TECHNICAL REPORT
R-240
January 1996

ELSEVIER

Artificial Intelligence 97 (1997) 9–43

Artificial Intelligence

Axioms of causal relevance

David Galles [1], Judea Pearl *

Cognitive Systems Laboratory, Computer Science Department, University of California, Los Angeles, CA 90024, USA

Received October 1995; revised May 1996

Abstract

This paper develops axioms and formal semantics for statements of the form "X is causally irrelevant to Y in context Z", which we interpret to mean "Changing X will not affect Y once Z is held constant". The axiomization of causal irrelevance is contrasted with the axiomization of informational irrelevance, as in "Finding X will not alter our belief in Y, once we know Z". Two versions of causal irrelevance are analyzed: probabilistic and deterministic. We show that, unless stability is assumed, the probabilistic definition yields a very loose structure that is governed by just two trivial axioms. Under the stability assumption, probabilistic causal irrelevance is isomorphic to path interception in cyclic graphs. Under the deterministic definition, causal irrelevance complies with all of the axioms of path interception in cyclic graphs except transitivity. We compare our formalism to that of Lewis (1973) and offer a graphical method of proving theorems about causal relevance. © 1997 Elsevier Science B.V.

Keywords: Causality; Graphoids; Causal models; Counterfactuals; Actions

1. Introduction

In [10], a set of axioms was developed for a class of relations called *graphoids*. These axioms characterize informational relevance [2] among observed events based on the semantics of conditional independence in probability calculus. This paper develops a parallel set of axioms for *causal relevance*, that is, the tendency of certain events to

* Corresponding author. Email: judea@cs.ucla.edu.

[1] Email: galles@cs.ucla.edu.

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

PII S0004-3702(97)00047-7

affect the occurrence of other events in the physical world, independent of the observer-reasoner. Informational irrelevance is concerned with statements of the form "X is independent of Y given Z", which means that, given the value of Z, gaining information about X gives us no new information about Y. Causal irrelevance is concerned with statements of the form "X is causally irrelevant to Y in context Z", which we take to mean "Changing X will not alter the value of Y, if Z is fixed".

The notion of causal relevance has its roots in the philosophical works of Good [12], Suppes [45], and Salmon [37], who attempted to give probabilistic interpretations to cause–effect relationships, and recognized the need to distinguish causal from statistical relevance. Although these attempts have not produced an algorithmic definition of causal relevance, they led to methods for testing the consistency of relevance statements against a given probability distribution and a given temporal ordering among the variables [3,5,32]. The current paper aims at axiomatizing relevance statements in themselves, with no reference to underlying probabilities or temporal orderings.

Axiomatic characterization of causal relevance may serve as a normative standard for analyzing theories of action as well as a guide for developing representation schemes (e.g., graphical models) for planning and decision-making applications. For example, instead of explicitly storing all possible effects of an action, as in STRIPS [6], such representation schemes should enable an agent to examine only direct effects of actions and to infer which actions are relevant for a given goal and which actions cease to be relevant once others are implemented.

Another application of causal relevance lies in the area of automatic language generation—for example, in complex diagnostic systems, where machine-generated explanations are loaded with causal utterances. The formalization of causal relevance and causal relationships in general should assist a machine in distinguishing and selecting proper linguistic nuances in causal conversations. Statements such as "A normally causes B", "B was caused by A", "A was the cause of B", "B occurred despite A", or "B would not have occurred if it were not for A" all express some form of causal relevance between A and B, yet these utterances are not entirely equivalent and making the appropriate choice may require careful understanding of the relation between A and B in the context of the discussion.

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

We provide two formal definitions of causal irrelevance. The probabilistic definition, which equates causal irrelevance with inability to change the probability of the effect variable, has intuitive appeal but is inferentially very weak; it does not support a very expressive set of axioms unless further assumptions are made about the underlying causal model. If we add the stability assumption (i.e., that no irrelevance can be destroyed by changing the nature of the individual processes in the system), then we obtain the same set of axioms for probabilistic causal irrelevance as the set governing path interception in directed graphs. The deterministic definition, which equates causal irrelevance with inability to change the effect variable in any state of the world, allows for a rich set of axioms without our making any assumptions about the causal model. All of the path-interception axioms for directed graphs, with the exception of transitivity, hold for deterministic causal irrelevance.

In Section 2, we define causal models, a formal system for interpreting causal statements. In Section 3, we provide a definition of probabilistic causal irrelevance and determine which of the graphoid axioms hold under this definition. Finally, in Section 4, we give a nonprobabilistic definition of causal irrelevance and offer a graphical method for proving statements about causal irrelevance.

2. Causal models

A causal model is a complete specification of the causal relationships that govern a given domain; namely, it is a mathematical object that provides an interpretation (and computation) of every causal query about the domain. Following [29] we will adopt here a definition that generalizes most of the causal models used in engineering and economics.

Definition 1 (*Causal model*). A *causal model* is a 3-tuple

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

where

(i) $V = \{X_1, \ldots, X_n\}$ is a set of endogenous variables determined within the system,
(ii) $U = \{U_1, \ldots, U_m\}$ is a set of exogenous or background variables that represent disturbances, abnormalities, assumptions, or boundary conditions, and
(iii) F is a set of n nontrivial functions $\{f_1, \ldots, f_n\}$, each having the form

$$x_i = f_i(\boldsymbol{pa}_i, u), \quad i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{pa}_i$ are the values of a set of variables $PA_i \subseteq V \setminus X_i$ (connoting parents), called the *direct causes* of X_i. We will assume that the set of equations in (iii) has a unique solution for $X_1, \ldots, X_n$, given any value of the background variables $U_1, \ldots, U_m$. Thus we can consider each variable $X \in V$ to be a function $X_M(u)$ of the background U in the causal model M.

The uniqueness assumption is always satisfied in *recursive models*, where PA_i are predecessors of X_i in some order, but may be violated in nonrecursive systems, that is,

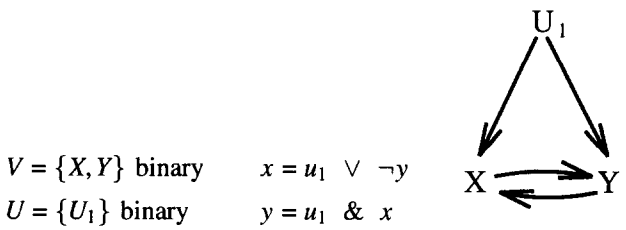


Fig. 1. A valid nonrecursive causal model, with unique values for X and Y for all values of U.

systems with feedback. For example, consider the equations $x = y \vee u$ and $y = x \vee u$. The state $U = 0$ permits two possible solutions for X and Y—namely, $(X = 1, Y = 1)$ and $(X = 0, Y = 0)$—so such functions would be disallowed in a causal model. The uniqueness requirement in nonrecursive models conveys the understanding that F represents a deterministic physical system in equilibrium. Indeed, if we assume that all relevant background conditions U were accounted for, such a system can only be in one state. Systems possessing several equilibrium states indicate the existence of dynamic factors, not modeled in U. Such factors often can be summarized by the notion of *previous state*, and incorporated into our analysis as a third kind of variables supplementing V and U [9].

The assumption that there is a unique solution for $X_1, \ldots, X_n$, while limiting the scope of Definition 1, does not prevent the use of causal models to describe feedback systems in stable equilibrium. The equations do not need to be recursive to ensure uniqueness. For example, the causal model shown in Fig. 1 dictates unique values for X and Y for $U_1 = 0$ and $U_1 = 1$.

Drawing arrows between the variables PA_i and X_i defines a directed graph $G(M)$, which we call the *causal graph* of M. In general, $G(M)$ can be cyclic. For some examples of causal models, see Section 2.1.

Definition 1 merely provides a description of the mathematical objects that enter into a causal model. To fulfill our requirement that a causal model be capable of computing answers for causal queries, we need to supplement Definition 1 with an interpretation of the sentence "$X = x$ causes $Y = y$". In ordinary discourse, such a sentence implies that we can bring about the condition $Y = y$ by locally enforcing the condition $X = x$. Thus, Definition 1 must be supplemented with a formal interpretation of the notion "locally enforcing $X = x$" that is compatible with its common usage.

External intervention normally implies changing some mechanisms in the domain. In a logical circuit, for example, the act of enforcing the condition $X_i = 0$ by connecting some intermediate variable X_i to ground amounts to changing the mechanism that normally determines X_i. If X_i is the output of an OR gate, then after the intervention, X_i would no longer be determined by the OR gate but by a new mechanism (involving the ground) that clamps X_i to 0 regardless of the input to the OR gate. In the equational representation, this amounts to replacing the equation $x_i = f_i(\boldsymbol{pa}_i, u)$ with a new equation, $X_i = 0$, that represents the grounding of X_i.

The replacement of just one equation, not several, reflects the principle of locality in the common understanding of imperative sentences such as "Raise taxes" or "Make him laugh". When told to clean his face, a child does not ask for a razor, nor does he jump into the swimming pool. The proper interpretation of the modal sentence "*do p*" corresponds to a a minimal perturbation of the existing state of affairs, and this, in the context of Definition 1, corresponds to the replacement of the minimal set of equations necessary to make p compatible with U.

In general, we will consider concurrent action of the form $do(X = x)$, where X involves several variables in V.[3] This leads to the following definitions.

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

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

where

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

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

Submodels are useful for representing the effect of local actions and changes. If we interpret each function f_i in F as an independent physical mechanism and define the action $do(X = x)$ as the minimal change in M required to make $X = x$ hold true under any u, then M_x represents the model that results from such a minimal change, since it differs from M by only those mechanisms that determine the variables in X.

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.

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 $Y_x(u)$.[4]

[3] The formalization of conditional actions of the form "$do(X = x)$ if $Z = z$" is straightforward [28].

[4] The connection between counterfactuals and local actions is made by Lewis [17] and is further elaborated by Balke and Pearl [1] and Heckerman and Shachter [14]. Readers who are disturbed by the impracticality of actions in the interpretation of some counterfactuals (e.g., "If I were young") are invited to replace the word "action" with the word "modification" (see [16]). Pearl [29, p.706] explains the advantage of using hypothetical external interventions, rather than spontaneous changes, in thinking about causation and counterfactuals.

The syntactical transformation described in Definition 4 corresponds to replacing the old functional mechanisms $x_i = f_i(PA_i, u)$ with new mechanisms $X_i = x_i$ that represent the external forces that set the values x_i for each $X_i \in X$. As before, we will assume each variable $Y \in V$ to be a unique function of the background U in any model M_x: $Y = Y_{M_x}(u)$. For brevity, the subscript M is often omitted, leaving $Y_x(u)$.

The notation $Y_x(u)$ is sometimes used in the statistical literature [36] to stand for the counterfactual sentence "The value that Y would take in person u had X been x", where X stands for a type of treatment that a person can receive. There is a strong connection between the sentence above and our formal interpretation of $Y_x(u)$ [29]. Definition 4 interprets this abstract, counterfactual sentence in terms of the processes responsible for Y taking on the value $Y_x(u)$ as X changes to x. It treats u not as merely the index of an individual but, rather, as the set of attributes u that characterize the individual, the experimental conditions under study, and so on. In fact, every causal model meeting the requirements of Definition 1 can be translated into a set of counterfactual statements of the type used in the statistical literature [29, p. 703]. In Section 4, we will further show that the process-based semantics given in Definition 4 will uncover new properties of $Y_x(u)$ that were not formalized in the statistical literature.

An explicit translation of intervention into "wiping out" equations in the causal model was first proposed by Strotz and Wold [43] and used by Fisher [7] and Sobel [40]. Graphical ramifications are explicated by Spirtes et al. [41] and Pearl [27]. Interpretations of causal and counterfactual utterances in terms of $Y_x(u)$ are provided by Pearl [31]. Other formulations of causality, in terms of event trees, are given by Robins [33] and Shafer [39].

Note that $Y_x(u)$ is well defined even when $U = u$ and $X = x$ are incompatible in M (i.e., $X(u) \neq x$), thus allowing for actions to enforce propositions that are not realized under normal conditions, or under the abnormalities modeled in U. For example, if M describes a logic circuit we might wish to intervene and set some voltage X to x, even though the input dictates $X \neq x$. It is for this reason that one must invoke some notion of mechanism breakdown or "surgery" in the definition of interventions.

The unique feature of our formulation of actions—the feature that sets it apart from the formulations in control theory or decision analysis [14, 38]—is that an action is treated as a *modality*, namely, it is not given an explicit name but acquires the names of the propositions that it enforces as true. This enables the model to predict the effects of a huge number of action combinations without the modeler having to attend to such combinations. Instead, the causal model is constructed by specifying the characteristics of each individual mechanism under normal conditions, free of intervention.

We can extend the notion of causal models to encode probabilistic information as follows:

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

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

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

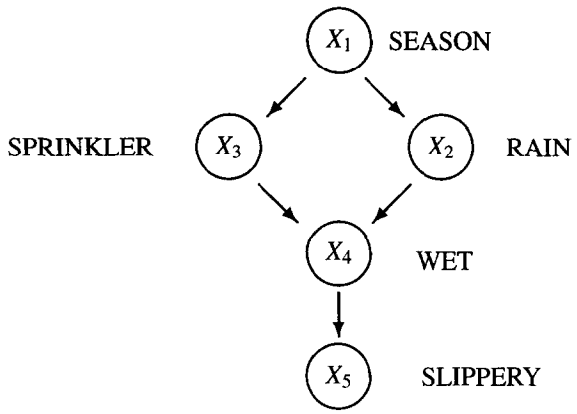


Fig. 2. Causal graph illustrating causal relationships among five variables.

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

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

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

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

We note that a causal model defines a joint distribution on all counterfactual statements, that is, $P(Y_x = y, Z_w = z)$ is defined for any sets of variables Y, X, Z, W, not necessarily disjoint. In particular, $P(Y_x = y, Y_{x'} = y')$ is well defined and is given by $\sum_{\{u|Y_x(u)=y \,\&\, Y_{x'}(u)=y'\}} P(u)$. Likewise, $P(Y_x = y, X = x')$ is well defined and is given by $\sum_{\{u|Y_x(u)=y \,\&\, X(u)=x'\}} P(u)$.[5]

2.1. Examples

Next we demonstrate the generality of the mathematical object defining causal models using two familiar applications: evidential reasoning and linear structural equation models.

[5] The existence of such joint distributions has prompted some of the objections to treating counterfactuals as random variables, because, when x and x' are incompatible, it is hard to attribute probability to the joint statement "Y would be y if X were x and X is actually x'". The definition of Y_x in terms of submodel not only avoids such problems but also illustrates that such joint probabilities can be encoded rather parsimoniously using $P(u)$ and F.

2.1.1. Sprinkler example

Fig. 2 is a simple yet typical causal graph used in common sense reasoning. It describes the causal relationships among the season of the year (X_1), whether rain falls (X_2) during the season, whether the sprinkler is on (X_3), whether the pavement is wet (X_4), and whether the pavement is slippery (X_5). All variables in this graph except the root variable X_1 take a value of either "True" or "False". X_1 takes one of four values: "Spring", "Summer", "Fall", or "Winter". Here, the absence of a direct link between, for example, X_1 and X_5, captures our understanding that the influence of the season on the slipperiness of the pavement is mediated by other conditions (e.g., the wetness of the pavement). The corresponding model consists of five functions, each representing an autonomous mechanism:

$$\begin{aligned} x_1 &= u_1 \\ x_2 &= f_2(x_1, u_2) \\ x_3 &= f_3(x_1, u_3) \\ x_4 &= f_4(x_3, x_2, u_4) \\ x_5 &= f_5(x_4, u_5) \end{aligned} \tag{5}$$

The disturbances $U_1, \ldots, U_5$ are not shown explicitly in Fig. 2 but are understood to govern the uncertainties associated with the causal relationships. The causal graph coincides with the Bayesian network associated with $P(x_1, \ldots, x_5)$ whenever the disturbances are assumed to be independent, $U_i \perp U \setminus U_i$.

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

$$\begin{aligned} x_2 &= [(X_1 = \text{Winter}) \vee (X_1 = \text{Fall}) \vee ab_2] \wedge \neg ab_2' \\ x_3 &= [(X_1 = \text{Summer}) \vee (X_1 = \text{Spring}) \vee ab_3] \wedge \neg ab_3' \\ x_4 &= (x_2 \vee x_3 \vee ab_4) \wedge \neg ab_4' \\ x_5 &= (x_4 \vee ab_5) \wedge \neg ab_5' \end{aligned} \tag{6}$$

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

To represent the action "turning the sprinkler ON", or $do(X_3 = \text{ON})$, we replace the equation $x_3 = f_3(x_1, u_3)$ in the model of Eq. (5) with $X_3 = \text{ON}$. The resulting submodel, $M_{X_3=\text{ON}}$, contains all the information needed for computing the effect of the action on the other variables. It is easy to see from this submodel that the only variables affected by the action are X_4 and X_5, that is, the descendants of the manipulated variable X_3. Note, however, that the operation $do(X_3 = \text{ON})$ stands in marked contrast to that of

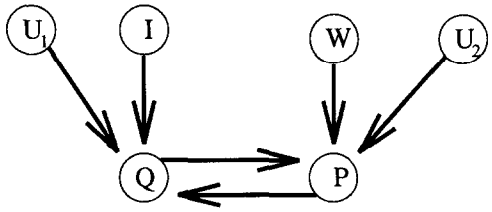


Fig. 3. Causal graph illustrating the relationship between supply and demand.

finding the sprinkler ON; the latter involves making the substitution X_3 = ON *without* removing the equation for X_3, and therefore may potentially influence (the belief in) every variable in the network. This mirrors the difference between seeing and doing: after observing that the sprinkler is ON, we may wish to infer that the season is dry, that it probably did not rain, and so on; no such inferences can be drawn about the reasons for the action "turning the sprinkler ON".

2.1.2. Policy analysis in linear econometric models

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

$$q = b_1 p + d_1 i + u_1, \tag{7}$$

$$p = b_2 q + d_2 w + u_2, \tag{8}$$

where q is the quantity of household demand for a product A, p is the unit price of product A, i is household income, w is the wage rate for producing product A, and u_1 and u_2 represent error terms, namely, unmodeled factors that affect quantity and price, respectively [11].

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

We can use this model to answer queries such as:

(1) Find the expected demand (Q) if the price is controlled at $P = p_0$.
(2) Find the expected demand (Q) if the price is reported to be $P = p_0$.
(3) Given that the current price is $P = p_0$, find the expected demand (Q) had the price been $P = p_1$.

To find the answer to the first query, we replace Eq. (8) with $p = p_0$, leaving

$$q = b_1 p + d_1 i + u_1. \tag{9}$$

$$p = p_0. \tag{10}$$

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

$$E[Q \mid \widehat{p_0}] = E[Q] + b_1(p - E[P]) + d_1(i - E[I]).$$

The answer to the second query is given by conditioning Eq. (7) on the current observation $\{P = p_0, I = i, W = w\}$ and taking the expectation,

$$E[Q \mid p_0, i, w] = b_i p_0 + d_1 i + E[U_1 \mid p_0, i, w]. \tag{11}$$

The computation of $E[U_1 \mid p_0, i, w]$ is a standard procedure once Σ_{ij} is given. Note that, although U_1 was assumed independent of I and W, this independence no longer holds once $P = p_0$ is observed [21]. Note also that Eqs. (7) and (8) both participate in the solution and that the observed value p_0 will affect the expected demand q (through $E[U_1 \mid p_0, i, w]$) even when $b_1 = 0$, which is not the case in the first query.

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

$$E[Q_{p=p_1} \mid p_0, i, w] = b_1 p_1 + d_1 i + E[U_1 \mid p_0, i, w]. \tag{12}$$

The expected value $E[U_1 \mid p_0, i, w]$ is the same in the solutions to the second and third queries; the latter differs only in the term $b_1 p_1$. A general method for solving such counterfactual queries is described in [2].

2.1.3. Linguistic notions of causality

Causal models provide a precise language for defining intuitive causal concepts. In this section, we provide some brief examples, all relating to a given causal model M.

- "X is a cause of Y", if there exist two values x and x' of X and a value u of U such that $Y_x(u) \neq Y_{x'}(u)$.
- "X is a cause of Y in context $Z = z$", if there exist two values x and x' of X and a value u of U such that $Y_{xz} \neq Y_{x'z}(u)$.
- "X is a direct cause of Y", if there exist two values x and x' of X, and a value u of U such that $Y_{xr}(u) \neq Y_{x'r}(u)$ where r is some realization of $V \setminus X$.
- "X is an indirect cause of Y", if X is a cause of Y, and X is not a direct cause of Y.
- "Event $X = x$ may have caused $Y = x$" if
  (i) $X = x$ and $Y = y$ are true, and
  (ii) there exists a value u of U such that $X(u) = x$, $Y(u) = y, Y_x(u) = y$ and $Y_{x'}(u) \neq y$ for some $x' \neq x$.
- "The unobserved event $X = x$ is a likely cause of $Y = y$" if
  (i) $Y = y$ is true, and
  (ii) $P(Y_x = y, Y_{x'} \neq y \mid Y = y)$ is high for some $x' \neq x$
- "Event $Y = y$ occurred despite $X = x$", if
  (i) $X = x$ and $Y = y$ are true, and
  (ii) $P(Y_x = y)$ is low.

The preceding list demonstrates that, by varying the quantifiers of U and X, we have the flexibility of finding appropriate formalization for many nuances of causal expressions.

1.1 (Symmetry) $(X \perp Y \mid Z) \Longrightarrow (Y \perp X \mid Z)$

1.2 (Decomposition) $(X \perp YW \mid Z) \Longrightarrow (X \perp Y \mid Z)$

1.3 (Weak union) $(X \perp YW \mid Z) \Longrightarrow (X \perp Y \mid ZW)$

1.4 (Contraction) $(X \perp Y \mid Z) \,\&\, (X \perp W \mid ZY) \Longrightarrow (X \perp YW \mid Z)$

1.5 (Intersection) $(X \perp W \mid ZY) \,\&\, (X \perp Y \mid ZW) \Longrightarrow (X \perp YW \mid Z)$
Intersection requires a strictly positive probability distribution.

Fig. 4. The graphoid axioms.

3. Probabilistic causal irrelevance

The existence of a probability distribution over all variables in a probabilistic causal model leads to a natural definition of the probabilistic version of causal irrelevance.

Definition 7 (*Probabilistic causal irrelevance*). X is *probabilistically causally irrelevant* to Y given Z, written $(X \not\rightarrow Y \mid Z)_P$, iff

$$\forall x, x', y, z \quad P(y \mid \widehat{z}, \widehat{x}) = P(y \mid \widehat{z}, \widehat{x'}). \tag{13}$$

Read: "Once we hold Z fixed (at z), changing X between any two values will not change the probability of Y".

3.1. Comparison to informational relevance

If we remove the "hats" from Definition 7, we get the standard definition of conditional independence in probability calculus, denoted $(X \perp Y \mid Z)$, which is governed by the graphoid axioms [10,24] given in Fig. 4. Dawid [4] and Spohn [42] introduced these axioms in a different form, and Pearl and Paz [24] conjectured that these axioms were complete. This conjecture has been refuted by Studeny [44], who also proved that conditional independence in probability theory has no finite axiomatization. Nevertheless, the graphoid axioms capture the most important features of informational relevance: "Learning irrelevant information should not alter the relevance status of other propositions in the system; what was relevant remains relevant, and what was irrelevant remains irrelevant" [26].

One of the salient differences between informational and causal relevance is the property of symmetry, Axiom 1.1. Informational relevance is symmetric, namely, if X is relevant to Y, then Y is relevant to X as well. For example, learning whether the sprinkler is on provides information on whether the pavement is wet, and, vice versa, learning whether the pavement is wet provides information on whether the sprinkler is on. This property is clearly violated in causal models: turning a sprinkler on tends to make the pavement wet, so turning on the sprinkler gives us information about the state of the pavement; conversely, wetting the pavement has no physical effect on the state of the sprinkler and gives us no information about whether the sprinkler was on or off.

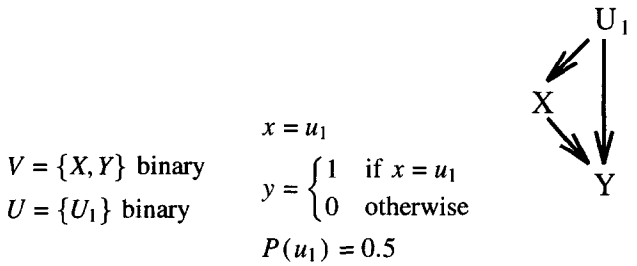


Fig. 5. An example of $P(y) > \mathrm{MAX}_x P(y \mid \hat{x})$.

Another basic difference between informational and causal relevance is that in the former, the rule of the *hypothetical middle* [26, p. 17] always holds:

$$\mathrm{MIN}_x P(y \mid x) \leqslant P(y) \leqslant \mathrm{MAX}_x P(y \mid x). \tag{14}$$

In causal relevance, $P(y)$ might be greater than $\mathrm{MAX}_x P(y \mid \hat{x})$ or less than $\mathrm{MIN}_x P(y \mid \hat{x})$. Fig. 5 illustrates such a possibility.

In Fig. 5, there are two endogenous variables X and Y, as well as an exogenous variable U_1. Without any intervention, X will always have the same value as U_1, thus, Y will have the value 1. If X and U_1 have different values, however, then Y will have the value 0. If we intervene and set $X = 1$, then Y will have the value 1 when $U_1 = 1$, which has a probability 0.5, and Y will have the value 0 when $U_1 = 0$, which has a probability 0.5: $P(Y = 0 \mid set(X = 1)) = P(Y = 1 \mid set(X = 1)) = 0.5$. Similarly, we can see that $P(Y = 0 \mid set(X = 0)) = P(Y = 1 \mid set(X = 0)) = 0.5$. Thus, $\mathrm{MAX}_x P(y \mid \hat{x}) = 0.5$, and $P(Y = 1) = 1 > 0.5 = \mathrm{MAX}_x P(y \mid \hat{x})$.

Note that, given this violation of the rule of the hypothetical middle (Eq. (14)), Definition 7 is not equivalent to

$$\forall x, y, z \quad P(y \mid \hat{z}, \hat{x}) = P(y \mid \hat{z}). \tag{15}$$

Read: "Once we hold Z fixed (at z), controlling X will not affect the probability of Y". In fact, Eq. (15 is stronger than Definition 7, furthermore, Statement 2.5.2 (left-intersection of Theorem 8, below) follows from the former but not from the latter.

The notion of probabilistic causal irrelevance may bring to mind the concept *ignorability* [35] which is extremely important in analyzing the effectiveness of treatments (e.g., drugs, diet, educational programs) from uncontrolled studies. The two concepts are related but different. Ignorability allows us to ignore *how* X obtained its value x, while irrelevance allows us to ignore *which* value X actually obtained. Ignorability is defined as the condition

$$P(Y_x = y \mid z) = P(Y = y \mid z, x), \tag{16}$$

which implies

$$P(y \mid \hat{x}) \doteq P(Y_x = y) = E_z(y \mid z, x). \tag{17}$$

Thus, ignorability allows an investigator to relate the potential response Y_x to observable conditional probabilities. Central in experimental design is the question of how to select a set of observables Z that would make Eq. (16) true, given causal knowledge of the domain. Ignorability in itself does not provide such a criterion although it does state the problem in formal counterfactual language: "Z can be selected if, for every x, the value that Y would obtain had X been x is conditionally independent of X, given Z". A practical criterion for selecting Z can be obtained from the graph $G(M)$ underlying a causal model (e.g., the back-door criterion in [29]).

The question we attempt to answer in this section is whether the relation of causal irrelevance, $(A \not\rightarrow B \mid C)_P$, is governed by a set of axioms similar to those governing the relation of informational irrelevance, $(A \perp B \mid C)$. More generally, one may ask whether there are any constraints that prohibit the assignment of arbitrary functions $P(y \mid \hat{x})$ to any pair (X, Y) of variable sets in V, in total disregard of the fact that $P(y \mid \hat{x})$ represents the probability of $(Y = y)$ induced by physically setting $X = x$ in some causal model M. Our finding indicate that, although the assignment $P(y \mid \hat{x})$ is not totally arbitrary, it is only weakly constrained by qualitative axioms such as those in Fig. 4.

3.2. Axioms of probabilistic causal relevance

We have found only two qualitative properties that constrain probabilistic causal irrelevance.

Theorem 8. *For any causal model, the following two properties must hold:*

2.2.1 (Right-Decomposition) $(X \not\rightarrow YW \mid Z)_P \Longrightarrow (X \not\rightarrow Y \mid Z)_P \,\&\, (X \not\rightarrow W \mid Z)_P$.

2.5.2 (Left-Intersection) $(X \not\rightarrow Y \mid ZW)_P \,\&\, (W \not\rightarrow Y \mid ZX)_P \Longrightarrow (XW \not\rightarrow Y \mid Z)_P$.

Property 2.2.1 is read: "If changing X has no effect on Y and W considered jointly, then it has no effect on either Y or W considered separately". This follows trivially from the fact that $P(\cdot)$ is a probability function, but it does not reflect any quality of causation.

Property 2.5.2 is read: "If changing X cannot affect $P(y)$ when W is fixed, and changing W cannot affect $P(y)$ when X is fixed, then changing X and W together cannot affect $P(y)$".

Many seemingly intuitive properties do not hold, however. For instance, none of the following sentences hold for all causal models.

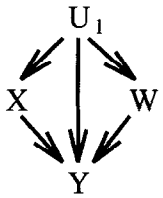
2.2.2 (Left-Decomposition-1) $(XW \not\rightarrow Y \mid Z)_P \Longrightarrow (X \not\rightarrow Y \mid Z)_P \vee (W \not\rightarrow Y \mid Z)_P$.

2.2.3 (Left-Decomposition-2) $(XW \not\rightarrow Y \mid Z)_P \Longrightarrow (X \not\rightarrow Y \mid Z)_P \vee (X \not\rightarrow W \mid Z)_P$.

2.2.4 (Left-Decomposition-3)
$(XW \not\rightarrow Y \mid Z)_P \,\&\, (XY \not\rightarrow Z \mid W)_P \Longrightarrow (X \not\rightarrow Y \mid Z)_P \vee (X \not\rightarrow W \mid Z)_P$.

2.3 (Weak Union) $(X \not\rightarrow WY \mid Z)_P \Longrightarrow (X \not\rightarrow Y \mid ZW)_P$.

2.4 (Contraction) $(X \not\rightarrow Y \mid Z)_P \,\&\, (X \not\rightarrow W \mid ZY)_P \Longrightarrow (X \not\rightarrow WY \mid Z)_P$.



$x = u_1$
$V = \{X, W, Y\}$ binary $\quad w = u_1$
$U = \{U_1\}$ binary $\quad y = Parity(x, w, u_1)$
$P(u_1) = 0.5$

Fig. 6. Counterexample to Property 2.2.2.

2.5.1 (Right-Intersection) $(X \not\to Y \mid ZW)_P \,\&\, (X \not\to W \mid XY)_P \Longrightarrow (X \not\to WY \mid Z)_P$.

2.6 (Transitivity)
$(X \not\to Y \mid Z)_P \Longrightarrow (a \not\to Y \mid Z)_P \vee (X \not\to a \mid Z)_P \; \forall a \notin X \cup Z \cup Y$.

The sentences above were tailored after the graphoid axioms (Fig. 4) with the provision that symmetry does not hold, which necessitates left and right versions of decomposition and intersection. Many of these sentences have intuitive appeal and yet are not sound relative to the semantics of $P(y \mid \hat{x})$. For example, Property 2.2.2 states, "If changing X has an effect on Y, and changing W has an effect on Y, then changing X and W simultaneously should also affect Y". A simple real-life example that refutes this assertion is difficult to come by. Still, as will be shown in Section 3.4 and in Appendix B, each of these sentences is refuted by some specific causal model.

3.3. Proofs of axioms of probabilistic causal irrelevance

We now prove the two sentences of Theorem 8.

2.2.1 $(X \not\to YW \mid Z)_P \Longrightarrow (X \not\to Y \mid Z)_P \,\&\, (X \not\to W \mid Z)_P$ holds trivially. $(X \not\to YW \mid Z)_P \Longrightarrow P(yw \mid \hat{z}, \hat{x}) = P(yw \mid \hat{z}, \hat{x}')$. We can sum over W to get $P(y \mid \hat{z}, \hat{x}) = P(y \mid \hat{z}, \hat{x}')$, which implies $(X \not\to Y \mid Z)_P$. A symmetric argument shows $(X \not\to W \mid Z)_P$. □

2.5.2 (By contradiction) Assume $(X \not\to Y \mid ZW)_P \,\&\, (W \not\to Y \mid ZX)_P \,\&\, \neg(XW \not\to Y \mid Z)_P$. Since $\neg(XW \not\to Y \mid Z)_P$, by definition of probabilistic causal irrelevance $\exists y, x, x', w, w', z \; P(y \mid \hat{z}, \hat{w}, \hat{x}) \neq P(y \mid \hat{z}, \hat{w}', \hat{x}')$. However, $(X \not\to Y \mid ZW)_P$ implies $\forall y, x, x', z, w \; P(y \mid \hat{z}, \hat{x}, \hat{w}) = P(y \mid \hat{z}, \hat{x}', \hat{w})$. Furthermore, $(W \not\to Y \mid ZX)_P$ implies $\forall y, x', w, w', z \; P(y \mid \hat{z}, \hat{x}', \hat{w}) = P(y \mid \hat{z}, \hat{x}', \hat{w}')$, so $\forall x, x', w, w', z \; P(y \mid \hat{z}, \hat{x}, \hat{w}) = P(y \mid \hat{z}, \hat{x}', \hat{w}) = P(y \mid \hat{z}, \hat{x}', \hat{w}')$. Thus $\forall x, x', w, w', z \; P(y \mid \hat{z}, \hat{x}, \hat{w}) = P(y \mid \hat{z}, \hat{x}', \hat{w}')$, which contradicts $\exists x, x', w, w', z \; P(y \mid \hat{z}, \hat{x}, \hat{w}) \neq P(y \mid \hat{z}, \hat{x}', \hat{w}')$. □

3.4. Counterexample to Property 2.2.2

We now disprove Property 2.2.2 by counterexample. This counterexample is not necessarily meant to model a common, real-life situation. Rather, it disproves the claim that *all possible* causal models must conform to the property.

2.2.2 $(XW \not\rightarrow Y \mid Z)_P \Longrightarrow (X \not\rightarrow Y \mid Z)_P \vee (W \not\rightarrow Y \mid Z)_P$.

Fig. 6 shows a counterexample to this sentence. In this model, $(XW \not\rightarrow Y \mid \emptyset)_P$ & $\neg(X \not\rightarrow Y \mid \emptyset)_P$ & $\neg(W \not\rightarrow Y \mid \emptyset)_P$. This counterexample is more clear when we consider its contrapositive form, which would state that changing W can affect the probability of Y, and changing X can affect the probability of Y, but changing W and X simultaneously has no effect on the probability of Y. This is extremely counterintuitive; if tweaking X has an effect on Y, and tweaking W has an effect on Y, we would expect the more flexible option of changing X and W simultaneously to also affect Y.

The key to this counterexample is the fact that setting W removes the connection between W and U_1. When we intervene on only X, W takes on the same value as U_1, and Y will always have the value of X. When we intervene on both X and W, there is no longer any connection between U_1 and W. Thus, the probability that W and U_1 will have the same value is 0.5, and $P(y) = 0.5$.

Counterexamples to the other six properties that do not hold for all causal models are in Appendix B.

3.5. Numeric constraints

Although Definition 7 imposes only weak constraints (Axioms 2.2.1 and 2.5.2) on the structure of probabilistic causal irrelevance, the probability assignments $P(y \mid \hat{x})$, which describe the effects of actions in the domain, are constrained nevertheless by nontrivial numerical bounds. For instance, the inequality

$$(y \mid \hat{x}, \hat{z}) \geqslant P(y, z \mid \hat{x}) \tag{18}$$

must hold in any causal model. This can easily be shown by the definition of $P(y, z \mid \hat{x})$ and $P(y \mid \hat{x}, \hat{z})$. Recall from Eq. (4) that

$$P(y, z \mid \hat{x}) = \sum_{\{u \mid Y_x(u)=y \,\&\, Z_x(u)=z\}} P(u), \qquad P(y \mid \hat{x}, \hat{z}) = \sum_{\{u \mid Y_{xz}(u)=y\}} P(u).$$

Consider U^{yz}, the set of all values u of U such that $Y_x(u) = Y$ and $Z_x(u) = z$, and U_z^y, the set of all values u' of U such that $Y_{xz}(u') = y$. Since all values u of U^{yz} already constrain Z to have the value z, fixing Z at z will not affect the value of Y. Thus, for all values u of U^{yz}, $Y_{xz}(u) = y$. Hence, $U_z^y \supseteq U^{yz}$ and $P(y \mid \hat{x}, \hat{z}) \geqslant P(yz \mid \hat{x})$. This can be shown more formally using Corollary 16 proven in Section 4.2. Additional constraints were explored in [30].

3.6. Axioms of causal relevance for stable models

The set of axioms we obtained for causal irrelevance is much smaller than we would expect from our intuition of cause–effect relations. We have two explanations for this discrepancy. One possibility is that our intuition of causal relevance is based on a deterministic rather than a probabilistic conception of physical reality. This possibility

will be explored in Section 4, which gives a deterministic definition of causal irrelevance that yields a more complete set of axioms. The other possibility is that the type of examples exploited in Section 3.4 and Appendix B are not commonly observed in everyday life. In this section, we explore what assumptions need to be made for probabilistic causal irrelevance to acquire properties that we intuitively associate with causal irrelevance.

A more expressive set of causal relevance axioms is obtained if we confine the analysis to *stable* causal models, that is, causal models whose irrelevances are implied by the structure of the causal model and, hence, remain invariant to changes in the forms of each individual function f_i. Our definition of stability employs the concept of a replacement class. A *replacement class* τ is the set of all models that have the same variables V and U, and the same functional arguments. In other words, the functions are allowed to change between members of τ, but the arguments of these functions are not allowed to vary. Formally, for any two models $M_1, M_2 \in \tau$ and any two functions $f_i(PA_i) \in M_1$ and $f_i'(PA_i') \in M_2$, $PA_i = PA_i'$. The class $\tau(M)$ represents the replacement class that contains the model M.

We now define stability using replacement classes (see also [25][6]).

Definition 9 (*Stability*). Let M be a causal model. An irrelevance $(X \not\rightarrow Y \mid Z)_P$ in M is *stable* if it is shared by all models in $\tau(M)$. The model M is *stable* if all of the irrelevances in M are stable.

Stability requires that irrelevance be determined by the structure of the equations, not merely by the parameters of the functions. Thus, a causal model is not stable if we can remove an irrelevance relationship by replacing an equation or set of equations to obtain a new model with fewer irrelevance statements. In each of the examples in Section 3.4 and Appendix B, for instance, a minor change in the form of one of the equations would destroy an irrelevance. Note that none of the models presented in Fig. 6 and the Appendix is stable.

There are, however, many stable causal models. All monotonic linear systems, for example, are stable. One might think that any causal models that contained only additive, monotonic functions f_i would be stable. The causal model of Fig. B.7 refutes that conjecture.

Definition 10 (*Path interception*). Let $(X \twoheadrightarrow Y \mid Z)_G$ stand for the statement "Every directed path from X to Y in graph G contains at least one element in Z".

Theorem 11. *If a causal model M is stable, then X is probabilistically causally irrelevant to Y, given Z, in M iff Z intercepts all directed paths from X to Y in the graph $G(M)$ defined by M. That is,*

$$(X \not\rightarrow Y \mid Z)_P \Longleftrightarrow (X \twoheadrightarrow Y \mid Z)_{G(M)}$$

[6] The probabilistic notion of stability (also called "DAG-isomorphism", "nondegeneracy" [26, p. 391], and "faithfulness" [41]) was used by Pearl and Verma [1991] to emphasize the invariance of certain independencies to functional form.

Proof. (i) $(X \not\to Y \mid Z)_P \Longrightarrow (X \nrightarrow Y \mid Z)_{G(M)}$. Assume that there exists a stable causal model M that induces a probabilistic causal irrelevance relation $(A \not\to B \mid C)_P$, and assume that, for some sets of variables X, Y, Z, $(X \not\to Y \mid Z)_P$ and $\neg(X \nrightarrow Y \mid Z)_{G(M)}$. Since there is a directed path from X to Y that is not intercepted by Z in $G(M)$, we can easily construct a model M' such that $G(M') = G(M)$ and $\neg(X \not\to Y \mid Z)_P$ in M'. We can do this by changing all of the functions that lie on the path from X to Y to disjunctions and then modifying the other functions to ensure that $P(y \mid \hat{z}) < 1$. Thus, if we force X to have the value 1, Y will also have the value 1, and $P(y \mid \hat{z}, \hat{x}) \neq P(y \mid \hat{z})$. By assumption, $(X \not\to Y \mid Z)_P$, so an irrelevance in M is not shared in a member of $\tau(M)$. Thus, M is not a stable causal model, a contradiction.

(ii) $(X \nrightarrow Y \mid Z)_{G(M)} \Longrightarrow (X \not\to Y \mid Z)_P$. We will use the following lemma:

Lemma 12. *For any structural equation f_Y in a causal model M, if a series of functional substitutions results in a new function g_Y such that X is an argument of g_Y, then there must be a directed path from X to Y in $G(M)$.*

We will prove this lemma by induction on the number of functional substitutions.

Base case: If we make no substitutions into f_Y, then every argument X of f_Y must be a parent of Y in $G(M)$, by our definition of $G(M)$. Thus, there is a directed path from each argument of f_Y to Y in $G(M)$.

Inductive case: Assume that $n-1$ functional substitutions into f_Y always results in the new function g_Y such that for each argument X of g_Y, there is a directed path from X to Y in $G(M)$. We use this assumption to prove that after n substitutions resulting in g'_Y, there is a directed path from every argument of g'_Y to Y in $G(M)$, as follows: When we do a single substitution, we replace a variable with a function of its parents in $G(M)$. So, for any new argument X' that is introduced into g'_Y by substituting in for X, X' must be a parent of X in $G(M)$. By the inductive hypothesis, there must be a directed path from X to Y in $G(M)$. Thus, there must be a directed path from X' to Y in $G(M)$.

We can now prove the implication $(X \nrightarrow Y \mid Z)_{G(M)} \Longrightarrow (X \not\to Y \mid Z)_P$. We will consider f_Y, the functional equation for Y in M_z. After we do a functional substitution for all variables in f_Y except for X and Z, we are left with a new function g_Y. By Lemma 12, since there is no directed path from X to Y in $G(M_z)$, X is not an argument of g_Y, so g_Y is a function of only Z and U. Since g_Y is a function of only Z and U, and not of X, $Y_{xz}(u) = Y_z(u)$, so $P(y \mid \hat{x}, \hat{z}) = P(y \mid \hat{z})$, and $(X \not\to Y \mid Z)_P$. □

Since $(X \not\to Y \mid Z)_P \Longleftrightarrow (X \nrightarrow Y \mid Z)_{G(M)}$ in stable causal models, probabilistic causal irrelevance is completely characterized by the axioms of path interception in directed graphs. A complete set of such axioms was developed in [22, 23] and is given in Fig. 7.

4. Deterministic causal relevance

The notion of causal irrelevance obtains a deterministic definition when we consider the effects of an action conditioned on a specific state of the world u.

3.2.1 (Right-Decomposition) $(X \dashrightarrow YW \mid Z)_G \Longrightarrow (X \dashrightarrow Y \mid Z)_G \,\&\, (X \dashrightarrow W \mid Z)_G$.

3.2.2 (Left-Decomposition) $(XW \dashrightarrow Y \mid Z)_G \Longrightarrow (X \dashrightarrow Y \mid Z)_G \,\&\, (W \dashrightarrow Y \mid Z)_G$.

3.4 (Strong Union) $(X \dashrightarrow Y \mid Z)_G \Longrightarrow (X \dashrightarrow Y \mid ZW)_G \; \forall W$.

3.5.1 (Right-Intersection) $(X \dashrightarrow Y \mid ZW)_G \,\&\, (X \dashrightarrow W \mid ZY)_G \Longrightarrow (X \dashrightarrow YW \mid Z)_G$.

3.5.2 (Left-Intersection) $(X \dashrightarrow Y \mid ZW)_G \,\&\, (W \dashrightarrow Y \mid ZX)_G \Longrightarrow (XW \dashrightarrow Y \mid Z)_G$.

3.6 (Transitivity) $(X \dashrightarrow Y \mid Z)_G \Longrightarrow (a \dashrightarrow Y \mid Z)_G \vee (X \dashrightarrow a \mid Z)_G \; \forall a \notin X \cup Z \cup Y$.

Fig. 7. Sound and complete axioms for path interception in directed graphs.

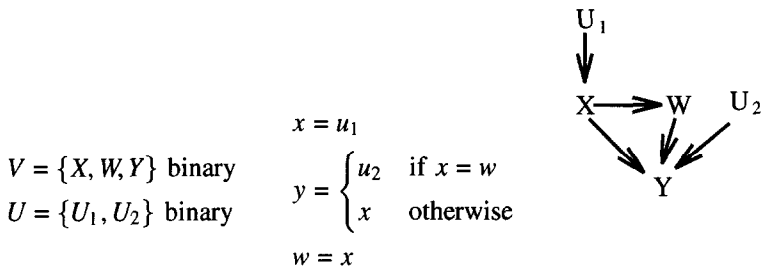


Fig. 8. Example of a causal model that requires the examination of submodels to determine causal relevance.

Definition 13 (*Causal irrelevance*). X is *causally irrelevant* to Y, given Z, written $(X \not\dashrightarrow Y \mid Z)$, if for every set W disjoint of $X \cup Y \cup Z$, we have:

$$\forall u, z, x, x', w \quad Y_{xzw}(u) = Y_{x'zw}(u). \tag{19}$$

This definition captures the intuition "If X is causally irrelevant to Y, then X cannot affect Y under any circumstance". It is stronger than the probabilistic definition, in that

$$(X \not\dashrightarrow Y \mid Z) \Longrightarrow (X \not\dashrightarrow Y \mid Z)_P.$$

Unlike the probabilistic definition of causal irrelevance (see Eq. (15)), the deterministic definition implies

$$\forall u, z, x \quad Y_{xz}(u) = Y_z(u). \tag{20}$$

To see why we require the equality $Y_{xzw}(u) = Y_{x'zw}(u)$ to hold in every context $W = w$, consider the causal model of Fig. 8. In this example, $Z = \{\emptyset\}$, W follows X and, hence, Y follows X, that is, $Y_{X=0}(u) = Y_{X=1}(u) = u_2$. However, since $y(x, w, u_2)$ is a nontrivial function of x, X is perceived to be causally relevant to Y. Only holding W constant would reveal the causal influence of X on Y. To capture this intuition, we must consider all contexts $W = w$ in Definition 13.

This definition of irrelevance bears some similarity to the idea of limited unresponsiveness presented in [14]. However, whereas Heckerman and Shacter define causality in terms of limited unresponsiveness to a specific set of actions, we view causal relevance as a property of the configuration of the mechanisms in a causal model. In fact, a version of their definition of causality, translated into our language, will be shown to be a theorem of causal relevance in Section 4.7.2 (see Eq. (27)).

4.1. Axioms of causal irrelevance

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

Theorem 14. *For any causal model, the following sentences must hold:*

4.2.1 (Right-Decomposition) $(X \not\rightarrow YW \mid Z) \Longrightarrow (X \not\rightarrow Y \mid Z)\ \&\ (X \not\rightarrow W \mid Z)$.

4.2.2 (Left-Decomposition) $(XW \not\rightarrow Y \mid Z) \Longrightarrow (X \not\rightarrow Y \mid Z)\ \&\ (W \not\rightarrow Y \mid Z)$.

4.4 (Strong Union) $(X \not\rightarrow Y \mid Z) \Longrightarrow (X \not\rightarrow Y \mid ZW)\ \forall W$.

4.5.1 (Right-Intersection) $(X \not\rightarrow Y \mid ZW)\ \&\ (X \not\rightarrow W \mid ZY) \Longrightarrow (X \not\rightarrow YW \mid Z)$.

4.5.2 (Left-Intersection) $(X \not\rightarrow Y \mid ZW)\ \&\ (W \not\rightarrow Y \mid ZX) \Longrightarrow (XW \not\rightarrow Y \mid Z)$.

Comparing to Fig. 7, we see that all axioms of path interception, except transitivity, are sound relative to deterministic causal relevance. The proof of Theorem 14 is in Section 4.4, below.

4.2. Properties of counterfactual statements

The axioms listed in the preceding section are based on three fundamental properties of counterfactuals, namely *composition*, *effectiveness*, and *reversibility*, which we will motivate using the action semantics of Definition 3.

Composition. For any two singleton variables Y and W and any set of variables Z in a causal model,

$$X_z(u) = x \Longrightarrow Y_{zw}(u) = Y_z(u) \tag{21}$$

Composition states that, in any context $Z = z$, if we force a variable X to a value x that it would have had without our intervention, then the intervention will have no effect on other variables in the system.

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

Definition 15 (*Null action*). $Y_\emptyset(u) \doteq Y(u)$.

Corollary 16 (Consistency). *For any variables Y and X in a causal model,*

$$X(u) = x \Longrightarrow Y(u) = Y_x(u) \tag{22}$$

Corollary 16 follows directly from composition and null action. The implication in Eq. (22) was called *consistency* by Robins [34].[7]

Effectiveness. For all variables X and W in a causal model,

$$X_{xw}(u) = x. \tag{23}$$

Effectiveness specifies the effect of an intervention on the manipulated variable itself, namely, that if we force a variable X to have the value x, then regardless of other enforcements $W = w$, X will indeed take on the value x.

Reversibility. For any two variables Y and W and any set of variables X in a causal model,

$$(Y_{xw}(u) = y) \ \& \ (W_{xy}(u) = w) \Longrightarrow Y_x(u) = y. \tag{24}$$

Reversibility reflects memoryless behavior—the state of the system, V, tracks the state of U, regardless of U's history. Given a context $X = x$ as in Eq. (24), if forcing W to a value w results in a value y for Y and forcing Y to y results, in turn, in $W = w$, then W and Y will have the values w and y, respectively, without any intervention. This follows from the requirement that the equations in every context $X = x$ have a unique solution. Thus, if we assume a solution $W = w$ and obtain $Y = y$ and, in turn, assuming a solution $Y = y$ yields $W = w$, then $(W = w, Y = y)$ is indeed the solution to the equations.

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

In recursive systems, reversibility follows directly from composition. This can easily be seen by noting that in a recursive system, either $Y_{xw}(u) = Y_x(u)$ or $W_{xy}(u) = W_x(u)$. Thus, reversibility reduces to $(Y_{xw}(u) = y) \ \& \ (W_x(u) = w) \Longrightarrow Y_x(u) = y$, which is another form of composition, or to $(Y_x(u) = y) \ \& \ (W_{xy}(u) = w) \Longrightarrow Y_x(u) = y$, which is trivially true. In non-recursive systems, reversibility is a property of causal loops. If forcing X to a value x results in a value y for Y, and forcing Y to the value y results

[7] Consistency and composition are tacitly used in economics [20] and statistics within the so-called Rubin's model [36]. To the best of our knowledge, Robins was the first to state consistency formally and to use it to derive other properties of counterfactuals [34]. Composition was brought to our attention by J. Robins (personal communication, February 1995). A weak version of composition is mentioned explicitly in [15, p. 968].

in X achieving the value x, then X and Y will have the values x and y, respectively, without any intervention.

4.3. Soundness of composition, effectiveness, and reversibility

Following the tradition of standard logic, we will consider a property of causal relationships to be *sound* if that property holds in all causal models.

Theorem 17. *Composition is sound.*

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

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

Theorem 18. *Effectiveness is sound.*

Proof. This theorem follows from Definition 1, where $Y_x(u)$ is interpreted as the unique solution for Y of a set of equations under $X = x$. □

Theorem 19. *Reversibility is sound.*

Proof. Reversibility follows from the assumption that the solution for V in every submodel is unique. Since $Y_x(u)$ has a unique solution, forming M_x and substituting out all other variables would yield a unique solution for Y, regardless of the order of substitution. So, we will form M_x and examine the structural equation for Y in M_x, which in general might be a function of X,W,U, and additional variables: $Y_x = f_Y(x, w, z, u)$, where Z stands for parents of Y not contained in $X \cup W \cup U$. We now solve for Z by substituting out all variables except X,Y, and W. That is, we substitute out all variables in M_x, without substituting into X, W, and Y and express Z as a function of x, w, and u. We then plug this solution into f_Y to get $Y_x = f_Y(x, w, Z(x, w, u), u)$, which we can write as $Y_x = f(x, w, u)$. We now consider what would happen if we solved for Y in M_{xw}. Since we avoided substituting anything into W when we solved for Y in M_x, we will get the same result as before, namely, $Y_{xw} = f(x, w, u)$. In the same way, we can show that $W_x = g(x, y, u)$ and $W_{xy} = g(x, y, u)$. So, solving for $y = Y_x(u)$, $w = W_x(u)$ is the same as solving for $y = f(x, w, u)$ and $w = g(x, y, u)$, which is the same as solving for $y = Y_{xw}(u)$, $w = W_{xy}(u)$. Thus, any solution y to $y = Y_{xw}(u), w = W_{xy}(u)$ is also a solution to $y = Y_x(u)$. □

Given a causal ordering of variables in V, that is, $Y_{xz}(u) = Y_z(u)$ for any set Z whenever Y precedes X in the ordering, one can show that effectiveness and composition are complete [9]. Joseph Halpern [13] has recently shown that composition, reversibility, and effectiveness are complete in all causal models, recursive as well as nonrecursive, as long as the uniqueness assumption holds.

4.4. Proofs of causal relevance axioms

Using the properties from Section 4.2, we can prove Theorem 14, that the axioms of causal relevance are sound.

4.2.1 Holds trivially. □

4.2.2 (By contradiction) Assume that there exists a causal model such that $(XW \not\rightarrow Y \mid Z) \,\&\, \neg((Z \not\rightarrow Y \mid Z) \,\&\, (W \not\rightarrow Y \mid Z))$. So, either $(XW \not\rightarrow Y \mid Z) \,\&\, \neg(X \not\rightarrow Y \mid Z)$ or $(XW \not\rightarrow Y \mid Z) \,\&\, \neg(W \not\rightarrow Y \mid Z)$.

First, we consider $(XW \not\rightarrow Y \mid Z) \,\&\, \neg(X \not\rightarrow Y \mid Z)$. By our definition of causal irrelevance, $\neg(X \not\rightarrow Y \mid Z)$ implies that there exist two values x, x' of X and some value u of U such that $Y_{xz}(u) \neq Y_{x'z}(u)$. Now, let us consider the values x, x', z, u such that $Y_{xz}(u) \neq Y_{x'z}(u)$. Using these values, we can determine w and w' as follows: Let $w = W_{xz}(u)$, and $w' = W_{x'z}(u)$. It does not matter whether $w = w'$ or $w \neq w'$. By composition, $Y_{xzw}(u) \neq Y_{x'zw}(u)$. Thus, $\exists x, w, z, u\ Y_{xwz}(u) \neq Y_{x'w'z}(u)$, which contradicts $(XW \not\rightarrow Y \mid Z)$. Thus, $(XW \not\rightarrow Y \mid Z) \,\&\, \neg(X \not\rightarrow Y \mid Z)$ leads to a contradiction.

We can use a symmetric argument to show that $(XW \not\rightarrow Y \mid Z) \,\&\, \neg(W \not\rightarrow Y \mid Z)$ also leads to a contradiction. □

4.4 By our definition of causal irrelevance, $(X \not\rightarrow Y \mid Z) \Longrightarrow Y_{xz}(u) = Y_{x'z}(u)$ for all submodels of M_{xz}. For an arbitrary W, we consider the submodel M_w where W is forced to have the value w. By our definition of causal irrelevance, $Y_{xzw}(u) = Y_{x'zw}$ for all values w. In addition, since $(X \not\rightarrow Y \mid Z) \Longrightarrow Y_{xz}(u) = Y_{x'z}(u)$ for all submodels of M, $Y_{xzw}(u) = Y_{x'zw}$ for all submodels of M_w. Since W was arbitrary, $(X \not\rightarrow Y \mid Z) \Longrightarrow (X \not\rightarrow Y \mid ZW)$ for all W. □

4.5.1 (By contradiction) Assume $(X \not\rightarrow Y \mid ZW) \& (X \not\rightarrow W \mid ZY) \& \neg(X \not\rightarrow YW \mid Z)$. $\neg(X \not\rightarrow YW \mid Z)$ implies $\exists x, x', z\ (Y_{xz}(u) \neq Y_{x'z}(u)) \vee (W_{xz}(u) \neq W_{x'z}(u))$. Since W and Y are symmetric, we will only consider Y. Consider the values of x, x', z, u such that $Y_{xz}(u) \neq Y_{x'z}(u)$. Let $y = Y_{xz}(u)$ and $y' = Y_{x'z}(u)$.

By composition, $Y_{xz}(u) = Y_{xzw}(u)$ for $w = W_{xz}(u)$. By assumption, $Y_{xzw}(u) = Y_{x'zw}(u)$. Also by composition, $W_{xz}(u) = W_{xzy}(u)$ for $y = Y_{xz}(u)$. By assumption, $W_{xzy}(u) = W_{x'zy}(u)$. By reversibility, since y is a solution to the simultaneous equations $y = Y_{x'zw}$ and $w = W_{x'zy}$, then y must also be a solution to $Y_{x'z}(u)$. Thus $y = y'$, a contradiction. We can use a symmetric argument to show that $W_{xz}(u) \neq W_{x'z}(u)$ also leads to a contradiction. □

4.5.2 (By contradiction) Assume $(X \not\to Y \mid ZW)$ & $(W \not\to Y \mid ZX)$ & $\neg(XW \not\to Y \mid Z)$. Since $\neg(XW \not\to Y \mid Z)$, by definition $\exists x, x', w, w', z\ Y_{xwz}(u) \neq Y_{x'w'z}(u)$. However, $(X \not\to Y \mid ZW)$ implies $\forall x, x', z, w\ Y_{xzw}(u) = Y_{x'zw}(u)$. Furthermore, $(W \not\to Y \mid ZX)$ implies $\forall x', w, w', z\ Y_{x'wz}(u) = Y_{x'w'z}(u)$. Thus, $\forall x, x', w, w', z\ Y_{xwz}(u) = Y_{x'wz}(u) = Y_{x'w'z}(u)$, thus $\forall x, x', w, w', z\ Y_{xwz}(u) = Y_{x'w'z}(u)$. This contradicts $\exists x, x', w, w', z\ Y_{xwz}(u) \neq Y_{x'w'z}(u)$. □

4.5. Causal relevance and Lewis' counterfactuals

It is instructive to compare our framework to that of Lewis [18]. We give here a version of Lewis' logic for counterfactual sentences (from [19]).

Rules

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

Axioms

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

The statement $A \,\square\!\!\rightarrow B$ stands for "In all closest worlds where A holds, B holds as well". Lewis is careful to not put any restrictions on definitions of closest worlds, beyond the obvious requirement that world w be no further from itself than any other $w' \neq w$. In essence, causal models with local interventions define an ordering among worlds that gives a metric by which to define what worlds are closest. As such, all of Lewis' axioms are true for causal models and follow from effectiveness, composition, and (for nonrecursive systems) reversibility.

In order to relate Lewis' axioms to our framework, we need to translate his syntax into the language of causal models. We will equate Lewis' "world" with an instantiation of all variables in a causal model, including the variables in U. Propositions, such as A and B in the statements above, will be limited to the assignment of values to subsets of variables in a model. Thus, the meaning of the statement $A \,\square\!\!\rightarrow B$ in causal models is "If we force a set of variables to have the values A, a second set of variables will have the values B". Let A stand for a set of values $x_1, \ldots, x_n$ of the variables $X_1, \ldots, X_n$, and let B stand for a set of values $y_1, \ldots, y_m$ of the variables $Y_1, \ldots, Y_m$. Then,

$$\begin{aligned} A \,\square\!\!\rightarrow B \equiv\ & Y_{1\,x_1 \ldots x_n}(u) = y_1 \,\& \\ & Y_{2\,x_1 \ldots x_n}(u) = y_2 \,\& \\ & \ldots \\ & Y_{2\,x_1 \ldots x_n}(u) = y_n \,\& \end{aligned} \tag{25}$$

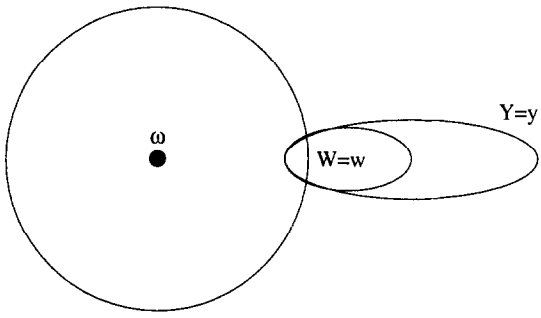


Fig. 9. Example of the failure of reversibility in Lewis' framework: $W = w$ holds in all closest y-worlds, and $Y = y$ holds in all closest w-worlds, yet $Y \neq Y$ and $W \neq w$.

Conversely, we need to define what statements such as $Y_x(u) = y$ mean in Lewis' notation. Let A stand for the proposition $X = x$, and B stand for the proposition $Y = y$. Then,

$$Y_x(u) = y \equiv A \,\square\!\!\rightarrow B \tag{26}$$

We can now examine each of Lewis' axioms in turn.

(1) Trivially true.

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

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

(4) Since actions in causal models are restricted to conjunctions of literals, this axiom does not apply. However, under the interpretation $do(A \vee B) \equiv do(A) \vee do(B)$, this axiom does hold.

(5) This axiom follows directly from composition.

(6) This axiom follows directly from composition.

Likewise, composition and effectiveness follow from Lewis' axioms. Composition is a consequence of Lewis' axiom (5) and rule (1), while effectiveness is Lewis' axiom (2). Thus, causal models do not add any restrictions to counterfactual statements above those imposed by Lewis' framework, when we are considering recursive models. When we consider nonrecursive systems, we see that reversibility is not enforced by Lewis' framework. Lewis' axiom (3), while similar, is not as strong as reversibility. For instance, $Y = y$ may hold in all closest w-worlds, $W = w$ may hold in all closest y-worlds and, still, $Y = y$ may not hold in our world. A graphical example violating reversibility in Lewis' framework is given in Fig. 9.

4.6. Why transitivity fails in causal relevance

Causal transitivity is a property that makes intuitive sense. If a variable A has a causal influence on B, and B has a causal influence on C, one would think that A would have

$$V = \{X, W, Y\},\ x, y \in \{0, 1\},\ w \in \{0, 1, 2, 3\}$$
$$U = \{U_1, U_2\},\ u_1, u_2 \in \{0, 1\}$$

$$w = x + 2 * u_2$$
$$x = u_1$$
$$y = (w > 1)$$

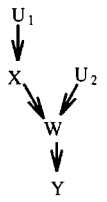


Fig. 10. Counterexample to transitivity in causal irrelevance.

$$V = \{X, W_1, W_2, W_3, W_4, Y\} \text{ binary}$$
$$U = \{U_1, U_2\} \text{ binary}$$

$$x = u_1$$
$$w_1 = \neg x\ \&\ \neg u_2$$
$$w_2 = x\ \&\ \neg u_2$$
$$w_3 = \neg x\ \&\ u_2$$
$$w_4 = x\ \&\ u_2$$
$$y = (w_3\ \&\ \neg w_1\ \&\ \neg w_2)\ \vee\ (w_4\ \&\ \neg w_1\ \&\ \neg w_2)$$
$$P(u_1) = P(u_2) = 0.5$$

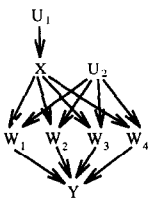


Fig. 11. Transitivity fails, even when a variable is more completely controlled by its parents than in Fig. 10.

a causal influence on C. This is not always the case, however, even in deterministic causality. Consider the causal model described in Fig. 10. In this example, X is causally relevant to W, and W is causally relevant to Y, but X is causally irrelevant to Y. The intuition behind this example is that changing X can only cause a minor change in W, while Y only responds to large changes in W. However, the failure of transitivity is deeper than this. Even when X has more complete control over the intermediate variable W, we still may not be able to achieve transitivity. Consider the causal model of Fig. 11.

This model is the same as the model of Fig. 10 except W has now been split into $W_1, \ldots, W_4$, corresponding to W's four possible values. That is, W_1 is true if $x + u_2 = 0$, W_2 is true if $x + u_2 = 1$, W_3 is true if $x + u_2 = 2$, and W_4 is true if $x + u_2 = 3$. Now, by fixing X, we can cause any of the intermediate variables $W_1, \ldots, W_4$ to be false in any given state of the world u. Likewise, each of the intermediate variables $W_1, \ldots, W_4$ can affect Y in any state u. However, X has no effect on Y in any state u.

4.7. Causal relevance and directed graphs

4.7.1. Causal graphs as irrelevance-maps

Comparing Axioms 3.2–3.5 to Axioms 4.2–4.5, we see that causal irrelevance is quite similar to path interception in directed graphs. Since people (and machines) can easily reason about graphs it would be useful to create a graph that represents all of the causal relevances and irrelevances of a given causal model. That is, we would like to create a graph $G^*(M)$ such that

(i) Each variable X in M corresponds to exactly one node X^* in $G^*(M)$,

(ii) For all subsets of nodes X^*, Y^*, Z^* in $G^*(M)$, $(X^* \twoheadrightarrow Y^* \mid Z^*)_{G^*(M)} \Longrightarrow (X \not\rightarrow Y \mid Z)$, and

(iii) For all subsets of variables X, Y, Z in M, $(X \not\rightarrow Y \mid Z) \Longrightarrow (X^* \twoheadrightarrow Y^* \mid Z^*)_{G^*(M)}$.

In such a graph $G^*(M)$, if all directed paths from X^* to Y^* were intercepted by some variables in Z, then X would be causally irrelevant to Y in the model M. Likewise, if a set of variables X was causally irrelevant to a set Y given fixed Z, then all paths from nodes in X^* to nodes in Y^* would be intercepted by some variables in Z.

The obvious choice for $G^*(M)$ is $G(M)$, the graph associated with the causal model itself, as defined by Eq. (1). If we use $G^*(M) = G(M)$, then implication (ii) holds, since in Section 3.6 we showed that $(X \twoheadrightarrow Y \mid Z)_{G(M)} \Longrightarrow Y_{xz}(u) = Y_z(u)$, and thus $(X \not\rightarrow Y \mid Z)$. However, since transitivity always holds in path interception but does not always in causal irrelevance, for a given model M there might be no graph $G^*(M)$ such that implications (ii) and (iii) hold simultaneously. Nonetheless, we can use directed graphs to validate candidate theorems of causal irrelevance, as we show below.

4.7.2. Directed graphs as theorem provers

Consider an oracle that takes in statements about path interception and returns YES if the statement holds in all directed graphs and NO otherwise. We will show that such an oracle can be used to validate or refute sentences about causal relevance.

First, let us consider a language of causal relevance in which the literals stand for simple irrelevance statements of the form $(X \not\rightarrow Y \mid Z)$, where X, Y and Z are sets of variables. Second, let the *canonical form* for sentences in the language of causal irrelevance be an implication $a_1 \,\&\, a_2 \,\&\, \cdots \,\&\, a_i \Longrightarrow b_1 \vee b_2 \vee \cdots \vee b_k$, whose antecedent consists of a conjunction of non-negated literals and whose consequent consists of non-negated literals. For instance, consider the sentence[8]

$$(X \not\rightarrow Y \mid Z) \,\&\, \neg(X \not\rightarrow Y \mid \emptyset) \Longrightarrow \neg(Z \not\rightarrow Y \mid \emptyset). \tag{27}$$

This sentence is not in canonical form because the second conjunct in the antecedent is negated and the statement in the consequent is negated. The canonical form of this sentence is

$$(X \not\rightarrow Y \mid Z) \,\&\, (Z \not\rightarrow Y \mid \emptyset) \Longrightarrow (X \not\rightarrow Y \mid \emptyset). \tag{28}$$

Any causal irrelevance sentence can be written in a unique canonical form using standard logical procedures.

[8] A version of this sentence was chosen in [14] as the definition of causality.

Definition 20 (*Horn component*). A *Horn component* H of a causal irrelevance sentence S is a sentence H such that

(i) H is in canonical form,
(ii) the consequent of H contains no disjunctions, and
(iii) $H \Longrightarrow S$.

If a sentence S is in the canonical form $a_1 \,\&\, a_2 \,\&\, \cdots \,\&\, a_i \Longrightarrow b_1 \vee b_2 \vee \cdots \vee b_k$, then a Horn component of S is any sentence of the form $a_1 \,\&\, a_2 \,\&\, \cdots \,\&\, a_i \Longrightarrow b_j$. For example, Eq. (28) has no disjunctions in its consequent and, hence, is itself a Horn component.

For any causal irrelevance statement A of the form $(X \not\rightarrow Y \mid Z)$, we will consider A_g, the *graphical translation* of A to be the corresponding path-interception statement $(X \nrightarrow Y \mid Z)_{G(M)}$. Using this convention, we can define

Theorem 21 (Graphical theorem verification). *A causal irrelevance sentence S is true for all causal models iff there exists a Horn component H of S such that H_g, the graphical translation of H, is true for all graphs.*

For example, consider the sentence in Eq. (27). The canonical form of this sentence is given in Eq. (28), and is itself a Horn component. The sentence corresponding to Eq. (28) for path interception in directed graphs, $(X \nrightarrow Y \mid Z)_G \,\&\, (Z \nrightarrow Y \mid \emptyset)_G \Longrightarrow (X \nrightarrow Y \mid \emptyset)_G$, states that if all paths from X to Y are intercepted by Z, and there are no paths from Z to Y, then there is no path from X to Y. This sentence is true for all directed graphs, so Eq. (27) is a valid theorem of causal relevance.

Next, consider transitivity, stated as $(X \not\rightarrow Y \mid Z) \Longrightarrow (a \not\rightarrow Y \mid Z) \vee (X \not\rightarrow a \mid Z)$. The Horn components of this sentence are

$$H^1\colon \quad (X \not\rightarrow Y \mid Z) \Longrightarrow (a \not\rightarrow Y \mid Z), \tag{29}$$

$$H^2\colon \quad (X \not\rightarrow Y \mid Z) \Longrightarrow (X \not\rightarrow a \mid Z). \tag{30}$$

Looking at each of the corresponding path-interception sentences in turn, we find that $H^1_g : (X \nrightarrow Y \mid Z)_G \Longrightarrow (a \nrightarrow Y \mid Z)_G$ is not true for all directed graphs G, and $H^2_g : (X \nrightarrow Y \mid Z)_G \Longrightarrow (X \nrightarrow a \mid Z)_G$ is also not true for all directed graphs G, that is, if Z intercepts all paths from X to Y, it is not the case that either Z intercepts all paths from any other variable to Y or Z intercepts all paths from X to any other variable. Thus, transitivity is not a theorem of causal relevance.

Proof of Theorem 21. First, we prove that if there are no disjunctions in the consequent of a canonical form sentence, then the sentence is true iff the corresponding sentence is true for path interception in directed graphs.

We will prove this by contradiction. Assume that there exists some theorem $A \Longrightarrow B$, where A and B are conjunctions of literals such that

- $A \Longrightarrow B$ is not a theorem in causal irrelevance, and
- $A_g \Longrightarrow B_g$ is a theorem in path interception in directed graphs.

Since $A_g \Longrightarrow B_g$ is a theorem in path interception, then we must be able to generate B_g from A_g using the axioms of path interception in directed graphs. However, since

$A \Longrightarrow B$ is not a theorem in causal irrelevance, every such generation of B_g from A_g must include the application of the axiom of transitivity. When the axiom of transitivity is used, a disjunction is created. This disjunction must be used in the generation of B_g. By assumption, B_g does not contain a disjunction. Also, none of the antecedents of any of the axioms of path interception contain disjunctions. Thus the only way to use this disjunction in the generation of B_g is to resolve the disjunction with a negated clause. Since A_g started with no negated statements, and none of the axioms of path interception can be used to create negated statements, we cannot resolve the disjunction with anything. Thus, generating B_g from A_g did not require an application of transitivity, a contradiction.

Next, we prove that if a theorem $A \Longrightarrow B \vee C$ is a theorem in causal irrelevance, then either $A \Longrightarrow B$ is a theorem in causal irrelevance or $A \Longrightarrow C$ is a theorem in causal irrelevance. If $A \Longrightarrow B \vee C$ is a theorem in causal irrelevance, then we must be able to generate $B \vee C$ from A using the axioms of causal irrelevance. Since no axiom creates a disjunction, to generate $B \vee C$ from A we must either generate B from A and add C or generate C from A and add B.

Thus, a causal irrelevance sentence is a theorem iff there is a path-interception theorem that corresponds to one of the Horn components of the original sentence. □

5. Conclusion

How do scientists predict the outcome of one experiment from the results of other experiments run under totally different conditions? Such transfer of experimental knowledge involves inferences that cannot easily be formalized in the standard languages of logic, physics, or probability.

The formalization of such inferences requires a language within which the experimental conditions prevailing in one experiment can be represented, such that the outcome of the experiment can be posed as constraint in the design and analysis of the next experiment. The description of experimental conditions, in turn, involves both observational and manipulative sentences, and it requires that manipulative phrases (e.g., "having no effect on", "holding Z fixed"), as distinct from observational phrases (e.g., "being independent of", "conditioning on Z"),[9] be given formal notation, semantical interpretation, and axiomatic characterization. It turns out that standard algebras, including the algebra of equations, Boolean algebra, and probability calculus, are all geared to serve observational but not manipulative sentences.

This paper bases the semantics of manipulative sentences on a set of structural equations that we call a *causal model*. Unlike ordinary algebraic equations, a causal model treats every equation as an independent mathematical object attached to one and only one variable. Actions are treated as modalities and are interpreted as the nonalgebraic operator of replacing equations.

[9] Philosophers, statisticians, and economists have often confused "holding Z constant" with "conditioning on a given Z" [29].

This semantics permits us to develop an axiomatic characterization of manipulative statements of the form "Changing X will not affect Y if we hold Z constant". This axiomatization highlights the differences between causal irrelevance, as in "X is causally irrelevant to Y in context Z", and informational irrelevance, as in "Finding X will not affect our belief in Y, once we know Z". The former shows a closer affinity to graphical representation than the latter. Under the deterministic definition, causal irrelevance complies with all of the axioms of path interception in cyclic graphs except transitivity. This affinity leads to graphical methods for proving theorems about causal relevance and explains, in part, why graphs are so prevalent in causal talk and causal modeling.

Outside of artificial intelligence, our results have interesting ramifications in the fields of statistics and epidemiology where, thus far, the only accepted formalization of causation has been Rubin's framework of counterfactuals [33,36], which is a rather cumbersome language for expressing causal knowledge. Graphical and structural equation models, popular as they are in econometrics and the social sciences, are viewed with suspicion by statisticians because the causal interpretation of these models has not been adequately formalized [8,46].

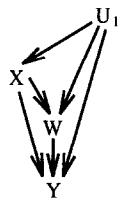
Our translation of counterfactuals into statements about structural equation models (Definition 5) generalizes and unifies the structural and counterfactual approaches, and greatly clarifies their conceptual and mathematical bases. The soundness of effectiveness and composition—the only properties of counterfactuals used in Rubin's framework—assures that every theorem in that framework is also a theorem in structural equations models. The completeness of effectiveness and composition in recursive models [9] further assures that the structural interpretation of counterfactuals introduces no extraneous properties beyond those embodied in Rubin's framework. Most significantly, this unification permits investigators to express causal knowledge in the intuitively appealing language of causal graphs, use the graphs as inferential machinery and be assured of the validity of the results.

Acknowledgments

This research was partially supported by Air Force grant #AFOSR/F496209410173, NSF grant #IRI-9420306, and Rockwell/Northrop Micro grant #94-100. We thank an anonymous reviewer for providing insightful and extremely helpful suggestions on the first draft of the paper. We also thank Joe Halpern for commenting on the first draft of this paper and for noting that Property 4.5.1 does not hold in Lewis' closest-world framework.

Appendix A. Independence of composition, effectiveness, and reversibility

We show that reversibility, composition, and effectiveness are independent by creating a table of counterfactual statements such that two of the properties hold but the third



$V = \{X, W, Y\}$ binary
$U = \{U_1\}$ binary

$x = u_1$
$w = (x \ \& \ u_1)$
$y = Parity(x, w, u_1)$
$P(u_1) = 0.5$

Fig. B.1. Counterexample to Property 2.2.3.

does not. We will consider a small model, one with only two binary variables X and Y and a single value for U.

A.1. Composition and effectiveness, not reversibility

$X = 0$	$Y = 0$		
$X_{X=0} = 0$	$Y_{X=0} = 0$	$X_{X=0,Y=0} = 0$	$Y_{X=0,Y=0} = 0$
$X_{X=1} = 1$	$Y_{X=1} = 1$	$X_{X=0,Y=1} = 0$	$Y_{X=0,Y=1} = 1$
$X_{Y=0} = 0$	$Y_{Y=0} = 0$	$X_{X=1,Y=0} = 1$	$Y_{X=1,Y=0} = 0$
$X_{Y=1} = 1$	$Y_{Y=1} = 1$	$X_{X=1,Y=1} = 1$	$Y_{X=1,Y=1} = 1$

A.2. Effectiveness and reversibility, not composition

$X = 0$	$Y = 1$		
$X_{X=0} = 0$	$Y_{X=0} = 1$	$X_{X=0,Y=0} = 0$	$Y_{X=0,Y=0} = 0$
$X_{X=1} = 1$	$Y_{X=1} = 0$	$X_{X=0,Y=1} = 0$	$Y_{X=0,Y=1} = 1$
$X_{Y=0} = 0$	$Y_{Y=0} = 0$	$X_{X=1,Y=0} = 1$	$Y_{X=1,Y=0} = 0$
$X_{Y=1} = 1$	$Y_{Y=1} = 1$	$X_{X=1,Y=1} = 1$	$Y_{X=1,Y=1} = 1$

A.3. Composition and reversibility, not effectiveness

$X = 0$	$Y = 1$		
$X_{X=0} = 0$	$Y_{X=0} = 1$	$X_{X=0,Y=0} = 0$	$Y_{X=0,Y=0} = 1$
$X_{X=1} = 0$	$Y_{X=1} = 1$	$X_{X=0,Y=1} = 0$	$Y_{X=0,Y=1} = 1$
$X_{Y=0} = 0$	$Y_{Y=0} = 1$	$X_{X=1,Y=0} = 0$	$Y_{X=1,Y=0} = 1$
$X_{Y=1} = 0$	$Y_{Y=1} = 1$	$X_{X=1,Y=1} = 0$	$Y_{X=1,Y=1} = 1$

Appendix B. Counterexamples

2.2.3 $(XW \not\rightarrow Y \mid Z)_P \Longrightarrow (X \not\rightarrow Y \mid Z)_P \vee (X \not\rightarrow W \mid Z)_P$.
In the causal model of Fig. B.1, we can see that

$$(XW \not\rightarrow Y \mid \emptyset)_P \ \& \ \neg(X \not\rightarrow W \mid \emptyset)_P \ \& \ \neg(X \not\rightarrow Y \mid \emptyset)_P.$$

$V = \{X, W, Y\}$ binary
$U = \{U_1, U_2\}$ binary

$$x = u_1$$
$$y = \begin{cases} u_1 & \text{if } u_2 = 0 \\ Parity(x, w, u_1) & \text{if } u_2 = 1 \end{cases}$$
$$w = \begin{cases} Parity(x, y, u_1) & \text{if } u_2 = 0 \\ u_1 & \text{if } u_2 = 1 \end{cases}$$
$$P(u_1) = P(u_2) = 0.5$$

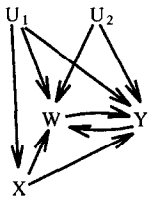


Fig. B.2. Counterexample to Property 2.2.4.

In this counterexample, changing X can affect the probability of Y, and changing X can affect the probability of W, but changing X and W together cannot affect the probability of Y. Since changing X affects the value of W, it makes sense to think that intervening on W while intervening on X would not interfere with the effect that X has on Y. However, X does not completely control W. That is, when we only intervene on X, U_1 still has some effect on W. Controlling both X and Y removes the influence of U_1 on W. As in the counterexample to Property 2.2.2, removing the connection between U_1 and W prevents X from having an effect on Y.

2.2.4 $(XW \not\rightarrow Y \mid Z)_P \,\&\, (XY \not\rightarrow W \mid Z)_P \Longrightarrow (X \not\rightarrow Y \mid Z)_P \vee (X \not\rightarrow W \mid Z)_P$.

In Fig. B.2, we can see that

- $P(w) = P(y) = 0.5$;
- $P(w \mid set(X = 1)) = P(y \mid set(X = 1)) = 0.75$;
- $P(w \mid \hat{x}, \hat{y}) = 0.5$ for all values of $\hat{x}$, $\hat{y}$; and
- $P(y \mid \hat{x}, \hat{w}) = 0.5$ for all values of $\hat{x}$, $\hat{w}$.

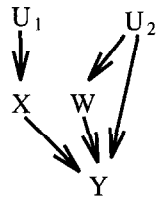
Thus, $(XW \not\rightarrow Y \mid \emptyset)_P \,\&\, (XY \not\rightarrow W \mid \emptyset)_P \,\&\, \neg((X \not\rightarrow Y \mid \emptyset)_P \vee (X \not\rightarrow W \mid \emptyset)_P)$.

This counterexample actually contains two causal models, each similar to the model of counterexample 2.2.2 (see Section 3.4, Fig. 6). In one, W is a function of X, Y, and U_1, and Y is a function of U_1. As for Property 2.2.2, X can affect W when Y has the same value as U_2, but X has no effect on $P(w)$ when Y is held constant. In the other, W is a function of U_1, and Y is a function of X, W, and U_1. Also as in Property 2.2.2, X can affect Y when W has the same value as U_1, but it has no effect on $P(w)$ when W is fixed. U_2 determines which model is in effect at any given time. While intervening on only X can affect $P(w)$ and $P(y)$, simultaneously changing X and Y has no effect on $P(w)$, and simultaneously changing X and W has no effect on $P(y)$.

2.3 $(X \not\rightarrow WY \mid Z)_P \Longrightarrow (X \not\rightarrow Y \mid ZW)_P$.

In the causal model of Fig. B.3, $(X \not\rightarrow YW \mid \emptyset)_P \,\&\, \neg(X \not\rightarrow W \mid Y)_P$.

In this counterexample, X does not have any effect on Y since $P(y) = 0$, and X can only act as an inhibitor of Y. When we intervene on W, then it is possible for Y to have the value 1, and X can affect the probability of Y. Thus, X can only affect Y when we intervene on W, and X has no effect on W.



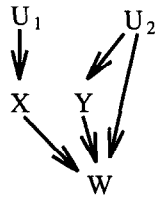
$x = u_1$

$V = \{X, W, Y\}$ binary $\quad w = u_2$

$U = \{U_1, U_2\}$ binary $\quad y = (x \ \& \ (w \text{ XOR } u_2))$

$P(u_1) = P(u_2) = 0.5$

Fig. B.3. Counterexample to Property 2.3.



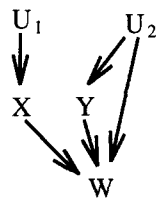
$x = u_1$

$V = \{X, W, Y\}$ binary $\quad y = u_2$

$U = \{U_1, U_2\}$ binary $\quad w = \mathit{Parity}(x, y, u_2)$

$P(u_1) = P(u_2) = 0.5$

Fig. B.4. Counterexample to Property 2.4.



$x = u_1$

$V = \{X, W, Y\}$ binary $\quad y = u_2$

$U = \{U_1, U_2\}$ binary $\quad w = \mathit{Parity}(x, y, u_2)$

$P(u_1) = P(u_2) = 0.5$

Fig. B.5. Counterexample to Property 2.5.1.

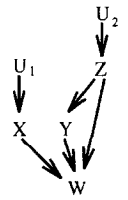
2.4 $(X \not\rightarrow Y \mid Z)_P \ \& \ (X \not\rightarrow W \mid ZY)_P \Longrightarrow (X \not\rightarrow WY \mid Z)_P$.

In the causal model of Fig. B.4, $(X \not\rightarrow Y \mid \emptyset)_P \ \& \ (X \not\rightarrow W \mid Y)_P \ \& \ \neg(X \not\rightarrow WY \mid \emptyset)_P$.

While changing X can affect $P(w)$ (and hence $P(y, w)$) when Y is not held fixed, and changing X has no effect on $P(y)$, fixing Y blocks the effect that X has on W.

2.5.1 $(X \not\rightarrow Y \mid ZW)_P \ \& \ (X \not\rightarrow W \mid ZY)_P \Longrightarrow (X \not\rightarrow WY \mid Z)_P$.

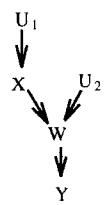
In the causal model of Fig. B.5, $(X \not\rightarrow Y \mid W)_P \ \& \ (X \not\rightarrow W \mid Y)_P \ \& \ \neg(X \not\rightarrow WY \mid \emptyset)_P$.



$V = \{X, W, Y, Z\}$ binary
$U = \{U_1, U_2\}$ binary

$x = u_1$
$y = u_2$
$w = Parity(x, y, z)$
$z = u_2$
$P(u_1) = P(u_2) = 0.5$

Fig. B.6. Counterexample to Property 2.5.1 in which each variable in U has a single child.



$V = \{X, W, Y\}$,
$x, y \in \{0, 1\}$, $w \in \{0, 1, 2, 3\}$
$U = \{U_1, U_2\}$, $u_1, u_2 \in \{0, 1\}$

$x = u_1$
$w = x + 2 * u_2$
$y = (w > 1)$
$P(u_1 = 1) = P(u_2 = 1) = 0.5$

Fig. B.7. Counterexample to Property 2.6.

Fixing W prevents X from altering the probability of Y, and fixing Y prevents X from altering the probability of W, but X can change the probability of W (and hence the probability of $W \& Y$) if there is no intervention on Y.

Up to this point, all of the counterexamples have relied on some exogenous variable from U having two different children in V. Obviously, this is not essential, since we could always create similar examples in which each exogenous variable has exactly one child. For example, in the model of Fig. B.5, we could replace U_2 with Z to get the model of Fig. B.6.

In this model, all of the exogenous variables U have exactly one child, yet Property 2.5.1 still does not hold. There is still an undirected cycle in the underlying causal graph, which is required for Property 2.5.1 to be false. Properties 2.2.1–2.6 are all true for all causal models whose causal graphs are trees. In addition, Properties 2.2.1–2.5.2 are true for all causal models whose causal graphs are polytrees. Property 2.6, as we will see now, is not always true, even when we restrict its causal graph to be a polytree.

2.6 $(X \not\rightarrow Y \mid Z)_P \Longrightarrow (a \not\rightarrow Y \mid Z)_P \vee (X \not\rightarrow a \mid Z)_P \; \forall a \notin X \cup Z \cup Y$.
In the causal model of Fig. B.7,

$$(X \not\rightarrow Y \mid \emptyset)_P \; \& \; \neg(W \not\rightarrow Y \mid \emptyset)_P \; \& \; \neg(X \not\rightarrow W \mid \emptyset)_P \; \& \; W \notin X \cup Z \cup Y.$$

X can only cause a minor change in W, while a large change in W is required to affect Y. Thus, X can affect W, and W can affect Y, but X has no effect on W. Even if we restrict all variables to be binary, transitivity will not hold. For this counterexample, W could be split into four binary variables $W_1, \ldots, W_4$, with $f_{W_1} = \neg(x \vee u_2)$, $f_{W_2} = x \,\&\, \neg u_2$, $f_{W_3} = \neg x \,\&\, u_2$, $f_{W_4} = x \,\&\, u_2$, and $f_y = w_3 \vee w_4$. Section 4.6 elaborates this counterexample.

References

[1] A. Balke and J. Pearl, Counterfactual probabilities: computation methods, bounds and applications, in: R. Lopez de Mantaras and D. Poole, eds., *Proceedings 10th Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann, San Francisco, CA, 1994) 46–54.

[2] A. Balke and J. Pearl, Counterfactuals and policy analysis in structural models, in: P. Besnard and S. Hanks, eds., *Proceedings 11th Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann, San Francisco, CA, 1995) 11–18.

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

[4] A.P. Dawid, Conditional independence in statistical theory, *J. Roy. Statist. Soc. Ser. A* 41 (1979) 1–31.

[5] E. Eells, *Probabilistic Causality* (Cambridge University Press, Cambridge, UK, 1991).

[6] R.E. Fikes and N.J. Nilsson, STRIPS: a new approach to the application of theorem proving to problem solving, *Artificial Intelligence* 3 (1972) 251–284.

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

[8] D. Freedman, As others see us: a case study in path analysis (with discussion), *J. Educational Statist.* 12 (1987) 101–223.

[9] D. Galles, Causal models: a formalism for modeling actions and counterfactuals, Ph.D. Thesis, University of California, Los Angeles, CA (1997).

[10] D. Geiger, T.S. Verma and J. Pearl, Identifying independence in Bayesian networks, in: *Networks*, Vol. 20 (John Wiley, Sussex, UK, 1990) 507–534.

[11] A.S. Goldberger, Models of substance [comment on N. Wermuth, "On block-recursive linear regression equations"], *Brazilian J. Probab. Statist.* 6 (1992) 1–56.

[12] I.J. Good, A causal calculus, *Philos. Sci.* 11 (1961) 305–318.

[13] J.Y. Halpern, Axiomatizing causal structures, Unpublished report, Cornell University, Ithaca, NY (1997).

[14] D. Heckerman and R. Shachter, A definition and graphical representation of causality, in: P. Besnard and S. Hanks, eds., *Proceedings 11th Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann, San Francisco, CA, 1995) 262–273.

[15] P.W. Holland, Statistics and causal inference (with discussion), *J. Amer. Statist. Assoc.* 81 (396) (1986) 945–970.

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

[17] D. Lewis, Causation, *J. Philos.* 70 (1973) 556–567.

[18] D. Lewis, *Counterfactuals* (Harvard University Press, Cambridge, MA, 1973).

[19] D. Lewis, Counterfactuals and comparative possibility, in: W.L. Harper, R. Stalnaker and G. Pearce, eds., *Ifs* (Reidel, Dordrecht, The Netherlands, 1981).

[20] C.F. Manski, Nonparametric bounds on treatment effects, *Amer. Economic Review, Papers and Proceedings* 80 (1990) 319–323.

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

[22] A. Paz and J. Pearl, Axiomatic characterization of directed graphs, Tech. Rept. R-234, Computer Science Department, University of California, Los Angeles, CA (1994).

[23] A. Paz, J. Pearl and S. Ur, A new characterization of graphs based on interception relations, *J. Graph Theory* 22 (2) (1996) 125–136.

[24] J. Pearl and A. Paz, Graphoids: a graph-based logic for reasoning about relevance relations, in: B. du Boulay and L. Steels, eds., *Advances in Artificial Intelligence—II* (North-Holland, Amsterdam, 1987) 357–363.

[25] J. Pearl and T. Verma, A theory of inferred causation, in: J.A. Allen, R. Fikes and E. Sandewall, eds., *Principles of Knowledge Representation and Reasoning: Proceedings 2nd International Conference* (Morgan Kaufmann, San Mateo, CA, 1991) 441–452; also in: D. Prawitz, B. Skyrms and D. Westertahl, eds., *Logic, Methodology and Philosophy of Science IX* (Elsevier, Amsterdam, 1994) 789–811.

[26] J. Pearl, *Probabilistic Reasoning in Intelligent Systems* (Morgan Kaufmann, San Mateo, CA, 1988).

[27] J. Pearl, Graphical models, causality and intervention, *Statist. Sci.* 8 (3) (1993) 266–273.

[28] J. Pearl, A probabilistic calculus of actions, in: R. Lopez de Mantaras and D. Poole, eds., *Proceedings 10th Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann, San Francisco, CA, 1994) 454–462.

[29] J. Pearl, Causal diagrams for empirical research (with discussion), *Biometrika* 82 (4) (1995) 669–709.

[30] J. Pearl, On the testability of causal models with latent and instrumental variables, in: D. Besnard and S. Hanks, eds., *Proceedings 11th Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann, San Francisco, CA, 1995) 435–443.

[31] J. Pearl, Causation, action and counterfactuals, in: Y. Shoham, ed., *Proceedings 6th Conference Theoretical Aspects of Reasoning about Knowledge (TARK 1996)* (Morgan Kaufmann, San Francisco, CA, 1996) 51–73.

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

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

[34] J. Robins, Addendum to "A new approach to causal inference in mortality studies with sustained exposure periods—application to control of the healthy worker survivor effect", *Comput. Math. Appl.* 14 (1987) 923–45.

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

[36] D.B. Rubin, Estimating causal effects of treatments in randomized and nonrandomized studies, *J. Educational Psychology* 66 (1974) 688–701.

[37] W. Salmon, *Scientific Explanation and the Causal Structure of the World* (Princeton University Press, Princeton, NJ, 1984).

[38] L.J. Savage, *The Foundations of Statistics*, Vol. 1 (John Wiley, New York, 1954).

[39] G. Shafer, *The Art of Causal Conjecture* (MIT Press, Cambridge, MA, 1996).

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

[41] P. Spirtes, C. Glymour and R. Schienes, *Causation, Prediction and Search* (Springer, New York, 1993).

[42] W. Spohn, Stochastic independence, causal independence and shieldability, *J. Philos. Logic* 9 (1980) 73–99.

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

[44] M. Studeny, Conditional independence relations have no complete characterization, in: *Information Theory, Statistical Decision Functions, Random: Trans. of the 11th Prague Conference, 1990* (Kluwer, Dordrecht, The Netherlands, 1992) 377–396.

[45] P. Suppes, *A Probabilistic Theory of Causation* (North-Holland, Amsterdam, 1970).

[46] N. Wermuth, On block-recursive linear regression equations, *Brazilian J. Probab. Statist.* 6 (1992) 1–56.