, \ldots, y_m$ of the variables $Y_1, \ldots, Y_m$. Then

$$A \mathrel{\Box\!\!\rightarrow} B \quad \equiv \quad \begin{array}{l} Y_{1 x_1 \ldots x_n}(u) = y_1 \ \& \\ Y_{2 x_1 \ldots x_n}(u) = y_2 \ \& \\ \ldots \\ Y_{m x_1 \ldots x_n}(u) = y_n \ \& \end{array} \tag{22}$$

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 $B$ stand for the proposition $Y = y$. Then

$$Y_x(u) = y \equiv A \mathrel{\Box\!\!\rightarrow} B \tag{23}$$

We can now examine each of Lewis's axioms 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 are restricted to conjunctions of literals, this axiom is irrelevant.

(5) This axiom follows directly from composition.

(6) This axiom follows directly from composition.

Likewise, composition and effectiveness follow from Lewis's axioms. 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 system of closest worlds is sufficient for recursive systems. 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, $Y = y$ may hold in all closest $w$-worlds, $W = w$ may hold in all closest $y$-worlds, and $Y = y$ 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 structure of counterfactual statements in recursive systems.

# 5 Applying Counterfactual Derivation: Example

Consider the century-old debate over the effect of smoking on the incidence of lung cancer. According to many, the tobacco industry has managed to block anti-smoking legislation by arguing that the observed correlation between smoking (X) and lung cancer (Y) could be explained by some sort of carcinogenic genotype ($U_1$) that involves inborn craving for nicotine.[9] However, according to the Surgeon General's report of 1964, there is a causal link between smoking and lung cancer that is mediated by the accumulation of tar ($Z$) deposited in a person's lungs. The two claims are combined in the graph of Figure 3, which represents causal models having the following structure:

$V = \{X \text{ (Smoking)}, Y \text{ (Lung Cancer)}, Z \text{ (Tar in Lungs)}\}$
$U = \{U_1, U_2\}, \qquad U_1 \perp\!\!\!\perp U_2$
$x = f_1(u_1)$
$z = f_2(x, u_2)$
$y = f_3(z, u_1)$

The graphical model embodies several assumptions. 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. To justify the missing link between $U_1$ and $U_2$, we must assume 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.

To demonstrate how counterfactual analysis can help assess the degree to which cigarette smoking increases (or decreases) lung cancer risk, imagine a study in which the three variables, $X$, $Y$, and $Z$, were measured simultaneously on a large, randomly selected sample from the population. From such data, we wish to assess the risk of lung cancer (for a randomly chosen person in the population) under two hypothetical policies: smoking ($X = 1$) and refraining from smoking ($X = 0$). In other words, we wish to derive an expression for the probability of $Y = y$ under the action $do(X = x), P(Y = y|do(x)) = P(Y_x = y)$, based on the joint distribution $P(x, y, z)$ and the assumptions embodied in the graphical model.

In [Pearl, 1995a] this problem was solved by a graphical method, using a set of axioms which, when certain conditions hold in the graph, transform expressions of the form $P(y|z, do(x))$ into other expressions of this type, so as to eliminate the $do(\cdot)$ operator. Here we show how the counterfactual expression $P(Y_x = y)$ can be reduced to ordinary probabilistic expression (involving no counterfactuals) by purely symbolic machinery, using only probability calculus and two rules of inference: effectiveness and composition. To this end,

[9] For an excellent historical account of this debate, see [Spirtes *et al.*, 1993, pp. 291–302].

we first need to translate the assumptions embodied in the graphical model into the language of counterfactuals. In [Pearl, 1995a, p. 704] it is shown that the translation can be accomplished systematically, using two simple rules:

Rule 1 *Exclusion restrictions.* For every variable $Y$ having parents $PA_Y$, and for every set of variables $Z$ disjoint of $PA_Y$, we have

$$Y_{pa_Y}(u) = Y_{pa_Y z}(u) \tag{24}$$

Rule 2 *Independence restrictions.* If $Z_1, \ldots, Z_k$ is any set of nodes in $V$ not connected to $Y$ via a path containing only $U$ variables, we have

$$Y_{pa_Y} \perp\!\!\!\perp \{Z_{1pa_{Z_1}}, \ldots, Z_{kpa_{Z_k}}\} \tag{25}$$

Rule 1 reflects the insensitivity of $Y$ to any manipulation, once its direct causes $PA_Y$ are held constant; it follows from the identity $v_i = f_i(pa_i, u)$ in Definition 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 $U_Y$.

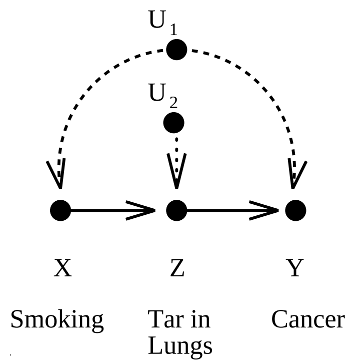


Figure 3: Causal graph illustrating the effect of smoking on lung cancer.

Applying these two rules, we see that the causal graph encodes the following assumptions:

$$
\begin{aligned}
Z_x(u) &= Z_{yx}(u) && (26)\\
X_y(u) &= X_{zy}(u) = X_z(u) = X(u) && (27)\\
Y_z(u) &= Y_{zx}(u) && (28)\\
Z_x &\perp\!\!\!\perp \{Y_z, X\} && (29)
\end{aligned}
$$

Eqs. (26-28) follow from the exclusion restrictions of Eq. (24), using:

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

Eq. (26), for instance, represents the absence of a causal link from $Y$ to $Z$, while Eq. (27) represents the absence of a causal link from $Z$ or $Y$ to $X$. In contrast, Eq. (29) follows from the independence restriction of Eq. (25), 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, and the properties of composition and effectiveness, to compute various tasks:

**Task 1** Compute $P(Z_x = z)$, i.e., the causal effect of smoking on tar.

$$\begin{aligned} P(Z_x = z) &= P(Z_x = z|x) && \text{from Eq. (29)} \\ &= P(Z = z|x) && \text{by composition} \\ &= P(z|x) \end{aligned} \tag{30}$$

**Task 2** Compute $P(Y_z = y)$, i.e., the causal effect of tar on cancer.

$$P(Y_z = y) = \sum_x P(Y_z = y|x)P(x) \tag{31}$$

and since Eq. (29) implies $Y_z \perp\!\!\!\perp Z_x | X$, we can write

$$\begin{aligned} P(Y_z = y|x) &= P(Y_z = y|x, Z_x = z) && \text{from Eq. (29)} \\ &= P(Y_z = y|x, z) && \text{by composition} \\ &= P(y|x, z) && \text{by composition} \end{aligned} \tag{32}$$

Substituting Eq. (32) in Eq. (31) gives

$$P(Y_z = y) = \sum_x P(y|x, z)P(x) \tag{33}$$

**Task 3** Compute $P(Y_x = y)$, i.e., the causal effect of smoking on cancer.

For any variable $Z$,

$$Y_x(u) = Y_{xz}(u), \quad \text{if} \quad Z_x(u) = z \qquad \text{by composition}$$

Since $Y_{xz}(u) = Y_z(u)$ (from Eq. (28)),

$$Y_x(u) = Y_{xz_x}(u) = Y_z(u) \quad \text{where} \quad z_x = Z_x(u) \tag{34}$$

Thus,

$$\begin{aligned} P(Y_x = y) &= P(Y_{z_x} = y) && \text{from Eq. (34)} \\ &= \textstyle\sum_z P(Y_{z_x} = y|Z_x = z)P(Z_x = z) \\ &= \textstyle\sum_z P(Y_z = y|Z_x = z)P(Z_x = z) && \text{by composition} \\ &= \textstyle\sum_z P(Y_z = y)P(Z_x = z) && \text{from Eq. (29)} \end{aligned} \tag{35}$$

$P(Y_x = y)$ and $P(Z_x = z)$ were computed in Eq. (30) and Eq. (33). Substituting gives us

$$P(Y_x = y) = \sum_z P(z|x) \sum_{x'} P(y|z, x')P(x') \tag{36}$$

The right hand side of Eq. (36) can be computed from $P(x, y, z)$ and coincides with the "front-door" formula derived in [Pearl, 1995a].

In general, a counterfactual quantity such as $P(Y_x = y)$ that can be reduced to expressions involving probabilities of observed variables is called *identifiable* [Fisher, 1966; Pearl, 1997b]. Our completeness result implies that any identifiable counterfactual quantity can be reduced to the correct expression by repeated application of composition and effectiveness.

# 6 Conclusions and Discussion

The completeness of composition and effectiveness in recursive causal models has two major implications. First, it shows that in systems with no feedback, the causal interpretation of counterfactuals adds no restrictions beyond those of Lewis's closest-world interpretation. Thus, the unstructured closest-worlds framework embodies all of the causal restrictions on counterfactuals that are not embodied already by the requirement of recursiveness. In non-recursive systems, however, there is a difference between the two formalisms; the causal reading of counterfactuals imposes the additional restriction of reversibility.

Second, the completeness result assures us that a deduction of counterfactual relationships in recursive models may safely be attempted with two axioms only, that is, all truths derivable by structural equation semantics are also derivable using effectiveness and composition. This establishes, in essence, the formal equivalence of structural equation modeling, popular in economics and the social sciences [Goldberger, 1992], and the potential-response framework, as used in statistics [Rubin, 1974; Holland, 1986; Robins, 1986].[10] In nonrecursive models, however, this is not necessarily the case. Attempts to evaluate counterfactual statements using only composition and effectiveness may fail to certify some statements that are true in all causal models but whose validity can only be recognized through the use of reversibility.[11]

The structural-counterfactual equivalence established in this paper does not in any way diminish the usefulness of structural equations and graphs in causal analysis. Graphs and equations are indispensable tools for expressing the assumptions that make up a causal model. Such assumptions must 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 from which we draw inferences about actions, changes, and their ramifications. These mechanisms are thus the building blocks from which judgments about

[10] This equivalence was anticipated in Holland (1988), Pratt and Schlaiffer (1988), Pearl (1995), and Robins (1995). Note, though, that the equation-deletion part of our model (Definition 2) is not made explicit in the standard literature on structural equation modeling.

[11] Joseph Halpern (1997) has recently shown that composition, reversibility, effectiveness, and definiteness are complete in recursive as well as nonrecursive models, as long as the uniqueness assumption holds. He also characterized systems in which uniqueness does not hold, using axioms of more elaborate syntax.

counterfactuals are derived. Structural equations $\{f_i\}$ and their graphical abstraction $G(M)$ provide faithful mapping for these mechanisms and constitute, therefore, the most natural language for articulating or verifying causal assumptions. Thus, graphical specification of assumptions, followed by translation into counterfactual notation and then by symbolic derivation, as exemplified in Section 5, should yield a more effective method of analysis than a method that insists on expressing assumptions directly as counterfactuals. Indeed, an assumption such as the one expressed in Eq. (29) is not easily comprehended by even skilled investigators. In contrast, its structural image $U_1 \perp\!\!\!\perp U_2$ evokes an immediate process-based interpretation.

Graphs may also assist symbolic proof procedures [Galles and Pearl, 1997b] by displaying independence relations (among counterfactuals as well as measured variables) that are not easily derived symbolically [Balke and Pearl, 1994]. For example, it is not straightforward to show that the assumptions of Eqs. (26)-(29) imply the conditional independence $(Y_z \perp\!\!\!\perp Z_x | \{Z, X\})$ but do not imply the conditional independence $(Y_z \perp\!\!\!\perp Z_x | Z)$. Such implications can be easily tested in the graph of Figure 3 or in the dual-graph method of [Balke and Pearl, 1994].

But perhaps the most compelling reason for molding causal assumptions in the language of graph is that such assumptions are needed before the data are gathered, at a stage when the model's parameters are still "free," that is, 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, hence of scientific legitimacy. 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 modeling 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 [Pearl, 1998b]. 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 $(X \perp\!\!\!\perp Y | Z, U_1)$, which is implied by the graph of Figure 3, 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 called *errors* or *disturbance terms*. For example, 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 in the equations that govern response variables. Naturally, such definitions attract criticism from strict empiricists, who regard unobservables as metaphysical or definitional [Richard, 1980; Engle et al., 1983; Holland, 1988], and from counterfactual analysts, who regard the use of equations as an unwarranted commitment to a particular functional form [Angrist *et al.*, 1996].

The analyses of this paper shed new light on this controversy by explicating the operational meaning of the "so-called disturbance terms" [Richard, 1980] and by clarifying the relationships among error-based, counterfactual, and graphical definitions. These three modes of description form a simple hierarchy. Since graph separation implies independence, but independence does not imply graph separation [Pearl, 1988], 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 hierarchy:

$$\textit{Graphical criteria} \Rightarrow \textit{Error-based criteria} \Rightarrow \textit{Counterfactual criteria}$$

The econometric notion of exogeneity may serve to illustrate this hierarchy. The pragmatic definition of exogeneity is best formulated in counterfactual or interventional terms, and reads:

**Counterfactual exogeneity:** $X$ is exogenous relative to $Y$ iff the effect of $X$ on $Y$ is identical to the conditional probability of $Y$ given $X$, namely, if

$$P(Y_x = y) = P(y|x) \tag{37}$$

or, equivalently,

$$P(Y = y|do(x)) = P(y|x) \tag{38}$$

which, in turns, is equivalent to the independence condition $Y_x \perp\!\!\!\perp X$, named "ignorability" in [Rosenbaum and Rubin, 1983].

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 into verifying, from substantive knowledge of the domain, whether the condition above holds in any given system, especially when many equations are involved. To facilitate such judgments, economists [e.g., Koopmans, 1950] have adopted the error-based definition:

**Error-based exogeneity:** $X$ 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 $X$.[12]

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 a scientist is familiar. Still, to judge whether such factors are statistically independent is a difficult mental task unless the independencies considered are dictated by topological considerations, which assures their stability. Indeed, the most popular conception of exogeneity is encapsulated in the notion of "common cause," formally:

**Graphical exogeneity:** $X$ is exogenous relative to $Y$ if $X$ and $Y$ have no common ancestor in $G(M)$.[13]

It is not hard to show that the graphical condition implies the error-based condition, which, in turns, implies the counterfactual (or pragmatic) condition of Eq. (38). The latter implication immediately rules out any contention that the error terms are metaphysical or

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

[13] The augmented graph $G(M)$ should be used in this test, where a latent common parent is added for every pair of dependent errors. This definition paraphrases the "back-door criterion" [Pearl, 1995a] in the special case of no covariates. The incorporation of observed covariates is straightforward in all three definitions.

definitional, as suggested by Hendry (1995, p. 62) and Holland (1988, p. 460). The equality in Eq. (38), and hence its error-independence implicant, is clearly within the realm of empirical verification, albeit requiring controlled experiments. From a narrow empiricist viewpoint, the meaning of an error term $u_Y$ is defined through the equation $Y_{pa_Y} = f_Y(pa_Y, u_Y)$, which states that the variable $U_Y$ is merely a convenient device for encoding variations in the functional mapping from $PA_Y$ to $Y$. The statistics of these variations are observable when $pa_Y$ is held fixed. From a broader perspective, however, the error terms can be viewed as (summaries of) a highly structured background knowledge, whose empirical basis may well lie outside the boundaries of specific study at hand [Pearl, 1998a].

A three-level hierarchy similarly characterizes the notion of *instrumental variables* [Bowden and Turkington, 1984; Pearl, 1995b]. The traditional definition qualifies a variable $Z$ as *instrument* (relative the pair $(X, Y)$) if $(i)$ $Z$ is independent of all terms in the equation for $Y$ (excluding $X$ and variables affected by $X$) and $(ii)$ $Z$ is not independent of $X$. The counterfactual definition replaces the former condition with $(i')$ $Z$ is independent of $Y_x$, while the corresponding graphical condition reads $(i'')$ every path connecting $Z$ and $Y$ must pass through $X$, unless it contains arrows pointing head-to-head.

Note that, in both examples, the graphical definitions are insensitive to the value of the model's parameters and can therefore be ascertained using our general, qualitative understanding of how mechanisms and processes are tied together. It is for this reason that graphical vocabulary guides and expresses so well our intuition about exogeneity, instruments, confounding, and even (I speculate) more technical notions such as randomness and statistical independence.

# Acknowledgment

This investigation has benefited significantly from the input of Joseph Halpern and from discussions with John Aldrich, Phil Dawid, David Freedman, Arthur Goldberger, Sander Greenland, Paul Holland, Guido Imbens, Ed Leamer, Charles Manski, Jamie Robins, Don Rubin, and Glenn Shafer. The research was partially supported by AFOSR, NSF, Northrop, and Rockwell.

# Appendix A

## Independence of Composition, Effectiveness, and Reversibility

We show that reversibility, composition, and effectiveness are independent by creating a table of counterfactual statements such that two of the properties hold, but the third does not. We will consider a small model, one with only two binary variables $X$ and $Y$, and a single value for $U$.

### A.1 Composition and Effectiveness, not Reversibility

$$
\begin{array}{cccc}
X = 0 & Y = 0 & & \\
X_{X=0} = 0 & Y_{X=0} = 0 & X_{X=0,Y=0} = 0 & Y_{X=0,Y=0} = 0 \\
X_{X=1} = 1 & Y_{X=1} = 1 & X_{X=0,Y=1} = 0 & Y_{X=0,Y=1} = 1 \\
X_{Y=0} = 0 & Y_{Y=0} = 0 & X_{X=1,Y=0} = 1 & Y_{X=1,Y=0} = 0 \\
X_{Y=1} = 1 & Y_{Y=1} = 1 & X_{X=1,Y=1} = 1 & Y_{X=1,Y=1} = 1
\end{array}
$$

### A.2 Effectiveness and Reversibility, not Composition

$$
\begin{array}{cccc}
X = 0 & Y = 1 & & \\
X_{X=0} = 0 & Y_{X=0} = 1 & X_{X=0,Y=0} = 0 & Y_{X=0,Y=0} = 0 \\
X_{X=1} = 1 & Y_{X=1} = 0 & X_{X=0,Y=1} = 0 & Y_{X=0,Y=1} = 1 \\
X_{Y=0} = 0 & Y_{Y=0} = 0 & X_{X=1,Y=0} = 1 & Y_{X=1,Y=0} = 0 \\
X_{Y=1} = 1 & Y_{Y=1} = 1 & X_{X=1,Y=1} = 1 & Y_{X=1,Y=1} = 1
\end{array}
$$

### A.3 Composition and Reversibility, not Effectiveness

$$
\begin{array}{cccc}
X = 0 & Y = 1 & & \\
X_{X=0} = 0 & Y_{X=0} = 1 & X_{X=0,Y=0} = 0 & Y_{X=0,Y=0} = 1 \\
X_{X=1} = 0 & Y_{X=1} = 1 & X_{X=0,Y=1} = 0 & Y_{X=0,Y=1} = 1 \\
X_{Y=0} = 0 & Y_{Y=0} = 1 & X_{X=1,Y=0} = 0 & Y_{X=1,Y=0} = 1 \\
X_{Y=1} = 0 & Y_{Y=1} = 1 & X_{X=1,Y=1} = 0 & Y_{X=1,Y=1} = 1
\end{array}
$$

# References


[Angrist *et al.*, 1996] J.D. Angrist, G.W. Imbens, and Rubin D.B. Identification of causal effects using instrumental variables (with comments). *Journal of the American Statistical Association*, 91(434):444–472, June 1996.

[Balke and Pearl, 1994] A. Balke and J. Pearl. Counterfactual probabilities: Computation methods, bounds, and applications. In R.L. de Mantaras and D. Poole, editors, *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pages 11–18, San Francisco, 1994. Morgan Kaufmann.

[Balke and Pearl, 1995] A. Balke and J. Pearl. Counterfactuals and policy analysis in structural models. In P. Besnard and S. Hanks, editors, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 11–18, San Francisco, 1995. Morgan Kaufmann.

[Bowden and Turkington, 1984] R.J. Bowden and D.A. Turkington. *Instrumental Variables.* Cambridge University Press, Cambridge, MA, 1984.

[Dawid, 1997] A.P. Dawid. Causal inference without counterfactuals. Technical Report, Department of Statistical Science, University College London, UK, 1997.

[Dhrymes, 1970] P.J. Dhrymes. *Econometrics.* Springer-Verlag, New York, 1970.

[Engle *et al.*, 1983] R.F. Engle, D.F. Hendry, and J.F. Richard. Exogeneity. *Econometrical*, 51(2):277–304, March 1983.

[Fagin *et al.*, 1995] R. Fagin, J.Y. Halpern, Y. Moses, and M.Y. Vardi. *Reasoning About Knowledge.* MIT Press, Cambridge, MA, 1995.

[Fisher, 1966] F.M. Fisher. *The Identification Problem in Econometrics.* McGraw-Hill, New York, 1966.

[Fisher, 1970] F.M. Fisher. A correspondence principle for simultaneous equation models. *Econometrica*, 38:73–92, 1970.

[Galles and Pearl, 1997a] D. Galles and J. Pearl. An axiomatic characterization of causal counterfactuals. Technical Report R-250-L, Computer Science Department, University of California, Los Angeles, March 1997. Prepared for *Foundations of Science*, Kluwer Academic Publishers.

[Galles and Pearl, 1997b] D. Galles and J. Pearl. Axioms of causal relevance. *Artificial Intelligence*, 97(1-2):9–43, 1997.

[Gibbard and Harper, 1981] A. Gibbard and L. Harper. Counterfactuals and two kinds of expected utility. In W.L. Harper, R. Stalnaker, and G. Pearce, editors, *Ifs.* D. Reidel, Dordrecht: Holland, 1981.

[Ginsberg and Smith, 1987] M.L. Ginsberg and D.E. Smith. Reasoning about action I: A possible worlds approach. In Frank M. Brown, editor, *The Frame Problem in Artificial Intelligence*, pages 233–258. Morgan Kaufmann, Los Altos, CA, 1987.

[Goldberger, 1992] Arthur S. Goldberger. Models of substance [comment on N. Wermuth, "On block-recursive linear regression equations"]. *Brazilian Journal of Probability and Statistics*, 6:1–56, 1992.

[Halpern, 1998] J. Halpern. Axiomatizing causal reasoning. Unpublished report, Cornell University, February, 1998.

[Heckerman and Shachter, 1995] D. Heckerman and R. Shachter. A definition and graphical representation of causality. In P. Besnard and S. Hanks, editors, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 262–273, San Francisco, 1995. Morgan Kaufmann.

[Hendry, 1995] David F. Hendry. *Dynamic Econometrics*. Oxford University Press, New York, 1995.

[Holland, 1986] P. W. Holland. Statistics and causal inference (with discussion). *Journal of the American Statistical Association*, 81(396):945–970, 1986.

[Holland, 1988] P.W. Holland. Causal inference, path analysis, and recursive structural equations models. In C. Clogg, editor, *Sociological Methodology*, pages 449–484. American Sociological Association, Washington, D.C., 1988.

[Katsuno and Mendelzon, 1991] H. Katsuno and A.O. Mendelzon. On the difference between updating a knowledge base and revising it. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pages 387–394, Boston, MA, 1991.

[Koopmans, 1950] T.C. Koopmans. When is an equation system complete for statistical purposes? In T.C. Koopmans, editor, *Statistical Inference in Dynamic Economic Models*, Cowles Commission, Monograph 10. Wiley, New York, 1950. Reprinted in D.F. Hendry and M.S. Morgan (Eds.), *The Foundations of Econometric Analysis*, Cambridge University Press, 527–537, 1995.

[Leamer, 1985] E. Leamer. Vector autoregression for causal inference? *Carnegie-Rochester Conference Series on Public Policy*, 22:255–304, 1985.

[Lewis, 1973a] D. Lewis. Causation. *Journal of Philosophy*, 70:556–567, 1973.

[Lewis, 1973b] D. Lewis. *Counterfactuals*. Harvard University Press, Cambridge, MA, 1973.

[Lewis, 1981] D. Lewis. Counterfactuals and comparative possibility. In W.L. Harper, R. Stalnaker, and G. Pearce, editors, *Ifs*. D. Reidel, Dordrecht, Holland, 1981.

[Manski, 1990] C.F. Manski. Nonparametric bounds on treatment effects. *American Economic Review, Papers and Proceedings*, 80:319–323, 1990.

[Meditch, 1969] J.S. Meditch. *Stochastic Optimal Linear Estimation and Control.* McGraw-Hill, New York, 1969.

[Pearl, 1988] J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Mateo, CA, 1988. (Revised 2nd printing, 1992).

[Pearl, 1994] J. Pearl. A probabilistic calculus of actions. In R.L. de Mantaras and D. Poole, editors, *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pages 454–462, San Francisco, 1994. Morgan Kaufmann.

[Pearl, 1995a] J. Pearl. Causal diagrams for empirical research (with discussion). *Biometrika*, 82(4):669–710, 1995.

[Pearl, 1995b] J. Pearl. On the testability of causal models with latent and instrumental variables. In P. Besnard and S. Hanks, editors, *Uncertainty in Artificial Intelligence 11*, pages 435–443. Morgan Kaufmann, 1995.

[Pearl, 1996] J. Pearl. Structural and probabilistic causality. *Psychology of Learning and Motivation*, 34:393–435, 1996.

[Pearl, 1997a] J. Pearl. The new challenge: From a century of statistics to an age of causation. *Computing Science and Statistics*, 29(2):415–423, 1997.

[Pearl, 1997b] J. Pearl. On the identification of nonparametric structural models. In M. Berkane, editor, *Latent Variable Modeling with Application to Causality*, pages 29–68. Springer-Verlag, 1997.

[Pearl, 1998a] J. Pearl. Graphs, causality, and structural equation models. Technical Report R-253, Department of Computer Science, University of California, Los Angeles, 1998. To appear in *Socioligical Methods and Research*, Special Issue on Causality.

[Pearl, 1998b] J. Pearl. Why there is no statistical test for confounding, why many think there is, and why they are almost right. Technical Report R-256, Department of Computer Science, University of California, Los Angeles, 1998.

[Pratt and Schlaifer, 1988] J.W. Pratt and R. Schlaifer. On the interpretation and observation of laws. *Journal of Econometrics*, 39:23–52, 1988.

[Richard, 1980] J.F. Richard. Models with several regimes and charges in exogeneity. *The Review of Economic Studies*, 47:1–20, 1980. from orion.

[Robins, 1986] J. Robins. A new approach to causal inference in mortality studies with a sustained exposure period – applications to control of the healthy workers survivor effect. *Mathematical Modeling*, 7:1393–512, 1986.

[Robins, 1987] J. Robins. Addendum to "A new approach to causal inference in mortality studies with sustained exposure periods – application to control of the healthy worker survivor effect". *Computers and Mathematics, with Applications.*, 14:923–45, 1987.

[Robins, 1995] J.M. Robins. Discussion of "Causal diagrams for empirical research" by J. Pearl. *Biometrika*, 82(4):695–698, 1995.

[Rosenbaum and Rubin, 1983] P. Rosenbaum and D. Rubin. The central role of propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[Rubin, 1974] D.B. Rubin. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

[Simon and Rescher, 1966] H.A. Simon and N. Rescher. Cause and counterfactual. *Philosophy and Science*, 33:323–340, 1966.

[Sobel, 1990] M.E. Sobel. Effect analysis and causation in linear structural equation models. *Psychometrika*, 55:495–515, 1990.

[Spirtes *et al.*, 1993] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Springer-Verlag, New York, 1993.

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

[Winslett, 1988] M. Winslett. Reasoning about action using a possible worlds approach. In *Proceedings of the Seventh American Association for Artificial Intelligence Conference*, pages 89–93, 1988.