Probabilities of causation: Three counterfactual interpretations and their identification

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

Abstract

According to common judicial standard, judgment in favor of plaintiff should be made if and only if it is "more probable than not" that the defendant's action was the *cause* for the plaintiff's damage (or death). This paper provides formal semantics, based on structural models of counterfactuals, for the probability that event x was a *necessary* or *sufficient* cause (or both) of another event y. The paper then explicates conditions under which the probability of necessary (or sufficient) causation can be learned from statistical data, and shows how data from both experimental and nonexperimental studies can be combined to yield information that neither study alone can provide. Finally, we show that necessity and sufficiency are two independent aspects of causation, and that both should be invoked in the construction of causal explanations for specific scenarios.

1 Introduction

The standard counterfactual definition of causation[1] (i.e., that E would not have occurred if it were not for C), captures the notion of "necessary cause." Competing notions such as "sufficient cause" and "necessary-and-sufficient cause" may be of interest in a number of applications,[2] and these, too, can be given concise counterfactual definitions. One advantage of casting aspects of causation in the language of counterfactuals is that the latter enjoys natural and formal semantics in terms of structural models [Galles and Pearl, 1997, 1998;

[1] This definition dates back to Hume (1748, p. 115) and Mill (1843) and has been formalized and advocated in the philosophical work of D. Lewis (1986).

[2] The distinction between necessary and sufficient causes goes back to J.S. Mill (1843), and has received semi-formal explications in the 1960s using the syntax of conditional probabilities [Good, 1961] and logical implications [Mackie, 1965]. The basic limitations of the logical and probabilistic accounts are discussed in Kim (1971) and Pearl (1996, 1998) and stem primarily from lacking syntactic distinction between formulas that represent stable mechanisms and those that represent transitory logical or probabilistic relationships.

Halpern, 1998; Pearl, forthcoming 2000], as well as effective procedures for computing probabilities of counterfactual expressions from a given causal theory [Balke and Pearl, 1994, 1995]. These developments are reviewed in Section 2.

The purpose of this paper is to explore the counterfactual interpretation of necessary and sufficient causes, to illustrate the application of structural-model semantics (of counterfactuals) to the problem of identifying probabilities of causes, and to present, by way of examples, new ways of estimating probabilities of causes from statistical data. Additionally, the paper will argue that necessity and sufficiency are two distinct facets of causation that should be kept apart in any explication of "actual cause" and, using these two facets, we will show how certain problems associated with the standard counterfactual account of causation [Lewis, 1986] can be resolved.

The results have applications in epidemiology, legal reasoning, artificial intelligence (AI), and psychology. Epidemiologists have long been concerned with estimating the probability that a certain case of disease is *attributable* to a particular exposure, which is normally interpreted counterfactually as "the probability that disease would not have occurred in the absence of exposure, given that disease and exposure did in fact occur." This counterfactual notion, which Robins and Greenland (1989) called the "probability of causation" measures how *necessary* the cause is for the production of the effect.[3] It is used frequently in lawsuits, where legal responsibility is at the center of contention. We shall denote this notion by the symbol PN, an acronym for Probability of Necessity.

A parallel notion of causation, capturing how *sufficient* a cause is for the production of the effect, finds applications in policy analysis, AI, and psychology. A policy maker may well be interested in the dangers that a certain exposure may present to the healthy population [Khoury *et al.*, 1989]. Counterfactually, this notion can be expressed as the "probability that a healthy unexposed individual would have gotten the disease had he/she been exposed," and will be denoted by PS (Probability of Sufficiency). A natural extension would be to inquire for the probability of necessary-and-sufficient causation, PNS, namely, how likely a given individual is to be affected both ways.

As the examples illustrate, PS assesses the presence of an active causal process capable of producing the effect, while PN emphasizes the absence of alternative processes, not involving the cause in question, still capable of sustaining the effect. In legal settings, where the occurrence of the cause (x) and the effect (y) are fairly well established, PN is the measure that draws most attention, and the plaintiff must prove that y would not have occurred *but for* x [Robertson, 1997]. Still, lack of sufficiency may weaken arguments based on PN [Good, 1993; Michie, 1997].

It is known that PN is in general non-identifiable, namely, non-estimatable from frequency data involving exposures and disease cases [Greenland and Robins, 1988; Robins and Greenland, 1989]. The identification is hindered by two factors:

1. **Confounding**: exposed and unexposed subjects may differ in several relevant factors

[3] Greenland and Robins (1988) further distinguish between two ways of measuring probabilities of causation: the first (called "excess fraction") concerns only *whether* the effect (e.g., disease) occurs by a particular time, while the second, (called "etiological fraction") requires consideration of *when* the effect occurs. We will confine our discussion here to binary events occurring within a specified time period, hence, will not be concerned with the temporal aspects of etiological fractions.

or, more generally, the cause and the effect may both be influenced by a third factor. In this case we say that the cause is not *exogenous* relative to the effect.

2. **Sensitivity to the generative process**: Even in the absence of confounding, probabilities of certain counterfactual relationships cannot be identified from frequency information unless we specify the functional relationships that connect causes and effects. Functional specification is needed whenever the facts at hand (e.g., disease) might be affected by the counterfactual antecedent (e.g., exposure) [Balke and Pearl, 1994b] (see example in Section 4.1).

Although PN is not identifiable in the general case, several formulas have nevertheless been proposed to estimate attributions of various kinds in terms of frequencies obtained in epidemiological studies [Breslow and Day, 1980; Hennekens and Buring, 1987; Cole, 1997]. Naturally, any such formula must be predicated upon certain implicit assumptions about the data-generating process. This paper explicates some of those assumptions and explores conditions under which they can be relaxed.[4] It offers new formulas for PN and PS in cases where causes are confounded (with outcomes) but their effects can nevertheless be estimated (e.g., from clinical trials or from auxiliary measurements). We further provide a general condition for the identifiability of PN and PS when functional relationships are only partially known (Section 5).

Glymour (1998) has raised a number of issues concerning the identifiability of causal relationships when the functional relationships among the variables *are* known, but some variables are unobserved. These issues surfaced in connection with the psychological model introduced by Cheng according to which people assess the "causal power" between two events by estimating the probability of the effect in a hypothetical model in which certain elements are suppressed [Cheng, 1997]. In the examples provided, Cheng's "causal power" coincides with PS and hence lends itself to counterfactual analysis. Accordingly we shall see that many of the issues raised by Glymour can be resolved and generalized using counterfactual analysis.

The distinction between *necessary* and *sufficient* causes has important implications in AI, especially in systems that generate verbal explanations automatically. As can be seen from the epidemiological examples above, necessary causation is a concept tailored to a specific event under consideration, while sufficient causation is based on the general tendency of certain event *types* to produce other event types. Adequate explanations should respect both aspects. If we base explanations solely on generic tendencies (i.e., *sufficient* causation), we lose important specific information. For instance, aiming a gun at and shooting a person from 1000 meters away will not qualify as an explanation for that person's death, due to the very low tendency of typical shots fired from such long distances to hit their marks. The fact that the shot did hit its mark on that singular day, regardless of the reason, should carry decisive weight when we come to assess whether the shooter is the culprit for the consequence. If, on the other hand, we base explanations solely on singular-event considerations (i.e., *necessary* causation), then various background factors that are normally present in the world would

[4] A set of sufficient conditions for the identification of etiological fractions are given in Robins and Greenland (1989). These conditions, however, are too restrictive for the identification of PN, which is oblivious to the temporal aspects associated with etiological fractions.

awkwardly qualify as explanations. For example, the presence of oxygen in the room would qualify as an explanation for the fire that broke out, simply because the fire would not have occurred if it were it not for the oxygen. Clearly, some balance must be made between the necessary and the sufficient components of causal explanation, and the present paper illuminates this balance by formally explicating some of the basic relationships between the two components. Section 6 further discusses ways of incorporating singular-event information in the definition and evaluation of sufficient causation.

2 Structural Model Semantics (A Review)

This section presents a brief summary of the structural-equation semantics of counterfactuals as defined in Balke and Pearl (1995), Galles and Pearl (1997, 1998), and Halpern (1998). Related approaches have been proposed in Simon and Rescher (1966), Rubin (1974) and Robins (1986). For detailed exposition of the structural account and its applications see [Pearl, 2000].

2.1 Definitions: Causal models, actions and counterfactuals

A causal model is a mathematical object that assigns truth values to sentences involving causal and counterfactual relationships. Basic of our analysis are sentences involving actions or external interventions, such as, "p will be true if we do q" where q is any elementary proposition. Structural models are generalizations of the structural equations used in engineering, biology, economics and social science.[5] World knowledge is represented as a collection of stable and autonomous relationships called "mechanisms," each represented as an equation, and changes due to interventions or hypothetical novel eventualities are treated as local modifications of those equations.

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, namely, variables in $U \cup V$.*

(iii) *F is a set of functions $\{f_1, f_2, \ldots, f_n\}$ where each f_i is a mapping from $U \times (V \setminus V_i)$ to V_i. In other words, each f_i tells us the value of V_i given the values of all other variables in $U \cup V$. Symbolically, the set of equations F can be represented by writing*

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

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

where pa_i is any realization of the unique minimal set of variables PA_i in V/V_i (connoting parents*) that renders f_i nontrivial. Likewise, $U_i \subseteq U$ stands for the unique minimal set of variables in U 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

$$F_x = \{f_i : V_i \notin X\} \cup \{X = x\} \tag{1}$$

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 dictated 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].[6]

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.*

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.*[7]

[6] 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 (1995). A similar notion of sub-model is introduced in Fine (1985), though not specifically for representing actions and counterfactuals.

[7] Galles and Pearl (1998) required that F_x has a unique solution, a requirement later relaxed by Halpern (1998). In this paper we are dealing with recursive systems (i.e., $G(M)$ is acyclic) where uniqueness of solution is ensured.

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

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 $Y_x(u)$.[8]

This 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

$$P(y) \triangleq P(Y = y) = \sum_{\{u \ | \ Y(u)=y\}} P(u) \tag{2}$$

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

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

Likewise a causal model defines a joint distribution on 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

$$P(Y_x = y, X = x') = \sum_{\{u|Y_x(u)=y \ \& \ X(u)=x'\}} P(u) \tag{4}$$

and

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

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 recent objections to treating counterfactuals as jointly distributed random variables [Dawid, 1997]. The definition

[8]The connection between counterfactuals and local actions (sometimes resembling "miracles") is made in Lewis (1986) and is further elaborated in Balke and Pearl (1994) and Heckerman and Shachter (1995).

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 Appendix A) and further illustrates that joint probabilities of counterfactuals can be encoded rather parsimoniously using $P(u)$ and F.

In particular, the probabilities of causation analyzed in this paper (see Eqs. (12)-(14)) require the evaluation of expressions of the form $P(Y_{x'} = y'|X = x, Y = y)$ with x and y incompatible with x' and y', respectively. Eq. (4) allows the evaluation of this quantity as follows:

$$\begin{aligned} P(Y_{x'} = y'|X = x, Y = y) &= \frac{P(Y_{x'} = y', X = x, Y = y)}{P(X = x, Y = y)} \\ &= \sum_u P(Y_{x'}(u) = y')P(u|x, y) \end{aligned} \tag{6}$$

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

2.2 Examples

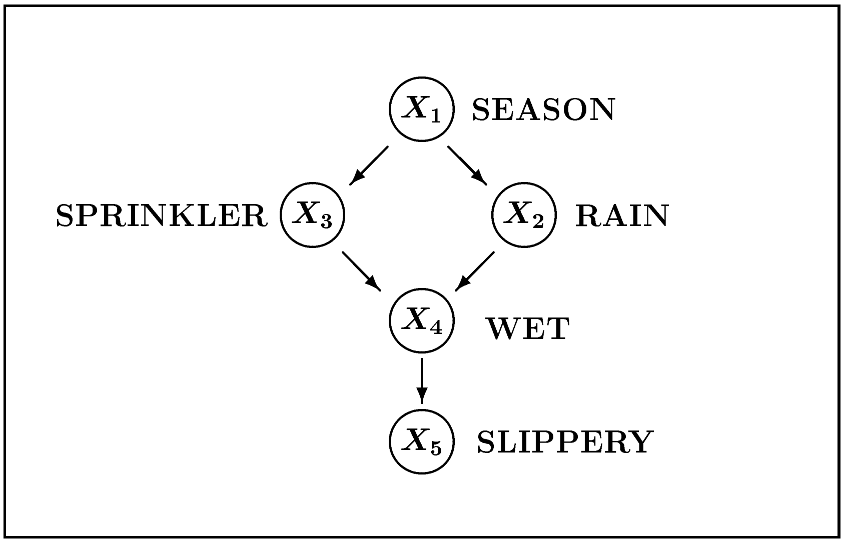

Figure 1: Causal graph illustrating causal relationships among five variables.

Figure 1 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" (encoded "1" and "0" for convenience.) 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 exogenous variables $U_1, \ldots, U_5$, represent factors omitted from the analysis. For example, U_4 may stand for (unspecified) events that would cause the pavement to get wet ($x_4 = 1$) when the sprinkler is off ($x_2 = 0$) and it does not rain ($x_3 = 0$) (e.g., a leaking water pipe). These factors are not shown explicitly in Figure 1 to communicate, by convention, that the U's are assumed independent of one another. When some of these factors are judged to be dependent, it is customary to encode such dependencies by augmenting the graph with double-headed arrows [Pearl, 1995].

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 $x_3 = 1$. The resulting submodel, $M_{X_3=\text{ON}}$, contains all the information needed for computing the effect of the action on the other variables. Note that 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 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 action "turning the sprinkler ON" that a person may consider taking.

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

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2, x_3)P(x_5|x_4) \tag{8}$$

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 $x_3 = 1$.

$$P(x_1, x_2, x_4, x_5|do(X_3 = \text{ON})) = P(x_1)\ P(x_2|x_1)\ P(x_4|x_2, x_3 = 1)\ P(x_5|x_4) \tag{9}$$

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. (9), while the latter by *conditioning* Eq. (8) on the observation, i.e.,

$$P(x_1, x_2, x_4, x_5|X_3 = \text{ON}) = \frac{P(x_1)\ P(x_2|x_1)\ P(x_3 = 1|x_1)P(x_4|x_2, x_3 = 1)P(x_5|x_4)}{P(x_3 = 1)}$$

Note that the conditional probabilities on the r.h.s. of Eq. (9) are the same as those in Eq. (8), and can therefore be estimated from pre-action observations, provided $G(M)$ is available. However, the pre-action distribution P together with the causal graph $G(M)$ is generally not sufficient for evaluating all counterfactuals sentences. For example, the probability that "the pavement would be slippery if the sprinkler were off, given 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].

To illustrate the evaluation of counterfactuals, consider a deterministic version of the model given by Eq. (7) assuming that the only uncertainty in the model lies in the identity of the season, summarized by a probability distribution $P(u_1)$ (or $P(x_1)$.) We observe the ground slippery and the sprinkler on and we wish to assess the probability that the ground would be slippery had the sprinkler been off. Formally, the quantity desired is given by

$$P(X_{5_{x_3=0}} = 1|X_5 = 1, X_3 = 1)$$

According to Eq. (6), the expression above is evaluated by summing over all states of U that are compatible with the information at hand. In our example, the only state compatible with the evidence $X_5 = 1$ and $X_3 = 1$ is that which yields $X_1 = \text{Summer} \vee \text{Spring}$, and in this state $X_2 = \text{no-rain}$, hence $X_{5_{x_3=0}} = 0$. Thus, matching intuition, we obtain

$$P(X_{5_{x_3=0}} = 1|X_5 = 1, X_3 = 1) = 0.$$

In general, the conditional probability of a counterfactual sentence "If it were A then B", given evidence e, can be computed in three steps:

1. **Abduction** – update $P(u)$ by the evidence e, to obtain $P(u|e)$.

2. **Action** – Modify M by the action $do(A)$, where A is the antecedent of the counterfactual, to obtain the submodel M_A.

3. **Deduction** – Use the updated probability $P(u|e)$ in conjunction with M_A to compute the probability of the counterfactual consequence B.

In temporal metaphors [Thomason and Gupta, 1980], this 3-step procedure can be interpreted as follows: Step-1 explains the past (U) in light of the current evidence e, Step-2 bends the course of history (minimally) to comply with the hypothetical condition $X = x$ and, finally, Step-3 predicts the future (Y) based on our new understanding of the past and our new starting condition, $X = x$. Effective methods of computing probabilities of counterfactuals are presented in Balke and Pearl (1994, 1995).

2.3 Relation to Lewis' counterfactuals

The structural model of counterfactuals is closely related to Lewis's account [Lewis, 1986][9], but differs from it in several important aspects. According to Lewis' account, one orders possible worlds by some measure of similarity, and the a counterfactual $A > B$ is true in

[9] $Y_x(u) = y$ can be translated to "$(X = x) > (Y = y)$ in world u."

a world w just in case B is true in all the closest A-worlds to w. This semantics leaves two questions unsettled and problematic: 1. What choice of similarity measure would make counterfactual reasoning compatible with ordinary conception of cause and effect? 2. What mental representation of worlds ordering would render the computation of counterfactuals manageable and practical (in both man and machine.)[10]

Kit Fine's celebrated example (of Nixon pulling the trigger [Fine, 1975]) demonstrates that similarity measures could not be arbitrary, but must respect our conception of causal laws.[11] Lewis (1979) has subsequently set up an intricate system of priorities among various dimensions of similarity: size of miracles (violations of laws), matching of facts, temporal precedence etc., to bring similarity closer to causal intuition. These difficulties do not enter the structural account. In contrast with Lewis' theory, counterfactuals are not based on abstract notion of similarity among hypothetical worlds, but rests directly on the mechanisms (or "laws," to be fancy) that produce those worlds, and on the invariant properties of those mechanisms. Lewis' elusive "miracles" are replaced by principled mini-surgeries, $do(X = x)$, which represent the minimal change (to a model) necessary for establishing the antecedent $X = x$ (for all u). Thus, similarities and priorities, if they are ever needed, may be read into the $do(*)$ operator (see [Goldszmidt and Pearl, 1992]), but do not govern the analysis.

The structural account answers the mental representational question by offering a parsimonious encoding of knowledge, from which causes, counterfactual and probabilities of counterfactuals can be derived by effective algorithms. 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."

2.4 Relation to probabilistic causality

The relation between the structural and probabilistic accounts of causality is best demonstrated when we make the Markov assumption (see Definition 15): 1. The equations $\{f_i\}$ are recursive (i.e., no feedback), and 2. The exogenous terms u_i are mutually independent. Under this assumption, which implies the "screening-off" condition in the probabilistic accounts of causality, it can be shown (e.g., [Pearl, 1995]) that the causal effect of a set X of decision variables on outcome variables Y is given by the formula:

$$P(Y = y|do(X = x)) = \sum_{pa_X} P(y|x, pa_X)P(pa_X) \tag{10}$$

where PA_X is the set of all parents of variables in X. Eq. (10) calls for conditioning $P(y)$ on the event $X = x$ as well as on the parents of X, then averaging the result, weighted by the prior probabilities of those parents. This operation is known as "adjusting for PA_X."

Variations of this adjustment have been advanced by several philosophers as *definitions* of causality or of causal effects. Good (1961), for example, calls for conditioning on "the state of the universe just before" the occurrence of the cause. Suppes (1970) calls for conditioning on the entire past, up to the occurrence of the cause. Skyrms (1980, p. 133) calls for conditioning

[10] Since matching human intuition is the ultimate success criterion in most philosophical theories of causation, questions of cognitive compatibility must be considered an integral part of any such theory.

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

on "... maximally specific specifications of the factors outside of our influence at the time of the decision which are causally relevant to the outcome of our actions ...". The aim of conditioning in these proposals is, of course, to eliminate spurious correlations between the cause (in our case $X = x$) and the effect (in our case $Y = y$) and, clearly, the set PA_X of direct causes accomplishes this aim with great economy. However, the averaged conditionalization operation is not attached here as an add-on *adjustment*, aimed at irradicating spurious correlations. Rather, it emerges purely formally from the deeper principle of discarding the obsolete and preserving all the invariant information that the pre-action distribution can provide. Thus, while probabilistic causality first confounds causal effects $P(y|do(x))$ with epistemic conditionalization $P(y|x)$, then gets rid of spurious correlations through remedial steps of adjustment, the structural account defines causation directly in terms of Nature's invariants (i.e., submodel M_x in Definition 3).

One tangible benefit of this conception is the ability to process commonplace causal statements in their natural deterministic habitat, without having to immerse them in non-deterministic decor. In other words, an event $X = x$ for which $P(x|pa_X) = 1$ (e.g., the output of a logic circuit), may still be a *cause* of some other event, $Y = y$. Consequently, probabilities of single-case causation are well defined, free of the difficulties that plague explications based on conditional probabilities. A second benefit lies in the generality of the structural equation model vis a vis probabilistic causality; interventions, causation and counterfactuals are well defined without invoking the Markov assumptions. Additionally, and most relevant to the topic of this paper, such ubiquitous notions as "probability of causation" cannot easily be defined in the language of probabilistic causality (see discussion after Corollary 1, and Section 4.1).

Finally, we should note that the structural model, as it is presented in Section 2.1, is quasi-deterministic or Laplacian; chance arises only from unknown prior conditions as summarized in $P(u)$. Those who frown upon this classical approximation should be able to extend the results of this paper along more fashionable lines (see appendix for an outline). However, considering that Laplace's illusion still governs human conception of cause and effect, I doubt that significant insight will be gained by such exercise.

2.5 Relation to Neyman-Rubin model

Several concepts defined in Section 2.1 bear similarity to concepts in the potential-outcome model used by Neyman (1923) and Rubin (1974) in the statistical analysis of treatment effects. In that model, $Y_x(u)$ stands for the outcome of experimental unit u (e.g., an individual, or an agricultural lot) under experimental condition $X = x$, and is taken as a primitive, that is, as an undefined relationship, in terms of which one must express assumptions about background knowledge. In the structural model framework, the quantity $Y_x(u)$ is not a primitive, but is derived mathematically from a set of equations F that is modified by the operator $do(X = x)$. Assumptions about causal processes are expressed naturally in the form of such equations. The variable U represents any set of exogenous factors relevant to the analysis, not necessarily the identity of a specific individual in the population.

Using this semantics, it is possible to derive a complete axiomatic characterization of the constraints that govern the potential response function $Y_x(u)$ vis-a-vis those that govern directly observed variables, such as $X(u)$ and $Y(u)$ [Galles and Pearl, 1998; Halpern, 1998].

These basic axioms include or imply relationships that were taken as given, and used extensively by statisticians who pursue the potential-outcome approach. Prominent among these we find the consistency condition [Robins, 1987]:

$$(X = x) \Rightarrow (Y_x = Y) \tag{11}$$

stating that if we intervene and set the experimental conditions $X = x$ equal to those prevailing before the intervention, we should not expect any change in the response variable Y. (For example, a subject who selects treatment $X = x$ by choice and responds with $Y = y$ would respond in exactly the same way to treatment $X = x$ under controlled experiment.) This condition is a proven theorem is structural-model semantics [Galles and Pearl, 1998] and will be used in several of the derivations of Section 3. Rules for translating the topology of a causal diagram into counterfactual sentences are given in [Pearl, 2000, Chapter 7].

3 Necessary and Sufficient Causes: Conditions of Identification

3.1 Definitions, notation, and basic relationships

Using the counterfactual notation and the structural model semantics introduced in Section 2.1, we give the following definitions for the three aspects of causation discussed in the introduction.

Definition 7 (*Probability of necessity* (*PN*))
Let X and Y be two binary variables in a causal model M, let x and y stand for the propositions $X =$ true and $Y =$ true, respectively, and x' and y' for their complements. The probability of necessity is defined as the expression

$$\begin{aligned} PN &\triangleq P(Y_{x'} = \text{false} \mid X = \text{true}, Y = \text{true}) \\ &\triangleq P(y'_{x'}|x, y) \end{aligned} \tag{12}$$

In other words, PN stands for the probability that event y would not have occurred in the absence of event x, ($y'_{x'}$), given that x and y did in fact occur.

Note a slight change in notation relative to that used Section 2. Lower case letters (e.g., x, y) denoted values of variables in Section 2, and now stand for propositions (or events). Note also the abbreviations y_x for $Y_x = true$ and y'_x for $Y_x = false$.[12] Readers accustomed to writing "$A > B$" for the counterfactual "B if it were A" can translate Eq. (12) to read $PN \triangleq P(x' > y'|x, y)$.

Definition 8 (*Probability of sufficiency* (*PS*))

$$PS \triangleq P(y_x|y', x') \tag{13}$$

[12]These were proposed by Peyman Meshkat in class homework, and substantially simplify the derivations.

PS measures the capacity of x to *produce* y and, since "production" implies a transition from the absence to the presence of x and y, we condition the probability $P(y_x)$ on situations where x and y are both absent. Thus, mirroring the necessity of x (as measured by PN), PS gives the probability that setting x would produce y in a situation where x and y are in fact absent.

Definition 9 (*Probability of necessity and sufficiency* (*PNS*))

$$PNS \triangleq P(y_x, y'_{x'}) \tag{14}$$

PNS stands for the probability that y would respond to x both ways, and therefore measures both the sufficiency and necessity of x to produce y.

Associated with these three basic notions, there are other counterfactual quantities that have attracted either practical or conceptual interest. We will mention two such quantities, but will not dwell on their analyses, since these can be easily inferred from our treatment of PN, PS, and PNS.

Definition 10 (*Probability of disablement* (*PD*))

$$PD \triangleq P(y'_{x'}|y) \tag{15}$$

PD measures the probability that y would have been prevented if it were not for x; it is therefore of interest to policy makers who wish to assess the social effectiveness of various prevention programs [Fleiss, 1981, pp. 75–76].

Definition 11 (*Probability of enablement* (*PE*))

$$PE \triangleq P(y_x|y')$$

PE is similar to PS, save for the fact that we do not condition on x'. It is applicable, for example, when we wish to assess the danger of an exposure on the entire population of healthy individuals, including those who were already exposed.

Although none of these quantities is sufficient for determining the others, they are not entirely independent, as shown in the following lemma.

Lemma 1 *The probabilities of causation, PNS, PN and PS satisfy the following relationship:*

$$PNS = P(x, y)PN + P(x', y')PS \tag{16}$$

Proof of Lemma 1
Using the consistency conditions of Eq. (11),

$$x \Rightarrow (y_x = y), \qquad x' \Rightarrow (y_{x'} = y)$$

we can write

$$\begin{aligned} y_x \wedge y'_{x'} &= (y_x \wedge y'_{x'}) \wedge (x \vee x') \\ &= (y \wedge x \wedge y'_{x'}) \vee (y_x \wedge y' \wedge x') \end{aligned}$$

Taking probabilities on both sides, and using the disjointness of x and x', we obtain:

$$\begin{aligned} P(y_x, y'_{x'}) &= P(y'_{x'}, x, y) + P(y_x, x', y') \\ &= P(y'_{x'}|x, y)P(x, y) + P(y_x|x', y')P(x', y') \end{aligned}$$

which proves Lemma 1. □

To put into focus the aspects of causation captured by PN and PS, it is helpful to characterize those changes in the causal model that would leave each of the two measures invariant. The next two lemmas show that PN is insensitive to the introduction of potential inhibitors of y, while PS is insensitive to the introduction of alternative causes of y.

Lemma 2 *Let $PN(x, y)$ stand the for the probability that x is a necessary cause of y, and $z = y \wedge q$ a consequence of y, potentially inhibited by q'. If $q \perp\!\!\!\perp \{X, Y_x, Y_{x'}\}$, then*

$$PN(x, z) \triangleq P(z'_{x'}|x, z) = P(y'_{x'}|x, y) \triangleq PN(x, y)$$

Cascading the process $Y_x(u)$ with the link $z = y \wedge q$ amounts to inhibiting y with probability $P(q')$. Lemma 2 asserts that we can add such a link without affecting PN, as long as q is randomized. The reason is clear; conditioning on the event x and y implies that, in the scenario under consideration, the added link was not inhibited by q'.

Proof of Lemma 2

$$\begin{aligned} PN(x, z) &= P(z'_{x'}|x, z) = \frac{P(z'_{x'}, x, z)}{P(x, z)} = \\ &= \frac{P(z'_{x'}, x, z|q)P(q) + P(z'_{x'}, x, z|q')P(q')}{P(z, x, q) + P(z, x, q')} \end{aligned} \tag{17}$$

Using $z = y \wedge q$, we have

$$q \Rightarrow (z = y), \;\; q \Rightarrow (z'_{x'} = y'_{x'}), \;\; \text{and } q' \Rightarrow z'$$

therefore

$$\begin{aligned} PN(x, z) &= \frac{P(y'_{x'}, x, y|q)P(q) + 0}{P(y, x, q) + 0} \\ &= \frac{P(y'_{x'}, x, y)}{P(y, x)} = P(y'_{x'}|xy) = PN(x, y) \end{aligned}$$

□

Lemma 3 *Let $PS(x,y)$ stand the for the probability that x is a sufficient cause of y, and let $z = y \vee r$ be a consequence of y, potentially triggered by r. Then*

$$PS(x,z) = P(z_x|x',z') = P(y_x|x',y') = PS(x,y)$$

Lemma 3 asserts that we can add alternative independent causes (r), without affecting PS. The reason again is clear; conditioning on the event x' and y' implies that the added causes (r) were not active. The proof of Lemma 3 is similar to that of Lemma 2.

Definition 12 (*Identifiability*)
Let $Q(M)$ be any quantity defined on a causal model M. Q is identifiable *in a class $\boldsymbol{M}$ of models iff any two models M_1 and M_2 from $\boldsymbol{M}$ that satisfy $P_{M_1}(v) = P_{M_2}(v)$ also satisfy $Q(M_1) = Q(M_2)$. In other words, Q is identifiable if it can be determined uniquely from the probability distribution $P(v)$ of the endogenous variables V.*

The class $\boldsymbol{M}$ that we will consider when discussing identifiability will be determined by assumptions that one is willing to make about the model under study. For example, if our assumptions consist of the structure of a causal graph G_0, $\boldsymbol{M}$ will consist of all models M for which $G(M) = G_0$. If, in addition to G_0, we are also willing to make assumptions about the functional form of some mechanisms in M, $\boldsymbol{M}$ will consist of all models M that incorporate those mechanisms, and so on.

Since all the causal measures defined above invoke conditionalization on y, and since y is presumed affected by x, the antecedent of the the counterfactual y_x, we know that none of these quantities is identifiable from knowledge of the structure $G(M)$ and the data $P(v)$ alone, even under condition of no confounding. Moreover, none of these quantities determines the others in the general case. However, simple interrelationships and useful bounds can be derived for these quantities under the assumption of no-confounding, an assumption that we call *exogeneity*.

3.2 Bounds and basic relationships under exogeneity

Definition 13 (*Exogeneity*)
A variable X is said to be exogenous relative to Y in model M iff

$$P(y_x, y_{x'}|x) = P(y_x, y_{x'}) \tag{18}$$

namely, the way Y would potentially respond to conditions x or x' is independent of the actual value of X.

Eq. (18) has been given a variety of (equivalent) definitions and interpretations. Epidemiologists refer to this condition as "no-confounding" [Robins and Greenland, 1989], statisticians call it "as if randomized," and Rosenbaum and Rubin (1983) call it "ignorability." A graphical criterion ensuring exogeneity is the absence of a common ancestor of X and Y

in $G(M)$. The classical econometric criterion for exogeneity (e.g., Dhrymes (1970, p. 169) states that X be independent of the error term in the equation for Y.[13]

The importance of exogeneity lies in permitting the identification of $P(y_x)$, the *causal effect* of X on Y, since (using $x \Rightarrow (y_x = y)$)

$$P(y_x) = P(y_x|x) = P(y|x) \tag{19}$$

with similar reduction for $P(y_{x'})$.

Theorem 1 *Under condition of exogeneity, PNS is bounded as follows:*

$$\max[0, P(y|x) + P(y'|x') - 1] \leq PNS \leq \min[P(y|x), P(y'|x')] \tag{20}$$

Both bounds are sharp in the sense that for every joint distribution $P(x, y)$ there exists a model $y = f(x, u)$, with u independent of x, that realizes any value of PNS permitted by the bounds.

Proof of Theorem 1:
For any two events A and B we have the tight bounds:

$$\max[0, P(A) + P(B) - 1] \leq P(A, B) \leq \min[P(A), P(B)] \tag{21}$$

Eq. (20) follows from (21) using $A = y_x, B = y'_{x'}, P(y_x) = P(y|x)$ and $P(y'_{x'}) = P(y'|x')$ $\square$

Clearly, if exogeneity cannot be ascertained, then PNS is bound by inequalities similar to those of Eq. (20), with $P(y_x)$ and $P(y'_{x'})$ replacing $P(y|x)$ and $P(y'|x')$, respectively.

Theorem 2 *Under condition of exogeneity, the probabilities PN, PS, and PNS are related to each other as follows:*

$$PN = \frac{PNS}{P(y|x)} \tag{22}$$

$$PS = \frac{PNS}{1 - P(y|x')} \tag{23}$$

Thus, the bounds for PNS in Eq. (20) provide corresponding bounds for PN and PS.

The resulting bounds for PN

$$\frac{\max[0, P(y|x) + P(y'|x') - 1]}{P(y|x)} \leq PN \leq \frac{\min[P(y|x), P(y'|x')]}{P(y|x)} \tag{24}$$

have significant implications relative to both our ability to identify PN by experimental studies and the feasibility of defining PN in stochastic causal models. Replacing the conditional probabilities with causal effects (licensed by exogeneity), Eq. (24) implies the following:

[13]This criterion has been the subject of relentless objections by modern econometricians [Engle et al., 1983; Hendry, 1995; Imbens, 1997], but see Aldrich (1993) and Galles and Pearl (1998) for a reconciliatory perspective on this controversy.

Corollary 1 *Let $P(y_x)$ and $P(y'_{x'})$ be the causal effects established in an experimental study. For any point p in the range*

$$\frac{\max[0, P(y_x) + P(y'_{x'}) - 1]}{P(y_x)} \leq p \leq \frac{\min[P(y_x), P(y'_{x'})]}{P(y_x)} \tag{25}$$

we can find a causal model M that agrees with $P(y_x)$ and $P(y'_{x'})$ and for which $PN = p$.

This corollary implies that probabilities of causation cannot be defined uniquely in stochastic (non-Laplacian) models where, for each u, $Y_x(u)$ is specified in probability $P(Y_x(u) = y)$ instead of a single number.[14] (See Example-1, Section 4.1.)

Proof of Theorem 2:

Using $x \Rightarrow (y_x = y)$, we can write $x \wedge y_x = x \wedge y$, and obtain

$$\begin{aligned} PN = P(y'_{x'}|x, y) &= P(y'_{x'}, x, y)/P(x, y) && (26)\\ &= P(y'_{x'}, x, y_x)/P(x, y) && (27)\\ &= P(y'_{x'}, y_x)P(x)/P(x, y) && (28)\\ &= \frac{PNS}{P(y|x)} && (29) \end{aligned}$$

which establishes Eq. (22). Eq. (23) follows by identical steps. □

For completion, we note the relationship between PNS and the probabilities of enablement and disablement:

$$PD = \frac{P(x)\ PNS}{P(y)}, \quad PE = \frac{P(x')\ PNS}{P(y')} \tag{30}$$

3.3 Identifiability under monotonicity and exogeneity

Before attacking the general problem of identifying the counterfactual quantities in Eqs. (12)–(14) it is instructive to treat a special condition, called *monotonicity*, which is often assumed in practice, and which renders these quantities identifiable. The resulting probabilistic expressions will be recognized as familiar measures of causation that often appear in the literature.

Definition 14 (*Monotonicity*)
A variable Y is said to be monotonic relative to variable X in a causal model M iff the function $Y_x(u)$ is monotonic in x for all u. Equivalently, Y is monotonic relative to X iff

$$y'_x \wedge y_{x'} = \mathit{false} \tag{31}$$

[14]Robins and Greenland (1989), who used a stochastic model of $Y_x(u)$, defined the probability of causation as

$$PN(u) = [P(y|x, u) - P(y|x', u)]/P(y|x, u)$$

instead of the counterfactual definition in Eq. (12).

Monotonicity expresses the assumption that a change from $X = false$ to $X = true$ cannot, under any circumstance make Y change from *true* to *false*.[15] In epidemiology, this assumption is often expressed as "no prevention," that is, no individual in the population can be helped by exposure to the risk factor. Angrist, Imbens, and Rubin (1996) used this assumption to identify treatment effects from studies involving non-compliance (see also Balke and Pearl (1997)). Glymour (1998) and Cheng (1997) resort to this assumption in using disjunctive or conjunctive relationships between causes and effects, excluding functions such as exclusive-or, or parity.

Theorem 3 (*Identifiability under exogeneity and monotonicity*)
If X is exogenous and Y is monotonic relative to X, then the probabilities PN, PS, and PNS are all identifiable, and are given by Eqs. (22)–(23) with

$$PNS = P(y|x) - P(y|x') \tag{32}$$

The r.h.s. of (32) is called "risk-difference" in epidemiology, and is also misnomered "attributable risk" [Hennekens and Buring, 1987, p. 87].

From (22) we see that the probability of necessity, PN, is identifiable and given by the *excess-risk-ratio*

$$PN = [P(y|x) - P(y|x')]/P(y|x) \tag{33}$$

often misnomered as the *attributable fraction* [Schlesselman, 1982], *attributable-rate percent* [Hennekens and Buring, 1987, p. 88], *attributed fraction for the exposed* [Kelsey *et al.*, 1996, p. 38], or *attributable proportion* [Cole, 1997]. Taken literally, the ratio presented in (33) has nothing to do with attribution, since it is made up of statistical terms and not of causal or counterfactual relationships. However, the assumptions of exogeneity and monotonicity together enable us to translate the notion of attribution embedded in the definition of PN (Eq. (12)) into a ratio of purely statistical associations. This suggests that exogeneity and monotonicity were tacitly assumed by authors who proposed or derived Eq. (33) as a measure for the "fraction of exposed cases that are attributable to the exposure."

Robins and Greenland (1989) have analyzed the identification of PN under the assumption of stochastic monotonicity (i.e., $P(Y_x(u) = y) > P(Y_{x'}(u) = y)$) and have shown that this assumption is too weak to permit such identification; in fact, it yields the same bounds as in Eq. (24). This indicates that stochastic monotonicity imposes no constraints whatsoever on the functional mechanisms that mediate between X and Y.

The expression for PS (Eq. (23)), is likewise quite revealing

$$PS = [P(y|x) - P(y|x')]/[1 - P(y|x')], \tag{34}$$

as it coincides with what epidemiologists call the "relative difference" [Shep, 1958], which is used to measure the *susceptibility* of a population to a risk factor x. Susceptibility is defined

[15] Our analysis remains invariant to complementing x or y (or both), hence, the general condition of monotonicity should read: either $y'_x \wedge y_{x'} = false$ or $y'_{x'} \wedge y_x = false$. For simplicity, however, we will adhere to the definition in Eq. (31).

as the proportion of persons who possess "an underlying factor sufficient to make a person contract a disease following exposure" [Khoury *et al.*, 1989]. PS offers a formal counterfactual interpretation of susceptibility, which sharpens this definition and renders susceptibility amenable to systematic analysis. Khoury et al. (1989) have recognized that susceptibility in general is not identifiable, and have derived Eq. (34) by making three assumptions: no confounding, monotonicity,[16] and independence (i.e., assuming that susceptibility to exposure is independent of susceptibility to background not involving exposure). This last assumption is often criticized as untenable, and Theorem 3 assures us that independence is in fact unnecessary; Eq. (34) attains its validity through exogeneity and monotonicity alone.

Eq. (34) also coincides with what Cheng calls "causal power" (1997), namely, the effect of x on y after suppressing "all other causes of y." The counterfactual definition of PS, $P(y_x|x', y')$, suggests another interpretation of this quantity. It measures the probability that setting x would produce y in a situation where x and y are in fact absent. Conditioning on y' amounts to selecting (or hypothesizing) only those worlds in which "all other causes of y" are indeed suppressed.

It is important to note, however, that the simple relationships among the three notions of causation (Eqs. 22–23) only hold under the assumption of exogeneity; the weaker relationship of Eq. (16) prevails in the general, non-exogenous case. Additionally, all these notions of causation are defined in terms of the global relationships $Y_x(u)$ and $Y_{x'}(u)$ which is too crude to fully characterize the many nuances of causation; the detailed structure of the causal model leading from X to Y is often needed to explicate more refined notions, such as "actual cause," (see Section 6).

Proof of Theorem 3:
Writing $y_{x'} \vee y'_{x'} = true$, we have

$$y_x = y_x \wedge (y_{x'} \vee y'_{x'}) = (y_x \wedge y_{x'}) \vee (y_x \wedge y'_{x'}) \tag{35}$$

and

$$y_{x'} = y_{x'} \wedge (y_x \vee y'_x) = (y_{x'} \wedge y_x) \vee (y_{x'} \wedge y'_x) = y_{x'} \wedge y_x \tag{36}$$

since monotonicity entails $y_{x'} \wedge y'_x = \mathit{false}$. Substituting (36) into (35) yields

$$y_x = y_{x'} \vee (y_x \wedge y'_{x'}) \tag{37}$$

Taking the probability of (37), and using the disjointness of $y_{x'}$ and $y'_{x'}$, we obtain

$$P(y_x) = P(y_{x'}) + P(y_x, y'_{x'})$$

or

$$P(y_x, y'_{x'}) = P(y_x) - P(y_{x'}) \tag{38}$$

Eq. (38) together with the assumption of exogeneity (Eq. (19)) establish Eq. (32). □

[16] Monotonicity is not mentioned in [Khoury *et al.*, 1989], but it must have been assumed implicitly to make their derivations valid.

3.4 Identifiability under monotonicity and non-exogeneity

The relations established in Theorems 1–3 were based on the assumption of exogeneity. In this section, we relax this assumption and consider cases where the effect of X on Y is confounded, i.e., $P(y_x) \neq P(y|x)$. In such cases $P(y_x)$ may still be estimated by auxiliary means (e.g., through adjustment of certain covariates, or through experimental studies) and the question is whether this added information can render the probability of causation identifiable. The answer is affirmative.

Theorem 4 *If Y is monotonic relative to X, then PNS, PN, PS are identifiable whenever the causal effect $P(y_x)$ is identifiable and are given by*

$$PNS = P(y_x, y'_{x'}) = P(y_x) - P(y_{x'}) \tag{39}$$

$$PN = P(y'_{x'}|x, y) = \frac{P(y) - P(y_{x'})}{P(x, y)} \tag{40}$$

$$PS = P(y_x|x', y') = \frac{P(y_x) - P(y)}{P(x', y')} \tag{41}$$

To appreciate the difference between Eqs. (40) and (33) we can expand $P(y)$ and write

$$\begin{aligned} PN &= \frac{P(y|x)P(x) + P(y|x')P(x') - P(y_{x'})}{P(y|x)P(x)} \\ &= \frac{P(y|x) - P(y|x')}{P(y|x)} + \frac{P(y|x') - P(y_{x'})}{P(x, y)} \end{aligned} \tag{42}$$

The first term on the r.h.s. of (42) is the familiar excess-risk-ratio as in (33), and represents the value of PN under exogeneity. The second term represents the correction needed to account for X's non-exogeneity, i.e. $P(y_{x'}) \neq P(y|x')$.

Eqs. (39)–(41) thus provide more refined measures of causation, which can be used in situations where the causal effect $P(y_x)$ can be identified through auxiliary means (see Example 4, Section 4.4). Note however that these measures are no longer governed by the simple relationships given in Eqs. (22)–(23). Instead, the governing relation is Eq. (16).

Remarkably, since PS and PN must be non-negative, Eqs. (40)-(41) provide a simple necessary test for the assumption of monotonicity

$$P(y_x) \geq P(y) \geq P(y_{x'}) \tag{43}$$

which strengthen the standard inequalities

$$P(y_x) \geq P(x, y), \quad P(y_{x'}) \geq P(x', y)$$

It can be shown that these inequalities are in fact sharp, that is, every combination of experimental and nonexperimental data that satisfy these inequalities can be generated from some causal model in which Y is monotonic in X. That the commonly made assumption of "no-prevention" is not entirely exempt from empirical scrutiny should come as a relief to many epidemiologists. Alternatively, if the no-prevention assumption is theoretically unassailable, the inequalities of Eq. (43) can be used for testing the compatibility of the experimental and non-experimental data, namely, whether subjects used in clinical trials are representative of the target population, characterized by the joint distribution $P(x,y)$.

Proof of Theorem 4:

Eq. (39) was established in (38). To prove (41), we write

$$P(y_x|x',y') = \frac{P(y_x,x',y')}{P(x',y')} = \frac{P(y_x,x',y'_{x'})}{P(x',y')} \tag{44}$$

because $x' \wedge y' = x' \wedge y'_{x'}$ (by consistency). To calculate the numerator of (44), we conjoin (37) with x'

$$x' \wedge y_x = (x' \wedge y_{x'}) \vee (y_x \wedge y'_{x'} \wedge x')$$

and take the probability on both sides, which gives (since $y_{x'}$ and $y'_{x'}$ are disjoint)

$$\begin{aligned} P(y_x, y'_{x'}, x') &= P(x', y_x) - P(x', y_{x'}) \\ &= P(x', y_x) - P(x', y) \\ &= P(y_x) - P(x, y_x) - P(x', y) \\ &= P(y_x) - P(x, y) - P(x', y) \\ &= P(y_x) - P(y) \end{aligned}$$

Substituting in (44), we finally obtain

$$P(y_x|x',y') = \frac{P(y_x) - P(y)}{P(x',y')}$$

which establishes (41). Eq. (40) follows through identical steps. □

One common class of models which permits the identification of $P(y_x)$ under conditions of non-exogeneity is called *Markovian.*

Definition 15 (*Markovian models*)
A causal model M is said to be Markovian if the graph $G(M)$ associated with M is acyclic, and if the exogenous factors u_i are mutually independent. A model is semi-Markovian iff $G(M)$ is acyclic and the exogenous variables are not necessarily independent. A causal model is said to be positive-Markovian if it is Markovian and $P(v) > 0$ for every v.

It is shown in Pearl (1993, 1995) that for every two variables, X and Y, in a positive-Markovian model M, the causal effect $P(y_x)$ is identifiable and is given by

$$P(y_x) = \sum_{pa_X} P(y|pa_X, x) P(pa_X) \tag{45}$$

where pa_X are (realizations of) the *parents* of X in the causal graph associate with M (see also Spirtes et al. (1993) and Robins (1986)). Thus, we can combine Eq. (45) with Theorem 4 and obtain a concrete condition for the identification of the probability of causation.

Corollary 2 *If in a positive-Markovian model M, the function $Y_x(u)$ is monotonic, then the probabilities of causation PNS, PS and PN are identifiable and are given by Eqs. (39)–(41), with $P(y_x)$ given in Eq. (45).*

A broader identification condition can be obtained through the use of the back-door and front-door criteria [Pearl, 1995], which are applicable to semi-Markovian models. These were further generalized in Galles and Pearl (1995)[17] and lead to the following corollary:

Corollary 3 *Let **GP** be the class of semi-Markovian models that satisfy the graphical criterion of Galles and Pearl (1995). If $Y_x(u)$ is monotonic, then the probabilities of causation PNS, PS and PN are identifiable in **GP** and are given by Eqs. (39)–(41), with $P(y_x)$ determined by the topology of $G(M)$ through the GP criterion.*

4 Examples and Applications

4.1 Example-1: Betting against a fair coin

We must bet heads or tails on the outcome of a fair coin toss; we win a dollar if we guess correctly, lose if we don't. Suppose we bet heads and we win a dollar, without glancing at the outcome of the coin, was our bet a necessary cause (respectively, sufficient cause, or both) for winning?

Let x stand for "we bet on heads," y for "we win a dollar," and u for "the coin turned up heads." The functional relationship between y, x and u is

$$y = (x \wedge u) \vee (x' \wedge u') \tag{46}$$

which is not monotonic but nevertheless permits us to compute the probabilities of causation from the basic definitions of Eqs. (12)–(14). To exemplify,

$$PN = P(y'_{x'}|x, y) = P(y'_{x'}|u) = 1$$

because $x \wedge y \Rightarrow u$, and $Y_{x'}(u) =$ *false*. In words, knowing the current bet (x) and current win (y) permits us to infer that the coin outcome must have been a head (u), from which we can further deduce that betting tails (x') instead of heads, would have resulted in a loss. Similarly,

$$PS = P(y_x|x', y') = P(y_x|u) = 1$$

because $x' \wedge y' \Rightarrow u$, and

$$\begin{aligned} PNS &= P(y_x, y'_{x'}) \\ &= P(y_x, y'_{x'}|u)P(u) + P(y_x, y'_{x'}|u')P(u') \\ &= 1\frac{1}{2} + 0\frac{1}{2} = \frac{1}{2} \end{aligned}$$

[17] Galles and Pearl (1995) provide an efficient method of deciding from the graph $G(M)$ whether $P(y_x)$ is identifiable and, if the answer is affirmative, deriving the expression for $P(y_x)$.

We see that betting heads has 50% chance of being a necessary-and-sufficient cause of winning. Still, once we win, we can be 100% sure that our bet was necessary for our win, and once we lose (say on betting tails) we can be 100% sure that betting heads would have been sufficient for producing a win. The empirical content of such counterfactuals is discussed in Appendix A.

Note that these counterfactual quantities cannot be computed from the joint probability of X and Y without knowledge of the functional relationship in Eq. (46) which tells us the (deterministic) policy by which a win or a loss is decided. This can be seen, for instance, from the conditional probabilities and causal effects associated with this example

$$P(y|x) = P(y|x') = P(y_x) = P(y_{x'}) = P(y) = \frac{1}{2}$$

because identical probabilities would be generated by a random payoff policy in which y is functionally independent of x, say by a bookie who watches the coin and ignores our bet. In such a random policy, the probabilities of causation PN, PS and PNS are all zero. Thus, according to our definition of identifiability (Definition 12), if two models agree on P and do not agree on a quantity Q, then Q is not identifiable. Indeed, the bounds delineated in Theorem 1 (Eq. (20)) read $0 \leq PNS \leq \frac{1}{2}$, meaning that the three probabilities of causation cannot be determined from statistical data on X and Y alone, not even in a controlled experiment; knowledge of the functional mechanism is required, as in Eq. (46).

It is interesting to note that whether the coin is tossed before or after the bet has no bearing on the probabilities of causation as defined above. This stands in contrast with some theories of probabilistic causality which attempt to avoid deterministic mechanisms by conditioning all probabilities on "the state of the world just before" the occurrence of the cause in question (x) (e.g., [Good, 1961]). In the betting story above, the intention is to condition all probabilities on the state of the coin (u), but it is not fulfilled if the coin is tossed after the bet is placed. Attempts to enrich the conditioning set with events occurring after the cause in question have led back to deterministic relationships involving counterfactual variables (see [Cartwright, 1989; Eells, 1991]).

One may argue, of course, that if the coin is tossed after the bet, then it is not at all clear what our winning would be had we bet differently; merely uttering our bet could conceivably affect the trajectory of the coin [Dawid, 1997]. This objection can be diffused by placing x and u in two remote locations and tossing the coin a split second after the bet is placed, but before any light ray could arrive from the betting room to the coin-tossing room. In such hypothetical situation the counterfactual statement: "our winning would be different had we bet differently" is rather compelling, even though the conditioning event (u) occurs after the cause in question (x). We conclude that temporal descriptions such as "the state of the world just before x" cannot be used to properly identify the appropriate set of conditioning events (u) in a problem; a deterministic model of the mechanisms involved is needed for such identification.

4.2 Example-2: The firing squad

Consider a 2-man firing squad (see Figure 2) in which A and B are riflemen, C is the squad's Captain who is waiting for the court order, U, and T is a condemned prisoner. Let u be

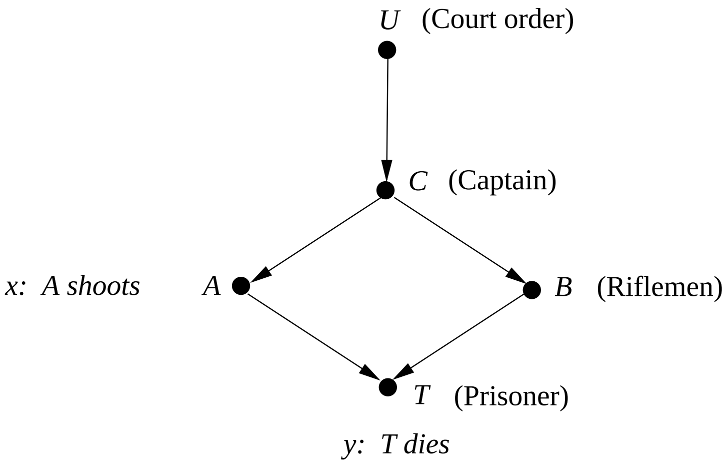

Figure 2: Causal relationships in the 2-man firing squad example.

the proposition that the court has ordered an execution, x the proposition stating that A pulled the trigger, and y that T is dead. Assume that $P(u) = \frac{1}{2}$, that A and B are perfectly accurate marksmen who are alert and law abiding, and that T is not likely to die from fright or other extraneous causes. We wish to compute the probability that x was a necessary (or sufficient, or both) cause for y (i.e., PN, PS, and PNS).

Definitions (7)–(9) permit us to compute these probabilities directly from the given causal model, since all functions and all probabilities are specified, with the truth value of each variable tracing that of U. Accordingly, we can write[18]

$$\begin{aligned} P(y_x) &= P(Y_x(u) = true)P(u) + P(Y_x(u') = true)P(u') \\ &= \frac{1}{2}(1+1) = 1 \end{aligned} \tag{47}$$

Similarly, we have

$$\begin{aligned} P(y_{x'}) &= P(Y_{x'}(u) = true)P(u) + P(Y_{x'}(u') = true)P(u') \\ &= \frac{1}{2}(1+0) = \frac{1}{2} \end{aligned} \tag{48}$$

To compute PNS, we need to evaluate the probability of the joint event $y_{x'} \wedge y_x$. Considering that these two events are jointly true only when $U = true$, we have

$$\begin{aligned} PNS &= P(y_x, y_{x'}) \\ &= P(y_x, y_{x'}|u)P(u) + P(y_x, y_{x'}|u')P(u') \\ &= \frac{1}{2}(1+0) = \frac{1}{2} \end{aligned} \tag{49}$$

The calculation of PS and PN, likewise, are simplified by the fact that each of the conditioning events, $x \wedge y$ for PN and $x' \wedge y'$ for PS, is true in only one state of U. We thus

[18]Recall that $P(Y_x(u') = true)$ involves the submodel M_x, in which X is set to *true* independently of U. Thus, although under condition u' the captain has not given a signal, the potential outcome $Y_x(u')$ calls for hypothesizing rifleman-A pulling the trigger (x) despite a court order to stay the execution.

have

$$PN = P(y'_{x'}|x, y) = P(y'_{x'}|u) = 0$$

reflecting the fact that, once the court orders an execution (u), T will die (y) from the shot of rifleman B, even if A refrains from shooting (x'). Indeed, upon learning of T's death, we can categorically state that rifleman-A's shot was *not* a necessary cause of the death.

Similarly,

$$PS = P(y_x|x', y') = P(y_x|u') = 1$$

matching our intuition that a shot fired by an expert marksman would be sufficient for causing the death of T, regardless of the court decision.

Note that Theorems 1 and 2 are not applicable to this example, because x is not exogenous; events x and y have a common cause (the Captain's signal) which renders $P(y|x') = 0 \neq P(y_{x'}) = \frac{1}{2}$. However, the monotonicity of Y (in x) permits us to compute PNS, PS and PN from the joint distribution $P(x, y)$ (using Eq. (39)–(41)), instead of consulting the basic model. Indeed, writing

$$P(x, y) = P(x', y') = \frac{1}{2} \tag{50}$$

$$P(x, y') = P(x', y) = 0 \tag{51}$$

we obtain

$$PN = \frac{P(y) - P(y_{x'})}{P(x, y)} = \frac{\frac{1}{2} - \frac{1}{2}}{\frac{1}{2}} = 0 \tag{52}$$

$$PS = \frac{P(y_x) - P(y)}{P(x', y')} = \frac{1 - \frac{1}{2}}{\frac{1}{2}} = 1 \tag{53}$$

as expected.

4.3 Example-3: The effect of radiation on leukemia

Consider the following data (adapted from Finkelstein and Levin[19] (1990)) comparing leukemia deaths in children in Southern Utah with high and low exposure to radiation from fallout from nuclear tests in Nevada. Given these data, we wish to estimate the probabilities that high exposure to radiation was a necessary (or sufficient or both) cause of death due to leukemia.

Assuming that exposure to nuclear radiation had no remedial effect on any individual in the study (i.e., monotonicity), the process can be modeled by a simple disjunctive mechanism represented by the equation

$$y = f(x, u, q) = (x \wedge q) \vee u \tag{54}$$

where u represents "all other causes" of y, and q represents all "enabling" mechanisms that must be present for x to trigger y. Assuming q and u are both unobserved, the question we

[19]The data in Finkelstein and Levin (1990) are given in person-year units. For the purpose of illustration we have converted the data to absolute numbers (of deaths and non-deaths) assuming a 10-year observation period.

		Exposure	
		High	**Low**
		x	x'
Deaths	y	30	16
Survivals	y'	69,130	59,010

Table 1:

ask is under what conditions we can identify the probability of causation, PNS, PN, and PS, from the joint distribution of X and Y.

Since Eq. (54) is monotonic in x, Theorem 3 states that all three quantities would be identifiable provided X is exogenous, namely, x should be independent of q and u. Under this assumption, Eqs. (32)–(34) further permit us to compute the probabilities of causation from frequency data. Taking fractions to represent probabilities, the data in Table 1 imply the following numerical results

$$PNS = P(y|x) - P(y|x') = \frac{30}{30 + 69,130} - \frac{16}{16 + 59,010} = .0001625 \tag{55}$$

$$PN = \frac{PNS}{P(y|x)} = \frac{PNS}{30/(30 + 69,130)} = .37535 \tag{56}$$

$$PS = \frac{PNS}{1 - P(y|x')} = \frac{PNS}{1 - 16/(16 + 59,010)} = .0001625 \tag{57}$$

Statistically, these figures mean: There is a 1.625 in ten thousand chance that a randomly chosen child would both die of leukemia if exposed and survive if not exposed. There is a 37.535% chance that a child who died from leukemia after exposure would have survived had he/she not been exposed. There is a 1.625 in ten-thousand chance that any unexposed surviving child would have died of leukemia had he/she been exposed.

Glymour (1998) analyzes this example with the aim of identifying the probability $P(q)$ (Cheng's "causal power") which coincides with PS (see Lemma 3). Glymour concludes that $P(q)$ is identifiable and is given by Eq. (34), provided x, u, and q are mutually independent. Our analysis shows that Glymour's result can be generalized in several ways. First, since Y is monotonic in X, the validity of Eq. (34) is assured even when q and u are dependent, because exogeneity merely requires independence between x and $\{u, q\}$ jointly. This is important in epidemiological settings, because an individual's susceptibility to nuclear radiation is likely to be associated with his/her susceptibility to other potential causes of leukemia (e.g., natural kinds of radiation).

Second, Theorem 2 assures us that the relationships between PN, PS and PNS (Eqs. (22)–(23)), which Glymour derives for independent q and u, should remain valid even when u and q are dependent.

Finally, Theorem 4 assures us that PN and PS are identifiable even when x is not independent of $\{u, q\}$, provided only that the mechanism of Eq. (54) is embedded in a larger causal structure which permits the identification of $P(y_x)$. For example, assume that exposure to

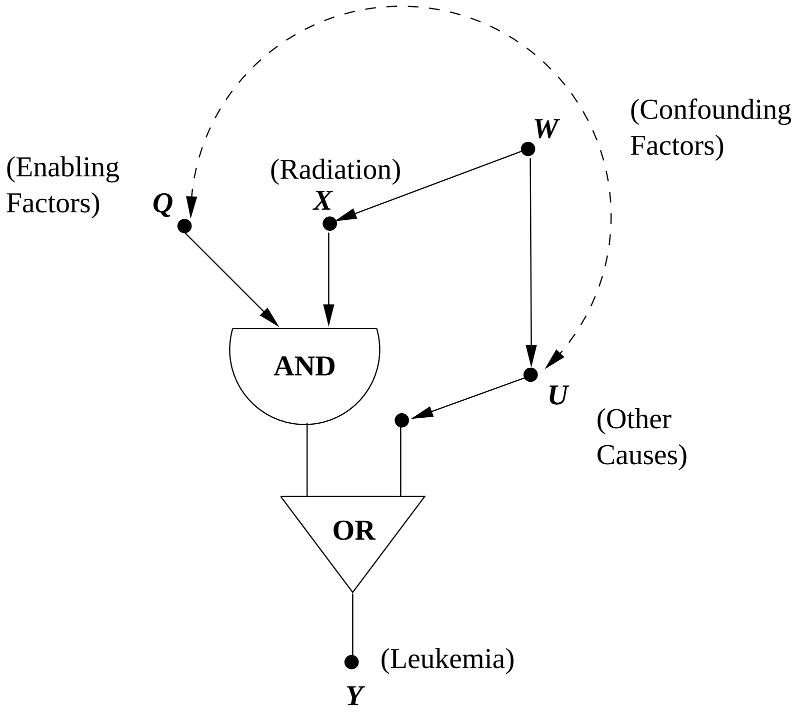

Figure 3: Causal relationships in the Radiation-Leukemia example. W represents confounding factors.

nuclear radiation (x) is suspect of being associated with terrain and altitude, which are also factors in determining exposure to cosmic radiation. A model reflecting such consideration is depicted in Figure 3, where W represents factors affecting both X and U. A natural way to correct for possible confounding bias in the causal effect of X on Y would be to adjust for W, that is, to calculate $P(y_x)$ using the adjustment formula

$$P(y_x) = \sum_w P(y|x,w)P(w) \tag{58}$$

(instead of $P(y|x)$) where the summation runs over levels of W. This adjustment formula, which follows from Eq. (45), is correct regardless of the mechanisms mediating X and Y, provided only that W represents *all* common factors affecting X and Y [Pearl, 1995]. Theorem 4 instructs us to evaluate PN and PS by substituting (58) into Eqs. (40) and (41), respectively, and it assures us that the resulting expressions constitute consistent estimates of PN and PS. This consistency is guaranteed jointly by the assumption of monotonicity and by the (assumed) topology of the causal graph.

Note that monotonicity as defined in Eq. (31) is a global property of all pathways between x and y. The causal model may include several nonmonotonic mechanisms along these pathways without affecting the validity of (31). Arguments for the validity of monotonicity, however, must be based on substantive information, as it is not testable in general. For example, Robins and Greenland (1989) argue that exposure to nuclear radiation may conceivably be of benefit to some individuals, since such radiation is routinely used clinically in treating cancer patients.

4.4 Example-4: Legal responsibility from experimental and non-experimental data

A lawsuit is filed against the manufacturer of drug x, charging that the drug is likely to have caused the death of Mr. A, who took the drug to relieve symptom S associated with disease D. The manufacturer claims that experimental data on patients with symptom S show conclusively that drug x may cause only negligible increase in death rates. The plaintiff argues, however, that the experimental study is of little relevance to this case, because it represents the effect of the drug on *all* patients, not on patients like Mr. A who actually died while using drug x. Moreover, argues the plaintiff, Mr. A is unique in that he used the drug on his own volition, unlike subjects in the experimental study who took the drug to comply with experimental protocols. To support this argument, the plaintiff furnishes non-experimental data indicating that most patients who chose drug x would have been alive if it were not for the drug. The manufacturer counter-argues by stating that: (1) counterfactual speculations regarding whether patients would or would not have died are purely metaphysical and should be avoided [Dawid, 1997], and (2) non-experimental data should be dismissed a priori, on the ground that such data may be highly biased; for example, incurable terminal patients might be more inclined to use drug x if it provides them greater symptomatic relief. The court must now decide, based on both the experimental and non-experimental studies, what the probability is that drug x was in fact the cause of Mr. A's death.

The (hypothetical) data associated with the two studies are shown in Table 2 below.

Experimental

		x	x'
Deaths	y	16	14
Survivals	y'	984	986

Non-Experimental

		x	x'
Deaths	y	2	28
Survivals	y'	998	972

Table 2:

The experimental data provide the estimates

$$P(y_x) = 16/1000 = 0.016 \tag{59}$$

$$P(y_{x'}) = 14/1000 = 0.014 \tag{60}$$

The non-experimental data provide the estimates

$$P(y) = 30/2000 = 0.015 \tag{61}$$

$$P(y,x) = 2/2000 = 0.001 \tag{62}$$

Assuming that drug x can only cause, never prevent, death, Theorem 4 is applicable and Eq. (40) gives

$$PN = \frac{P(y) - P(y_{x'})}{P(y,x)} = \frac{0.015 - 0.014}{0.001} = 1.00 \tag{63}$$

Thus, the plaintiff was correct; barring sampling errors, the data provide us with 100% assurance that drug x was in fact responsible for the death of Mr. A. Note that a straightforward use of the experimental excess-risk-ratio would yield a much lower (and incorrect) result:

$$\frac{P(y_x) - P(y_{x'})}{P(y_x)} = \frac{0.016 - 0.014}{0.016} = 0.125 \tag{64}$$

Evidently, what the experimental study does not reveal is that, given a choice, terminal patients stay away from drug x. Indeed, if there were any terminal patients who would choose x (given the choice), then the control group (x') would have included some such patients (due to randomization) and then the proportion of deaths among the control group $P(y_{x'})$ should have been higher than $P(x', y)$, the population proportion of terminal patients avoiding x. However, the equality $P(y_{x'}) = P(y, x')$ tells us that no such patients were included in the control group, hence (by randomization) no such patients exist in the population at large and, therefore, none of the patients who freely chose drug x was a terminal case; all were susceptible to x.

The numbers in Table 2 were obviously contrived to represent an extreme case, so as to facilitate a qualitative explanation of the validity of Eq. (40). Nevertheless, it is instructive to note that a combination of experimental and non-experimental studies may unravel what experimental studies alone will not reveal and, in addition, that such combination may provide a test for the assumption of no-prevention, as outlined in Section 3.4 (Eq. (43)).

5 Identification in Non-monotonic Models

In this section we discuss the identification of probabilities of causation without making the monotonicity assumption. We will assume that we are given a causal model M in which all functional relationships are known, but since the exogenous variables U are not observed, their distributions are not known.

A straightforward way to identify any causal or counterfactual quantity (including PN, PS and PNS) would be to infer the probability distribution of the exogenous variables – that would amount to inferring the entire model, from which all quantities can be computed. Thus, our first step would be to study under what conditions the function $P(u)$ can be identified.

If M is Markovian, the problem can be analyzed by considering each parents-child family separately. Consider any arbitrary equation in M

$$\begin{aligned} y &= f(pa_Y, u_Y) \\ &= f(x_1, x_2, \ldots, x_k, u_1, \ldots, u_m) \end{aligned} \tag{65}$$

where $U_Y = \{U_1, \ldots, U_m\}$ is the set of exogenous, possibly dependent variables that appear in the equation for Y. In general, the domain of U_Y can be arbitrary, discrete, or continuous, since these variables represent unobserved factors that were omitted from the model. However, since the observed variables are binary, there is only a finite number ($2^{(2^k)}$) of functions from PA_Y to Y and, for any point $U_Y = u$, only one of those function is realized. This defines a partition of the domain of U_Y into a set S of equivalence classes, where each equivalence class $s \in S$ induces the same function $f^{(s)}$ from PA_Y to Y. Thus, as u varies over its domain, a set S of such functions is realized, and we can regard S as a new exogenous variable, whose values are the set $\{f^{(s)} : s \in S\}$ of functions from PA_Y to Y that are realizable in U_Y. The number of such functions will usually be smaller than $2^{(2^k)}$.[20]

For example, consider the model described in Figure 3. As the exogenous variables (q, u) vary over their respective domains, the relation between X and Y spans three distinct functions

$$Y = \textit{true}, \quad Y = \textit{false}, \text{ and } Y = X$$

The fourth possible function, $Y = \textit{not-}X$, is never realized because $f_Y(\cdot)$ is monotonic. The cells (q, u) and (q', u) induce the same function between X and Y, hence they belong to the same equivalence class.

If we are given the distribution $P(u_Y)$, we can compute the distribution $P(s)$ and this will determine the conditional probabilities $P(y|pa_Y)$ by summing $P(s)$ over all those functions $f^{(s)}$ that map pa_Y into the value *true*,

$$P(y|pa_Y) = \sum_{s:f^{(s)}(pa_Y) \;=\; true} P(s) \tag{66}$$

To insure model identifiability it is sufficient that we can invert the process and determine $P(s)$ from $P(y|pa_Y)$. If we let the set of conditional probabilities $P(y|pa_Y)$ be represented by a vector $\boldsymbol{p}$ (of 2^k), and $P(s)$ by a vector $\boldsymbol{q}$, then the relation between $\boldsymbol{q}$ is $\boldsymbol{p}$ is linear and can be represented as a matrix multiplication [Balke and Pearl, 1994b]

$$\boldsymbol{p} = \boldsymbol{R}\boldsymbol{q} \tag{67}$$

where $\boldsymbol{R}$ is a 0-1 matrix, with dimension $2^k \times |S|$. Thus, a sufficient condition for identification is simply that $\boldsymbol{R}$, together with the normalizing equation $\sum_j \boldsymbol{q}_j = 1$, be invertible.

In general, $\boldsymbol{R}$ will not be invertible because the dimensionality of $\boldsymbol{q}$ can be much larger than that of $\boldsymbol{p}$. However, in many cases, such as the Noisy-OR mechanism

$$Y = U_0 \bigvee_{i=1,\ldots,k} (X_i \wedge U_i), \tag{68}$$

symmetry permits $\boldsymbol{q}$ to be identified from $P(y|pa_Y)$ even when the exogenous variables $U_0, U_1, \ldots, U_k$ are not independent. This can be seen by noting that every point u for which $U_0 =$ *false* defines a unique function $f^{(s)}$ because, if T is the set of indices i for which U_i is true, the relationship between PA_Y and Y becomes

$$Y = U_0 \bigvee_{i \in T} X_i \tag{69}$$

and, for $U_0 =$ *false*, this equation defines a distinct function for each T. The number of induced functions is $2^k + 1$, which (subtracting 1 for normalization) is exactly the number of distinct realizations of PA_Y. Moreover, it is easy to show that the matrix connecting $\boldsymbol{p}$ and $\boldsymbol{q}$ is invertible. We thus conclude that the probability of every counterfactual sentence

[20] Balke and Pearl (1994) called these S variables "response variables," and Heckerman and Shachter (1995) called them "mapping variables."

can be identified in any Markovian model composed of Noisy-OR mechanisms, regardless of whether the exogenous variables in each family are mutually independent. The same holds of course for Noisy-AND mechanisms or any combination thereof, including negating mechanisms, provided that each family consists of one type of mechanism.

To generalize this results to mechanisms other than Noisy-OR and Noisy-AND, we note that although $f_Y(\cdot)$ in this example was monotonic (in each X_i), it was the redundancy of $f_Y(\cdot)$, not its monotonicity, that ensured identifiability. The following is an example of a monotonic function for which the $\boldsymbol{R}$ matrix is not invertible

$$Y = (X_1 \wedge U_1) \vee (X_2 \wedge U_1) \vee (X_1 \wedge X_2 \wedge U_3)$$

It represents a Noisy-OR gate for $U_3 = \mathit{false}$, and becomes a Noisy-AND gate for $U_3 = \mathit{true}, U_1 = U_2 = \mathit{false}$. The number of equivalence-classes induced is six, which would require five independent equations to determine their probabilities; the data $P(y|pa_Y)$ provide only four such equations.

In contrast, the mechanism governed by the equation below, although non-monotonic, is invertible:

$$Y = XOR(X_1, XOR(U_2, \ldots, XOR(U_{k-1}, XOR(X_k, U_k))))),$$

where $XOR(*)$ stands for Exclusive-OR. This equation induces only two functions from PA_Y to Y;

$$Y = \begin{cases} XOR(X_1, ..., X_k) & if \quad XOR(U_1, \ldots, U_k) = \mathit{false} \\ \neg XOR(X_1, \ldots, X_k) & if \quad XOR(U_1, \ldots, U_k) = \mathit{true} \end{cases}$$

A single conditional probability, say $P(y|x_1, \ldots, x_k)$, would therefore suffice for computing the one parameter needed for identification: $P[XOR(U_1, \ldots, U_k) = \mathit{true}]$.

We summarize these considerations with a theorem.

Definition 16 (*Local invertability*)
A model M is said to be locally invertible if for every variable $V_i \in V$ the set of $2^k + 1$ equations

$$P(y|pa_i) = \sum_{s: f^{(s)}(pa_i) = \mathit{true}} q_i(s) \tag{70}$$

$$\sum_s q_i(s) = 1 \tag{71}$$

has a unique solution for $q_i(s)$, where each $f_i^{(s)}(pa_i)$ corresponds to the function $f_i(pa_i, u_i)$ induced by u_i in equivalence-class s.

Theorem 5 *Given a Markovian model $M = < U, V, \{f_i\} >$ in which the functions $\{f_i\}$ are known and the exogenous variables U are unobserved, if M is locally invertible, then the probability of every counterfactual sentence is identifiable from the joint probability $P(v)$.*

Proof:
If Eq. (70) has a unique solution for $q_i(s)$, we can replace U with S and obtain an equivalent model

$$M' = < S, V, \{f_i'\} > \text{ where } f_i' = f_i^{(s)}(pa_i).$$

M' together with $q_i(s)$ completely specifies a probabilistic model $< M', P(s) >$ (due to the Markov property) from which probabilities of counterfactuals are derivable by definition. $\square$

Theorem 5 provides a sufficient condition for identifying probabilities of causation, but of course does not exhaust the spectrum of assumptions that are helpful in achieving identification. In many cases we might be justified in hypothesizing additional structure on the model, for example, that the U variables entering each family are themselves independent. In such cases, additional constraints are imposed on the probabilities $P(s)$ and Eq. (70) may be solved even when the cardinality of S far exceeds the number of conditional probabilities $P(y|pa_Y)$.

6 From Necessity and Sufficiency to "Actual Cause"

6.1 The Role of Structural Information

In Section 3, we alluded to the fact that both PN and PS are global (i.e., input-output) features of a causal model, depending only on the function $Y_x(u)$, but not on the structure of the process mediating between the cause (x) and the effect (y). That such structure plays a role in causal explanation is seen in the following example.

Consider an electric circuit consisting of a light bulb and two switches, and assume that the light is turned on whenever either switch-1 or switch-2 is on. Assume further that, internally, when switch-1 is on it not only activates the light, but also disconnects switch-2 from the circuit, rendering it inoperative. From an input-output viewpoint, the light responds symmetrically to the two switches; either switch is sufficient to turn the light on. However, with both switches on, we would not hesitate to proclaim switch-1 as the "actual cause" of the current flowing in the light bulb, knowing that, internally, switch-2 is totally disconnected in this particular state of affairs. There is nothing in PN and PS that could possibly account for this asymmetry; each is based on the response function $Y_x(u)$, and is therefore oblivious to the internal workings of the circuit.

This example is isomorphic to Suppes' Desert Traveler, and belongs to a large class of counterexamples that were brought up against Lewis' counterfactual account of causation. It illustrates how an event (e.g., switch-1 being on) can be considered a cause although the effect persists in its absence. Lewis' (1986) answer to such counterexamples was to modify the counterfactual criterion and let x be a cause of y as long as there exists a counterfactual-dependence chain of intermediate variables between x to y, that is, the output of every link in the chain is counterfactually dependent on its input. Such a chain does not exist for switch-2, since it is disconnected when both switches are on.

Lewis' chain criterion retains the connection between causation and counterfactuals, but it is rather ad-hoc; after all, why should the existence of a counterfactual-dependence chain be taken as a defining test for such crucial concepts as "actual cause," by which we decide the guilt or innocence of defendants in a court of law? Another problem with Lewis' chain is its failure to capture symmetric cases of overdetermination. For example, consider two switches connected symmetrically, such that each participates equally in energizing the light bulb. In this situation, our intuition regards each of the switches as a contributory actual cause of the light, though none passes the counterfactual test and none supports a counterfactual-dependence chain in the presence of the other.

An alternative way of using counterfactuals to define actual causes is proposed in [Pearl, 1998]. An event x is defined as the "actual cause" of event y (in a world u), if x passes the standard counterfactual test (i.e., $Y_{x'}(u) = \mathit{false}$) in some mutilated model M', minimally removed from M. In the symmetric two-switch example, we declare each switch to be an actual cause of the light because the light would be off if that switch were off, when we consider a slightly mutilated circuit, one in which the other switch is disconnected from the power source. The mutilated model M', called a "causal beam," is carefully constructed in [Pearl, 1998] to ensure minimal deviation from the actual causal model M, considering the actual history of the world u.

The concept of causal sufficiency offers yet a third way of rescuing the counterfactual account of causation. Consider again the symmetric two-switch example (or the firing squad example of Section 4.2). Both switches enjoy high PS value, because each would produce light from a state (u') of darkness, namely, a state in which the other switch is off. Likewise, the shot of each riflemen in Example-2 (Section 4.2) enjoys a PS value of unity (see Eq. (53)), because each shot would cause the prisoner's death in the state u' in which the prisoner is alive, namely, the court orders no execution. Thus, if our intuition is driven by some strange mixture of sufficiency and necessity considerations, it seems plausible that we could formulate an adequate criterion for actual causation using the right mixture of PN and PS components.

Similar expectations are expressed in Hall (1998). In analyzing problems faced by the counterfactual approach, Hall makes the observation that there are two concepts of causation, only one of which is captured by the counterfactual account, and that failure to capture the second concept may well explain its clashes with intuition. Hall calls the first concept "dependence" and the second "production." In the symmetrical two-switch example (an instance of "over-determination"), intuition considers each switch to be an equal "producer" of the light, while the counterfactual account tests for "dependence" only, and fails because the light does not "depend" on either switch alone.

The notions of dependence and production closely parallel those of necessity and sufficiency, respectively. Thus, our formulation of PS could well provide the formal basis for Hall's notion of production, and serve as a step toward the formalization of actual causation. For this program to succeed, several hurdles must be overcome, the most urgent being the problems of incorporating singular event information and structural information into PS. These will be discussed next.

6.2 Singular sufficient causes

So far we have explicated the necessity and sufficiency conceptions of causation in terms of their probabilities, but not as properties of a given specific scenario, dictated by a specific state of U. This stands in contrast with standard practice of first defining truth values of sentences in each specific world, then evaluating probabilities of sentences from probabilities of worlds. Lewis (1986) counterfactual account of causation, for example, assigns a truth value to the sentence "x is a cause of y" in each specific world (u), given by the conjunction $x \wedge y \wedge y'_{x'}$. The question arises whether sentences about sufficient causation can likewise be given world-level truth values and, if they do, which worlds should provide those values, and how evidential information about those worlds should enter probability calculations.

Necessary causation can be formulated deterministically (at the world-level) in the standard counterfactual way:

Definition 17 (*Deterministic necessity*)
Event x is said to be a necessary cause of event y in a world u just in case the following hold in u:

1. $Y(u) = y$ *and* $X(u) = x$
2. $Y_{x'}(u) \neq y$ *for every* $x' = x$.

Accordingly, if additional evidence e is available about our current world, it can easily be incorporated into the evaluation of PN as follows:

$$PN(x \rightarrow y|e) = P(y'_{x'}|x, y, e)$$

where $PN(x \rightarrow y|e)$ is the probability that x was a necessary cause of y, given evidence e.

Sufficient causation, on the other hand, requires a nonstandard deterministic (i.e., world-level) formulation.

Definition 18 (*Deterministic sufficiency*)
Event x is said to be a sufficient cause of event y in a world u just in case the following hold in u:

1. $Y(u) \neq y$ *and* $X(u) \neq x$
2. $Y_x(u) = y$

In words, x is a sufficient cause for y if x would produce y (counterfactually) in world u in which x and y are absent.

The nonstandard feature of this definition lies in requiring both the explanation (x) and the explanadum (y) to be false in any world u where the former pertains to cause the latter. Thus, it appears that nothing could possibly explain (by consideration of sufficiency) events that happened to materialize in the actual world. This feature reflects, of course, our commitment to interpret sufficiency as the capacity to produce an effect and, as strange as it may sound, it is indeed impossible to talk about "x producing y" in a world (say ours) in which x and y are already true. The word "production" implies the establishment of new facts. Therefore, to test production, we must step outside our world momentarily, imagine a new world with x and y absent, apply x, and see if y sets in.

This peculiar feature of sufficiency leads to difficulties in incorporating world-specific findings into the analysis. Consider a 1-man firing squad in which rifleman A has a hit rate of 99% and the prisoner has a small chance p of dying from fear. Our analysis of Section 3 indicates that PS equals 99%, independent of p. Now suppose we find that the bullet fired hit the prisoner's leg, from which we conclude that the prisoner must have died from fear. Would this finding change our assessment of how sufficient A's shot was for causing T's death? There are grounds for arguing that it should: although, in general, a shot from a rifleman like Mr. A would be 99% sufficient for the job, this particular shot was evidently of a different type, a peculiar type that scores zero on the accuracy and sufficiency scale.

However it is not at all trivial to formalize this argument using the logical machinery at our disposal. First, to properly incorporate the new piece of evidence, e: "The bullet was found in the prisoner's leg" we need to know the structure of the causal process; the function $Y_x(u)$ in itself would be insufficient, for it does not tell us how the location of the bullet alters the chance of death. But even given the structure, say in the form of an intermediate variable denoting "Location of bullet," we cannot simply add e to the conditioning part in the expression for PS, forming $P(y_x|x', y', e)$, as we did for PN. The location of the bullet was observed in the actual world, that is, after x was enacted and y verified, while the conditioning events, x' and y', pertain to hypothetical world that existed prior to the action (x). Mixing the two without making this distinction leads to contradictions and misinterpretations. The expression $P(y_x|x', y', e)$ amounts to evaluating the probability that a living prisoner carrying a bullet in his leg would die if shot by Mr. A. This is certainly not the intended interpretation of PS and would not evaluate to zero as it should. As another example, if e stands for "bullet in the heart," which conflicts with y', we would be instructed into conditioning $P(y_x)$ on a contradictory event.

An attempt to place e in the consequent part of the counterfactual, forming $P(y_x, e|x', y')$, again does not accomplish our mission.[21] It expresses the probability that, both, the shot would be sufficient to cause death and that a living prisoner would have a bullet in his leg; still far from the probability that a shot in the leg will suffice to cause death.

These difficulties stem from dealing with the dynamic process of "production" using a syntax that does not allow explicit reference to time. Fortunately, the difficulty can be resolved even in the confines of this syntax. Since the evidence e was obtained in a world created by the action x, and since events in such worlds are governed by the submodel M_x (see Section 2.1), the proper syntax for introducing such evidence would be to condition on the subscripted symbol e_x. This leads to:

Definition 19 (*Singular-event sufficiency*)
The probability that x was a sufficient cause of y given evidence e is defined as[22]*:*

$$PS(x \rightarrow y|e) = P(y_x|e_x, x', y') \tag{72}$$

To illustrate, assume Z stands for a 2-state variable "Location of bullet," with z denoting "bullet in chest" and z' denoting "bullet not in chest." Assuming further that the flow of causation is governed by the causal chain $X \rightarrow Z \rightarrow Y$, and that a bullet would cause death if and only if it ends up in the chest. It is not hard to show that Definition 19 yields

$$PS(x \rightarrow y|z) = 1$$

[21] Related attempt to modify the consequent part is reported in Michie (1997), using an adaptation of Good's measure of causal sufficiency, Q_{suf}.

[22] Other expressions are also possible, for example, $P(y_x, e_x|x', y')$, which captures the capacity of x to produce both y and e. This expression suffers, however, from sensitivity to detail; elaborate descriptions of e would yield extremely low probabilities.

$$\begin{aligned} PS(x &\rightarrow y|z') = 0 \\ PN(x &\rightarrow y|z) = 1 \\ PN(x &\rightarrow y|z') = 1 - P(\textit{death from fear}) \end{aligned} \tag{73}$$

as expected.

The next subsection illustrates the role of singular event information in a probabilistic analysis of Suppes' desert traveler story.

6.3 Example: The Desert Traveler (after P. Suppes)

A desert traveler T has two enemies. Enemy-1 poisons T's canteen, and Enemy-2, unaware of Enemy-1's action, shoots and empties the canteen. A week later, T is found dead and the two enemies confess to action and intention. A jury must decide whose action was the cause of T's death.

Let u be the proposition that traveler's first need of drink occurred after the shot was fired. Let x and p be the propositions "Enemy-2 shot", and "Enemy-1 poisoned the water," respectively, and let y denote "T is dead." In addition to these events we will make informal use of possible exceptions to the normal story, such as T surviving the ordeal or T suspecting that the water is poisoned.

The causal model underlying the story is depicted in Figure 4. The model is completely

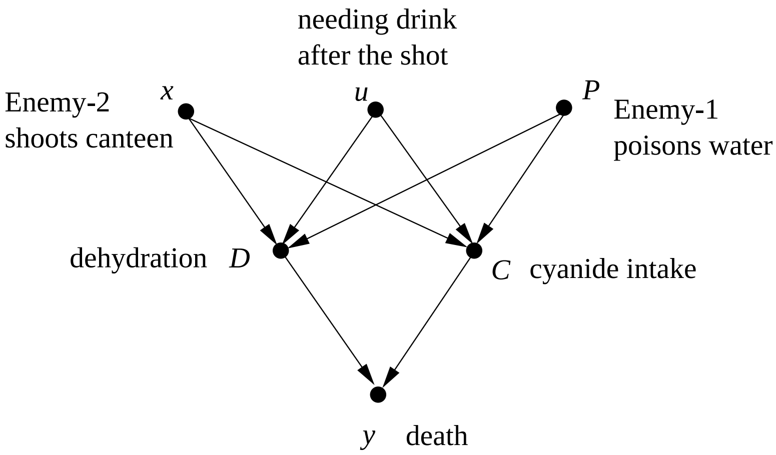

Figure 4: Causal relationships in the Desert-Traveler example.

specified through the functions $f_i(pa_i, u)$ which are not shown explicitly in Fig. 4, but are presumed to determine the value of each child variable from those of its parent variables in the graph, in accordance with our usual understanding of the story:

$$\begin{aligned} c &= p \wedge (u' \vee x') \\ d &= x \wedge (u \vee p') \\ y &= c \vee d \end{aligned}$$

(We assume that T will not survive with empty canteen (x) even after drinking some unpoisoned water before the shot ($p' \wedge u'$).)

6.3.1 Necessity and Sufficiency Ignoring Internal Structure

The global function $Y(x, p, u)$ is given by

$$y = x \vee p$$

which is symmetric in x and p.

The calculations of $PS(x \rightarrow y) = PS(p \rightarrow y)$ and $PN(p \rightarrow y) = PN(x \rightarrow y)$, can proceed directly from their definitions, without resorting to structural information.

$$PS(x \rightarrow y) = P(y_x | x', y') = 1$$

because (x', y') implies that no poison was added (p'), in which case $P(y_x)$ is 1, barring the unlikely event that T manages to survive with an empty canteen.

Similarly,

$$PN(x \rightarrow y) = P(y'_{x'} | x, y) = 0$$

If we wish to include the possibility of T surviving with either an empty canteen or a poisoned canteen, we have:

$$\begin{aligned} PS(x \rightarrow y) &= P(y_x | x', y') \\ &= 1 - P(\textit{survival with empty canteen}) \\ PS(p \rightarrow y) &= P(y_p | p', y') \\ &= 1 - P(\textit{survival with poisoned water}) \end{aligned} \quad (74)$$

Note that $PN(x \rightarrow y)$ and $PN(p \rightarrow y)$ remain zero, unaffected by the possibility of survival, because T's death (y) is taken as evidence that conditions necessary for survival did not in fact materialize (see Lemma 2).

6.3.2 Sufficiency and Necessity given Forensic Reports

Let c stand for: "Cyanide was found in T's body" and d for: "T's body showed signs of dehydration."

Incorporating the first evidence into the probability of sufficiency (Eq. 72), we have

$$PS(x \rightarrow y | c) = P(y_x | x', y', c_x)$$

The conditioning part instructs us to imagine a scenario in which Enemy-2 did not shoot, T did not die and cyanide would be found in T's body if Enemy-2 were to shoot. The one scenario which complies with these conditions is as follows: The water was poisoned, T drank the water before the time Enemy-2 was about to shoot (u'), (thus c_x is true despite x), and T was somehow rescued (y'). Under such scenario, Enemy-2 shooting would not produce T's death, hence, $PS(x \rightarrow y | c) = 0$. This matches our intuition; upon learning that T's body contains cyanide, emptying the canteen is no longer considered the cause of death.

Now consider the evidence d: "dehydration." To evaluate

$$PS(x \rightarrow y | d) = P(y_x | x', y', d_x)$$

we need first list all scenarios compatible with (x', y', d_x), namely: no shot fired, T is alive and T would be dehydrated if Enemy-2 were to shoot. Two scenarios come to mind, one natural, the other bizarre.

Scenario-1: No shot fired, the water is poisoned, the poisoned water would be emptied if Enemy-2 were to shoot (u), and T would suffer dehydration. In this scenario x would produce death, unless T is rescued.

Scenario-2: No shot was fired, T would come to drink before the shot (if any) but would somehow suspect that the water is poisoned and refrain from drinking. This would cause dehydration by choice, and death unless rescued.

Summing over both scenarios, we obtain

$$PS(x \to y|d) = 1 - P(T \ \textit{survives in dehydration}).$$

To summarize, we now have

$$\begin{aligned} PS(x \to y) &= 1 - P(\textit{survival with empty canteen}) \\ PS(x \to y|c) &= 0 \\ PS(x \to y|d) &= 1 - P(T \ \textit{will be rescued after dehydration}). \end{aligned} \tag{75}$$

Now consider the sufficiency of Enemy-1's action, in light of the two forensic reports. The conditioning part in

$$PS(p \to y|c) = P(y_p|p', y', c_p)$$

instructs us to imagine a scenario in which Enemy-1 did not poison the water, T did not die, but cyanide would be found in T's body if Enemy-1 were to poison the water. This is the natural scenario to evolve if Enemy-2 did not shoot – T would die if the water were poisoned (y_p) unless rescued before the cyanide exerts its effect. Thus,

$$PS(p \to y|c) = 1 - P(\textit{rescued after drinking cyanide})$$

Finally, consider the evidence d: "dehydration"

$$PS(p \to y|d) = P(y_p|p', y', d_p)$$

We need first to list all scenarios compatible with (p', y', d_p), namely: no poisoning occurred, T is alive and T would be dehydrated if enemy-1 were to poison the water. This is a bit hard to imagine, but not totally infeasible if we allow a special rescue operation: Enemy-2 shoots, the container is empty, T comes to drink after the shot is fired, dehydration occurs regardless of Enemy-1 action (d_p), but a rescue team revives T despite his state of dehydration.

In this scenario survival would occur even under p, therefore

$$PS(p \to y|d) = 0$$

Summarizing:

$$\begin{aligned} PS(p \to y) &= 1 - P(\textit{survival with poisoned canteen}) \\ PS(p \to y|c) &= 1 - P(\textit{rescue after drinking cyanide}) \\ PS(p \to y|d) &= 0 \end{aligned} \tag{76}$$

6.3.3 Necessity given Forensic reports

The probabilities associated with necessary causation are usually easier to evaluate than their sufficiency counterparts, because the former call for scenarios that actually materialized in the story. To illustrate, let us evaluate the probability that Enemy-2 was a necessary cause of T's death, given that cyanide was found in T's body,

$$PN(x \rightarrow y|c) = P(y'_{x'}|x, y, c)$$

The condition (x, y, c) can materialize only in state u', where T drinks the poisoned water before the shot. Assuming this state, it is clear that T is doomed regardless of Enemy-2 action, and $y'_{x'}$ is false. Thus,

$$PN(x \rightarrow y|c) = 0$$

Prospects of rescue, as we have mentioned before, do not alter this conclusion, because those are ruled out by the conditioning part.

A dehydration report would evoke the normal scenario, since

$$PN(x \rightarrow y|d) = P(y'_{x'}|x, y, d)$$

and condition (x, y, d) can materialize in state u: T reaches for drink after the shot is fired, finds the canteen empty, and suffers dehydration. In this state, $y'_{x'}$ is again false, because death would occur (from poison) even if Enemy-2 refrains from action (x'). Thus, as expected,

$$PN(x \rightarrow y|d) = 0$$

For completeness, we evaluate the necessity ascribed to Enemy-1 action,

$$\begin{aligned} PN(p \rightarrow y|c) &= P(y'_{p'}|p, y, c) \\ &= P(T \ survives \ if \ not \ p|u') = 0 \end{aligned} \tag{77}$$

because (p, y, c) implies that T drank the poisoned water before Enemy-2 fired and, in this state (u'), he would have died (from dehydration) even if Enemy-1 had not poisoned the water.

$$\begin{aligned} PN(p \rightarrow y|d) &= P(y'_{p'}|p, y, d) = \\ &= P(T \ survives \ if \ not \ p|u) = 0 \end{aligned} \tag{78}$$

because (p, y, d) implies that T reached for drink after Enemy-2 fired (u) and, in this state would have died (from dehydration) even if Enemy-1 had not poisoned the canteen.

Note that if we are not given any forensic report but assume, nevertheless, that such reports were available from the natural scenario in the story (i.e. u, x, p, d, y), then the probabilities of sufficiency would be (barring considerations of survival):

$$\begin{aligned} PS(x \rightarrow y|d) &= 1 \\ PS(p \rightarrow y|d) &= 0 \end{aligned} \tag{79}$$

These results coincide with those obtained from Lewis' analysis, using counterfactual-dependence chains. Whether this coincidence is universal, and whether it could serve as the basis for improving Lewis' account of causation remain a topic for future investigation.

7 Conclusion

This paper explicates and analyzes the necessary and sufficient components of causation. Using counterfactual interpretations that rest on structural-model semantics, the paper demonstrates how simple techniques of computing probabilities of counterfactuals can be used in computing probabilities of causes, deciding questions of identification, defining conditions under which probabilities of causes can be estimated from statistical data, and uncovering tests for assumptions that are routinely made (often unwittingly) by analysts and investigators.

On the practical side, the paper offers several useful tools to epidemiologists and health scientists. It formulates and calls attention to basic assumptions that must be ascertained before statistical measures such as excess-risk-ratio could represent causal quantities such as attributable-risk or probability of causes. It shows how data from both experimental and non-experimental studies can be combined to yield information that neither study alone can reveal. Finally, it provides tests for the commonly made assumption of "no prevention," and for the often asked question of whether a clinical study is representative of its target population.

On the conceptual side, we have seen that both the probability of necessity (PN) and probability of sufficiency (PS) play a role in our understanding of causation, and that both components have their logics and computational rules. Although the counterfactual concept of necessary cause (i.e., that an outcome would not have occurred "but for" the action) is predominant in legal settings [Robertson, 1997] and in ordinary discourse, the sufficiency component of causation has a definite influence on causal thoughts.

The sufficiency component plays a major role in scientific and legal explanations, as can be seen from examples where the necessary component is dormant. Why do we consider striking a match to be a more adequate explanation (of a fire) than the presence of oxygen? Recasting the question in the language of PN and PS, we note that, since both explanations are necessary for the fire, each will command a PN of unity. (In fact PN is higher for the oxygen, if we allow for alternative ways of igniting a spark). Thus, it must be the sufficiency component alone that endows the match with greater explanatory power than the oxygen. If the probabilities associated with striking a match and the presence of oxygen are p_m and p_o, respectively, the PS measures associated with these explanations evaluate to $PS(match) = p_o$ and $PS(oxygen) = p_m$, clearly favoring the match when $p_o >> p_m$. Thus, a robot instructed to explain why a fire broke out has no choice but to consider both PN and PS in its deliberations.

Should PS enter legal considerations in criminal and tort law? I believe that it should, (as does I.J. Good (1993)), because attention to sufficiency implies attention to the consequences of one's action. The person who lighted the match ought to have anticipated the presence of oxygen, whereas the person who supplied (or who could but failed to remove) the oxygen is not generally expected to have anticipated match-striking ceremonies.

However, what weight should the law assign to the necessary versus the sufficient component of causation? This question obviously lies beyond the scope of our investigation, and it is not at all clear who would be qualified to tackle the question or whether our legal system would be prepared to implement the recommendation. I am hopeful, however, that whoever undertakes to consider such questions will find the analysis in this paper to be of some use.

Acknowledgments

I am indebted to Sander Greenland for many suggestions and discussions concerning the treatment of causation in the epidemiological literature and potential applications of this analysis in practical epidemiological studies. Donald Michie and Jack Good are responsible for shifting my attention from PN to PS and PNS. Clark Glymour and Patricia Cheng have helped unravel some of the mysteries of causal power theory, and Michelle Pearl has provided useful pointers to the epidemiological literature. This investigation was supported in part by grants from NSF, AFOSR, ONR, and California MICRO program.

References

[Aldrich, 1993] J. Aldrich. Cowles' exogeneity and core exogeneity. Technical Report Discussion Paper 9308, Department of Economics, University of Southampton, England, 1993.

[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, 1994a] A. Balke and J. Pearl. Counterfactual probabilities: Computational methods, bounds, and applications. In R. Lopez de Mantaras and D. Poole, editors, *Uncertainty in Artificial Intelligence 10*, pages 46–54. Morgan Kaufmann, San Mateo, CA, 1994.

[Balke and Pearl, 1994b] A. Balke and J. Pearl. Probabilistic evaluation of counterfactual queries. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, volume Volume I, pages 230–237. MIT Press, Menlo Park, CA, 1994.

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

[Balke and Pearl, 1997] A. Balke and J. Pearl. Nonparametric bounds on causal effects from partial compliance data. *Journal of the American Statistical Association*, 92(439):1–6, September 1997.

[Breslow and Day, 1980] N.E. Breslow and N.E. Day. *Statistical Methods in Cancer Research; Vol. 1, The Analysis of Case-Control Studies.* IARC, Lyon, 1980.

[Cartwright, 1989] N. Cartwright. *Nature's Capacities and Their Measurement.* Clarendon Press, Oxford, 1989.

[Cheng, 1997] P.W. Cheng. From covariation to causation: A causal power theory. *Psychological Review*, 104(2):367–405, 1997.

[Cole, 1997] P. Cole. Causality in epidemiology, health policy, and law. *Journal of Marketing Research*, 27:10279–10285, 1997.

[Dawid, 1997] A.P. Dawid. Causal inference without counterfactuals. Technical report, Department of Statistical Science, University College London, UK, 1997. Forthcoming (with discussion), *Journal of the American Statistical Association.*

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

[Eells, 1991] E. Eells. *Probabilistic Causality.* Cambridge University Press, Cambridge, MA, 1991.

[Engle *et al.*, 1983] R.F. Engle, D.F. Hendry, and J.F. Richard. Exogeneity. *Econometrica*, 51:277–304, 1983.

[Fine, 1975] K. Fine. Review of lewis' counterfactuals. *Mind*, 84:451–458, 1975.

[Fine, 1985] K. Fine. *Reasoning with Arbitrary Objects.* B. Blackwell, New York, 1985.

[Finkelstein and Levin, 1990] M.O. Finkelstein and B. Levin. *Statistics for Lawyers.* Springer-Verlag, New York, 1990.

[Fisher, 1970] F.M. Fisher. A correspondence principle for simultaneous equations models. *Econometrica*, 38(1):73–92, January 1970.

[Fleiss, 1981] J.L. Fleiss. *Statistical Methods for Rates and Proportions.* John Wiley and Sons, New York, second edition, 1981.

[Galles and Pearl, 1995] D. Galles and J. Pearl. Testing identifiability of causal effects. In P. Besnard and S. Hanks, editors, *Uncertainty in Artificial Intelligence 11*, pages 185–195. Morgan Kaufmann, San Francisco, 1995. Also in [Pearl, 2000], Chapter 4.

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

[Galles and Pearl, 1998] D. Galles and J. Pearl. An axiomatic characterization of causal counterfactuals. *Foundations of Science*, 3(1):151–182, 1998.

[Glymour, 1998] C. Glymour. Psychological and normative theories of causal power and the probabilities of causes. In G.F. Cooper and S. Moral, editors, *Uncertainty in Artificial Intelligence*, pages 166–172. Morgan Kaufmann, San Francisco, CA, 1998.

[Goldszmidt and Pearl, 1992] M. Goldszmidt and J. Pearl. Rank-based systems: A simple approach to belief revision, belief update, and reasoning about evidence and actions. In B. Nebel, C. Rich, and W. Swartout, editors, *Proceedings of the Third International Conference on Knowledge Representation and Reasoning*, pages 661–672. Morgan Kaufmann, 1992.

[Good, 1961] I.J. Good. A causal calculus, I. *British Journal for the Philosophy of Science*, 11:305–318, 1961.

[Good, 1993] I.J. Good. A tentative measure of probabilistic causation relevant to the philosophy of the law. *J. Statist. Comput. and Simulation*, 47:99–105, 1993.

[Greenland and Robins, 1988] S. Greenland and J. Robins. Conceptual problems in the definition and interpretation of attributable fractions. *American Journal of Epidemiology*, 128:1185–1197, 1988.

[Hall, 1998] N. Hall. Two concepts of causation, 1998. In press.

[Halpern, 1998] J.Y. Halpern. Axiomatizing causal reasoning. In G.F. Cooper and S. Moral, editors, *Uncertainty in Artificial Intelligence*, pages 202–210. Morgan Kaufmann, San Francisco, CA, 1998.

[Heckerman and Shachter, 1995a] D. Heckerman and R. Shachter. Decision-theoretic foundations for causal reasoning. *Journal of Artificial Intelligence Research*, 3:405–430, 1995.

[Heckerman and Shachter, 1995b] D. Heckerman and R. Shachter. A definition and graphical representation for causality. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 262–273, San Mateo, CA, 1995. Morgan Kaufmann.

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

[Hennekens and Buring, 1987] C.H. Hennekens and J.E. Buring. *Epidemiology in Medicine.* Brown, Little, Boston, 1987.

[Hume, 1948] D. Hume. *An Enquiry concerning Human Understanding.* Open Court Press, LaSalle, 1948. Reprinted 1988.

[Imbens, 1997] G.W. Imbens. Book reviews. *Journal of Applied Econometrics*, 12, 1997.

[Kelsey *et al.*, 1996] J.L. Kelsey, A.S. Whittemore, A.S. Evans, and W.D Thompson. *Methods in Observational Epidemiology.* Oxford University Press, New York, 1996.

[Khoury *et al.*, 1989] M.J. Khoury, W.D Flanders, S. Greenland, and M.J. Adams. On the measurement of susceptibility in epidemiologic studies. *American Journal of Epidemiology*, 129(1):183–190, 1989.

[Kim, 1971] J. Kim. Causes and events: Mackie on causation. *Journal of Philosophy*, 68:426–471, 1971. Reprinted in E. Sosa and M. Tooley (Eds.), *Causation*, Oxford University Press, 1993.

[Lewis, 1979] D. Lewis. Counterfactual dependence and time's arrow. *Nous*, 13:418–446, 1979.

[Lewis, 1986] D. Lewis. *Philosophical Papers.* Oxford University Press, New York, 1986.

[Mackie, 1965] J.L. Mackie. Causes and conditions. *American Philosophical Quarterly*, 2/4:261–264, 1965. Reprinted in E. Sosa and M. Tooley (Eds.), *Causation*, Oxford University Press, 1993.

[Marschak, 1950] J. Marschak. Statistical inference in economics. In T. Koopmans, editor, *Statistical Inference in Dynamic Economic Models*, pages 1–50. Wiley, New York, 1950. Cowles Commission for Research in Economics, Monograph 10.

[Michie, 1997] D. Michie. Adapting Good's q theory to the causation of individual events. Technical report, University of Edinburgh, UK, 1997. Submitted for publication in *Machine Intelligence 15*.

[Mill, 1843] J.S. Mill. *System of Logic*, volume 1. John W. Parker, London, 1843.

[Neyman, 1923] J. Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. English Translation (1990) *Statistical Science*, 5(4):465–480, 1923.

[Pearl, 1993] J. Pearl. Comment: Graphical models, causality, and intervention. *Statistical Science*, 8:266–269, 1993.

[Pearl, 1994] J. Pearl. A probabilistic calculus of actions. In R. Lopez de Mantaras and D. Poole, editors, *Uncertainty in Artificial Intelligence 10*, pages 454–462. Morgan Kaufmann, San Mateo, CA, 1994.

[Pearl, 1995] J. Pearl. Causal diagrams for experimental research. *Biometrika*, 82:669–710, December 1995.

[Pearl, 1996a] J. Pearl. Causation, action, and counterfactuals. In Y. Shoham, editor, *Theoretical Aspects of Rationality and Knowledge, Proceedings of the Sixth Conference*, pages 51–73. Morgan Kaufmann, San Francisco, CA, 1996.

[Pearl, 1996b] J. Pearl. Structural and probabilistic causality. In D.R. Shanks, K.J. Holyoak, and D.L. Medin, editors, *The Psychology of Learning and Motivation*, volume 34, pages 393–435. Academic Press, San Diego, CA, 1996.

[Pearl, 1998] J. Pearl. On the definition of actual cause. Technical Report R-259, Department of Computer Science, University of California, Los Angeles, CA, 1998. Also in [Pearl, 2000], Chapter 10.

[Pearl, 2000] J. Pearl. *Causality*. Cambridge University Press, New York, 2000. Forthcoming.

[Robertson, 1997] D.W. Robertson. The common sense of cause in fact. *Texas Law Review*, 75(7):1765–1800, 1997.

[Robins and Greenland, 1989] J.M. Robins and S. Greenland. The probability of causation under a stochastic model for individual risk. *Biometrics*, 45:1125–1138, 1989.

[Robins, 1986] J.M. 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*, 7:1393–1512, 1986.

[Robins, 1987] J.M. Robins. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Diseases*, 40(Suppl 2):139S–161S, 1987.

[Rosenbaum and Rubin, 1983] P. Rosenbaum and D. Rubin. The central role of propensity score in observational studies for causal effects. *Biometrica*, 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.

[Schlesselman, 1982] J.J. Schlesselman. *Case-Control Studies: Design Conduct Analysis.* Oxford University Press, New York, 1982.

[Shep, 1958] M.C. Shep. Shall we count the living or the dead? *New England Journal of Medicine*, 259:1210–1214, 1958.

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

[Simon, 1953] H.A. Simon. Causal ordering and identifiability. In Wm. C. Hood and T.C. Koopmans, editors, *Studies in Econometric Method*, pages 49–74. Wiley and Sons, Inc., 1953.

[Skyrms, 1980] B. Skyrms. *Causal Necessity.* Yale University Press, New Haven, 1980.

[Sobel, 1990] M.E. Sobel. Effect analysis and causation in linear structural equation models. *Psychometrika*, 55(3):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 H.O.A. Wold. Recursive versus nonrecursive systems: An attempt at synthesis. *Econometrica*, 28:417–427, 1960.

[Suppes, 1970] P. Suppes. *A Probabilistic Theory of Causality.* North-Holland Publishing Co., Amsterdam, 1970.

[Thomason and Gupta, 1980] R. Thomason and A. Gupta. A theory of conditionals in the context of branching time. *Philosophical Review*, 88:65–90, 1980.

A APPENDIX: The empirical content of counterfactuals

The word "counterfactual" is a misnomer, as 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, *ceteras paribum*, at any future times $t > t_0$, if the current flow will be $I(t)$ the voltage drop 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 drop 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 while the counterfactual form merely speculates about events that have not, and could not have occurred; as it is impossible to apply two different currents into the same resistor at the same time. However, if we interpret the counterfactual form to mean no more nor less than a conversational short hand 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 (Ohm's law) which ascribes a time-invariant property (the ratio V/I) to any physical object.

I will adapt this predictive interpretation when I speak of counterfactuals, and I base this interpretation on the observation that counterfactuals, despite their a-temporal appearance, are invariably associated with some law-like, persistent relationships in the world. For example, the statement "had Germany not been punished so severely at the end world-war I, Hitler would not have come to power" would sound bizarre to anyone who does not share our understanding that, as a general rule, "humiliation breeds discontent."

But if counterfactual statements are merely a round-about way of stating sets of predictions, why do we resort to such convoluted modes of expression instead of using the predictive mode directly? The answer, I believe, rests with the qualification "ceteras paribum" 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 equipments? 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 (hence superfluous) when we use counterfactual expressions, especially when we agree over the underlying causal model. For example, we do not need to specify under what temperature and pressure future 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. That statement also implies that we do not really mean for anyone to hold the reading on the voltmeter constant – only variables that, according to our causal model, are not affected by the counterfactual antecedent (I) are expected to remain constant for the predictions to hold true.

To summarize, I interpret a counterfactual statement to convey a set of predictions under well defined set of conditions, those prevailing in the factual part of the statement. For these predictions to be valid, two components must remain invariants: 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 exogenous variables U. Thus, a precondition for the validity of the predictive interpretation of a counterfactual statement is the assumption that U will remain the same at the time where our predictive claim is to be applied or tested.

This is best illustrated using the betting example of Section 4.1. The predictive interpretation of the counterfactual "Had I bet differently I would have lost a dollar" is the 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. While the former is a plausible assumption in 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 hind-sighted nonsense. (It is not hard however to imagine a lottery in which the payoff policy and the outcome of the random device remain constant for a short period of time, during which additional bets are accepted and processed. Most those who play the stock market believe in strategies that allow an investor to quickly recover from a bad move.) At any rate, it is the persistence across time of U and $f(x, u)$ that endows counterfactual expressions with predictive power; take this persistence away, and the counterfactual loses its obvious economical utility.

I said "obvious" because there is an element of utility in counterfactuals that does not translate immediately to predictive payoff, and may explain, nevertheless, 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 like the one who decides which way to spin our atoms and electrons, but 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 to 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 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 submit that such unlikely operations are the norm in gauging the explanatory value of sentences. In fact, it is the nature of any explanation, especially causal, that its utility be amortized not over standard situations but, rather, over novel settings which require innovative manipulation of one's environment.

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), (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 on 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 dice 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 we disband all talk of counterfactuals (a strategy recommended by some researchers [Dawid, 1997]) or we continue to use counterfactuals but limit their usage to situations where they assume empirical meaning. This amounts to keeping in the analysis only U's that satisfy conditions (1) and (2) above. Instead of hypothesizing U's that completely remove all uncertainties, we admit only those U's that are either (1) persistent or (2) potentially observable.

Naturally, coarsening the granularity of the exogenous variables has its price tag; the mechanism equations $y = f(x, u)$ lose their deterministic character and 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 lies beyond the scope of the present paper, but its basic character should follow from the three steps of abduction-action-deduction, outlined in Section 2.2.