TECHNICAL REPORT
R-485
October 2018

Causal and Counterfactual Inference

Judea Pearl
University of California, Los Angeles
Computer Science Department
Los Angeles, CA, 90095-1596, USA
judea@cs.ucla.edu

October 2, 2018

Abstract

All accounts of rationality presuppose knowledge of how actions affect the state of the world and how the world would change had alternative actions been taken. The paper presents a framework called Structural Causal Model (SCM) which operationalizes this knowledge and explicates how it can be derived from both theories and data. In particular, we show how counterfactuals are computed and how they can be embedded in a calculus that solves critical problems in the empirical sciences.

1 Introduction - Actions, Physical, and Metaphysical

If the options available to an agent are specified in terms of their immediate consequences, as in "make him laugh," "paint the wall red," "raise taxes" or, in general, $do(X = x)$, then a rational agent is instructed to maximize the expected utility

$$EU(x) = \sum_y P_x(y)U(y) \tag{1}$$

over all options x. Here, $U(y)$ stands for the utility of outcome $Y = y$ and $P_x(y)$ – the focus of this paper - stands for the (subjective) probability that outcome $Y = y$ would prevail, had action $do(X = x)$ been performed so as to establish condition $X = x$.

It has long been recognized that Bayesian conditionalization, i.e., $P_x(y) = P(y|x)$, is inappropriate for serving in Eq. (1), for it leads to paradoxical results of several kinds (see (Skyrms, 1980; Pearl, 2000a, pp. 108–9)). For example, patients would avoid going to the doctor to reduce the probability that one is seriously ill; barometers would be manipulated to reduce the chance of storms; doctors would recommend a drug to male and female patients, but not to patients with undisclosed gender, and so on. Yet the question of what function should substitute for $P_x(y)$, despite decades of thoughtful debates (Jeffrey, 1965; Harper et al., 1981; Cartwright, 1983) seems to still baffle philosophers in the 21st century (Arlo-Costa, 2007; Weirich, 2008; Woodward, 2003).

Most studies of rationality have dealt with the utility function $U(y)$, its behavior under various shades of uncertainty, and the adequacy of the expectation operator in Eq. (1). Relatively little has been said about the probability $P_x(y)$ that governs outcomes when an action $do(X = x)$ is executed. Yet regardless of what criterion one adopts for rational behavior, it must incorporate knowledge of how our actions affect the world. We must therefore define the function $P_x(y)$ and explicate the process by which it is assessed or inferred. We must also ask what mental representation and thought processes would permit a rational agent to combine world knowledge with specific observations and compute $P_x(y)$.

Guided by ideas from structural econometrics (Haavelmo, 1943; Strotz and Wold, 1960; Spirtes et al., 1993), I have explored a conditioning operator called $do(x)$ (Pearl, 1995) that captures the intent of $P_x(y)$ by simulating an intervention in a causal model of interdependent variables (Pearl, 2009b).

The idea is simple. To model an action $do(X = x)$ one performs a "mini-surgery" on the causal model, that is, a minimal change necessary for establishing the antecedent $X = x$, while leaving the rest of the model intact. This calls for removing the mechanism (i.e., equation) that nominally assigns values to variable X, and replacing it with a new equation, $X = x$, that enforces the intent of the specified action. This mini-surgery (not unlike Lewis's "little miracle"), makes precise the idea of using a "minimal deviation from actuality" to define counterfactuals.

One important feature of this formulation is that the post-intervention probability, $P(y|do(x))$, can be derived from pre-interventional probabilities provided one possesses a diagrammatic representation of the processes that govern variables in the domain (Pearl, 2000a; Spirtes et al., 2001). Specifically the post-intervention probability reads:[1]

$$P(x, y, z|do(X = x^*)) = \begin{cases} P(x, y, z)/P(x|z) & \text{if } x = x^* \\ 0 & \text{if } x \neq x^* \end{cases} \tag{2}$$

Here z stands for any realization of the set Z of "past" variables, y is any realization of the set Y of "future" variables, and "past" and "future" refer to the occurrence of the action event $X = x^*$.[2]

This feature, to be further discussed in Section 2, is perhaps the key for the popularity of graphical methods in causal inference applications. It states that the effects of policies and interventions can be predicted without knowledge of the functional relationships (or mechanisms) among X, Y, and Z. The pre-interventional probability and a few qualitative features of the model (e.g., variable ordering) are sufficient for determining the pot-intervention probabilities as in Eq. (2).

[1]The relation between P_x and P takes a variety of equivalent forms, including the back-door formula, truncated factorization, adjustment for direct causes, or the inverse probability weighing shown in Eq. (2) (Pearl, 2000a, pp. 72–3). The latter form is the easiest to describe without appealing to graphical notation. But see Eq. (10), Section 3 for a more general formula.

[2]I will use "future" and "past" figuratively; "affected" and "unaffected" (by X) are more accurate technically (i.e., descendants and non-descendants of X, in graphical terminology). The derivation of Eq. (2) requires that processes be organized recursively (avoiding feedback loops); more intricate formulas apply to non-recursive models. See Pearl (2009b, pp. 72–3) or Spirtes et al. (2001) for a simple derivation of this and equivalent formulas.

The philosophical literature spawned a totally different perspective on the probability function $P_x(y)$ in Eq. (1). In a famous letter to David Lewis, Robert Stalnaker (1972) suggested to replace conditional probabilities with probabilities of conditionals, i.e., $P_x(y) = P(x > y)$, where $(x > y)$ stands for counterfactual conditional "Y would be y if X were x." Using a "closest worlds" semantics, Lewis (1973) defined $P(x > y)$ using a probability-revision operation called "imaging," in which probability mass "shifts" from worlds to worlds, governed by a measure of "similarity." Whereas Bayes conditioning $P(y|x)$ transfers the entire probability mass from worlds excluded by $X = x$ to all remaining worlds, in proportion to the latters' prior probabilities $P(\cdot)$, imaging works differently; each excluded world w transfers its mass individually to a select set of worlds $S_x(w)$ that are considered "closest" to w among those satisfying $X = x$. Joyce (1999) used the "\" symbol, as in $P(y\backslash x)$, to denote the probability resulting from such imaging process, and derived a formula for $P(y\backslash x)$ in terms of the selection function $S_x(w)$.

In Pearl (2000a, p. 73) I have shown that the transformation defined by the $do(x)$ operator, Eq. (2), can be interpreted as an imaging-type mass-transfer, if the following two provisions are met.

Provision 1 - the choice of "similarity" measure is not arbitrary; worlds with equal histories should be considered equally similar to any given world.

Provision 2 - the re-distribution of weight within each selection set $S_x(w)$ is not arbitrary either, equally-similar worlds should receive mass in proportion to their prior probabilities.

This tie-breaking rule is similar in spirit to the Bayesian policy, and permits us to generalize Eq. (2) to disjunctive actions, as in "exercise at least 30 minutes daily," or "paint the wall either green or purple (Pearl, 2017).

The theory that emerges from the *do*-operator (Eq. (2)) offers several conceptual and operational advantages over Lewis's closest-world semantics. First, it does not rest on a metaphysical notion of "similarity," which may be different from person to person and, thus, could not explain the uniformity with which people interpret causal utterances. Instead, causal relations are defined in terms of our scientific understanding of how variables interact with one another (to be explicated in Section 2). Second, it offers a plausible resolution of the "mental representation" puzzle: How do humans represent "possible worlds" in their minds and compute the closest one, when the number of possibilities is far beyond the capacity of the human brain? Finally, it results in practical algorithms for solving some of the most critical and difficult causal problems that have challenged data analysts and experimental researchers in the past century (see Pearl and Mackenzie (2018) for extensive historical account). I call this theory Structural Causal Model (SCM).

In the rest of the paper we will focus on the properties of SCM, and explicate how it can be used to define counterfactuals (Section 2), to control confounding and predict the effect of interventions and policies (Section 3), to define and estimate direct and indirect effects (Section 4) and, finally, to ensure generalizability of empirical results across diverse environments (Section 5).

2 Counterfactuals and SCM

At the center of the structural theory of causation lies a "structural model," M, consisting of two sets of variables, U and V, and a set F of functions that determine or simulate how values are assigned to each variable $V_i \in V$. Thus for example, the equation

$$v_i = f_i(v, u)$$

describes a physical process by which variable V_i is *assigned* the value $v_i = f_i(v, u)$ in response to the current values, v and u, of all variables in V and U. Formally, the triplet $< U, V, F >$ defines a SCM, and the diagram that captures the relationships among the variables is called the *causal graph* G (of M). The variables in U are considered "exogenous," namely, background conditions for which no explanatory mechanism is encoded in model M. Every instantiation $U = u$ of the exogenous variables uniquely determines the values of all variables in V and, hence, if we assign a probability $P(u)$ to U, it defines a probability function $P(v)$ on V. The vector $U = u$ can also be interpreted as an experimental "unit" which can stand for an individual subject, agricultural lot or time of day, since it describes all factors needed to make V a deterministic function of U.

The basic counterfactual entity in structural models is the sentence: "Y would be y had X been x in unit (or situation) $U = u$," denoted $Y_x(u) = y$. Letting M_x stand for a modified version of M, with the equation(s) of set X replaced by $X = x$, the formal definition of the counterfactual $Y_x(u)$ reads

$$Y_x(u) \triangleq Y_{M_x}(u). \tag{3}$$

In words, the counterfactual $Y_x(u)$ in model M is defined as the solution for Y in the "modified" submodel M_x. Galles and Pearl (1998) and Halpern (1998) have given a complete axiomatization of structural counterfactuals, embracing both recursive and non-recursive models (see also (Pearl, 2009b, Chapter 7)).[3]

Since the distribution $P(u)$ induces a well defined probability on the counterfactual event $Y_x = y$, it also defines a joint distribution on all Boolean combinations of such events, for instance "$Y_x = y$ AND $Z_{x'} = z$," which may appear contradictory, if $x \neq x'$. For example, to answer retrospective questions, such as whether Y would be y_1 if X were x_1, given that in fact Y is y_0 and X is x_0, we need to compute the conditional probability $P(Y_{x_1} = y_1 | Y = y_0, X = x_0)$ which is well defined once we know the forms of the structural equations and the distribution of the exogenous variables in the model.

In general, the probability of the counterfactual sentence $P(Y_x = y|e)$, where e is any propositional evidence, can be computed by the three-step process ((Pearl, 2009b, p. 207)):

Step 1 (***abduction***): Update the probability $P(u)$ to obtain $P(u|e)$.

Step 2 (***action***): Replace the equations determining the variables in set X by $X = x$.

Step 3 (***prediction***): Use the modified model to compute the probability of $Y = y$.

[3]The structural definition of counterfactual given in Eq. (3) was first introduced in Balke and Pearl (1995).

In temporal metaphors, Step 1 explains the past (U) in light of the current evidence e; Step 2 bends the course of history (minimally) to comply with the hypothetical antecedent $X = x$; finally, Step 3 predicts the future (Y) based on our new understanding of the past and our newly established condition, $X = x$.

2.1 Example: computing counterfactuals in linear SCM

We illustrate the working of this three-step algorithm using a linear structural equation model, depicted by the graph of Fig. 1.

To motivate the analysis, let X stands for the level of assistance (or "treatment") given to a student, Z stands for the amount of time the student spends studying, and Y, the outcome, stands for the student's performance on an exam. The algebraic version of this model takes the form of the following equations:

$$x = \epsilon_1$$
$$z = \beta x + \epsilon_2$$
$$y = \alpha x + \gamma z + \epsilon_3$$

The coefficient alpha, beta and gamma are called "structural coefficients," to be distinguished from regression coefficients, and represent direct causal effects of the corresponding variables. Under appropriate assumptions, say that the error terms ϵ_1, ϵ_2 and ϵ_3 are mutual independent, the structural coefficients can be estimated from data. Our task however is not to estimate causal effects but to answer counterfactual questions taking the model as given.

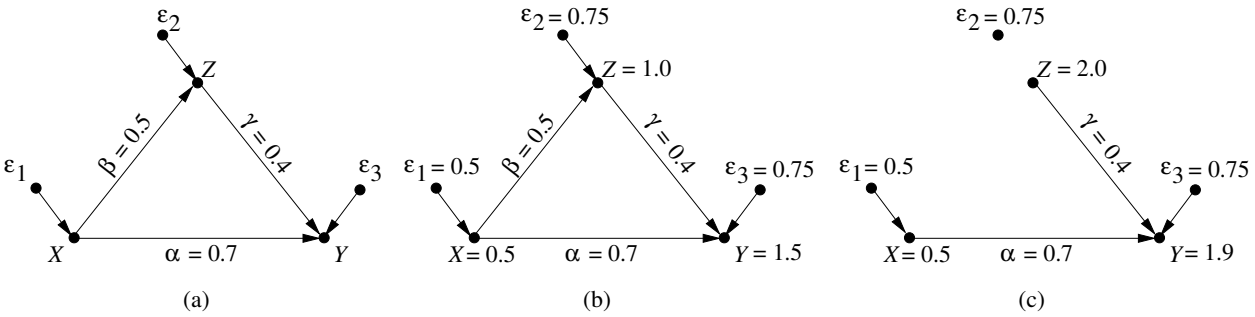

Figure 1: Structural models used for answering a counterfactual question about an individual $u = (\epsilon_1, \epsilon_2, \epsilon_3)$. (a) the generic model, (b) the u-specific model. (c) the modified model necessary to accommodate the antecedent $Z = 2$ of the counterfactual question Q_1.

Let us consider a student named Joe, for whom we measure $X = 0.5, Z = 1, Y = 1.5$, and about whom we ask a counterfactual question:

Q_1**:** What would Joe's score be had he doubled his study time?

Using our subscript notation, this question amounts to evaluating $Y_{Z=2}(u)$, with u standing for the distinctive characteristics of Joe, namely, $u = (\epsilon_1, \epsilon_2, \epsilon_3)$, as inferred from the observed data $\{X = 0.5, Z = 1, Y = 1.5\}$.

Following the algorithm above, the answer to this question is obtained in three steps.

1. Use the data to compute the exogenous factors $\epsilon_1, \epsilon_2, \epsilon_3$. (These are the invariant characteristics of unit u, and do not change by interventions or counterfactual hypothesizing.) In our model, we get (Figure 1(b)):

$$\begin{aligned}\epsilon_1 &= 0.5\\ \epsilon_2 &= 1 - 0.5 \times 0.5 = 0.75,\\ \epsilon_3 &= 1.5 - 0.5 \times 0.7 - 1 \times 0.4 = 0.75\end{aligned}$$

2. Modify the model, to form $M_{Z=2}$, in which Z is set to 2 and all arrows to Z are removed (Fig. 1(c)).

3. Compute the value of Y in the mutilated model formed in step 2, giving:

$$Y_{Z=2} = 0.5 \times 0.7 + 2.0 \times 0.4 + 0.75 = 1.90.$$

We can thus conclude that Joe's score would have been 1.90, instead of 1.5, had he doubled his study time. This example illustrates the need to modify the original model (Fig. 1(a)), in which the combination $(X = 1, \epsilon_2 = 0.75, Z = 2.0)$ constitutes a contradiction.

2.2 The two principles of causal inference

Before describing specific applications of the structural theory, it will be useful to summarize its implications in the form of two "principles," from which all other results follow.

Principle 1: "The law of structural counterfactuals."

Principle 2: "The law of structural independence."

The first principle is described in Eq. (3) and instructs us how to compute counterfactuals and probabilities of counterfactuals from a structural model. This, together with principle 2 will allow us (Section 3) to determine what assumptions one must make about reality in order to infer probabilities of counterfactuals from either experimental or passive observations.

Principle 2, defines how structural features of the model entail dependencies in the data. Remarkably, regardless of the functional form of the equations in the model and regardless of the distribution of the exogenous variables U, if the latters' are mutually independent and the model is recursive, the distribution $P(v)$ of the endogenous variables must obey certain conditional independence relations, stated roughly as follows: whenever sets X and Y are "separated" by a set Z in the graph, X is independent of Y given Z in the probability. This "separation" condition, called d-separation (Pearl, 2000a, pp. 16–18) constitutes the link between the causal assumptions encoded in the causal graph (in the form of missing arrows) and the observed data. It is defined formally as follows:

Definition 1 *(d-separation)*
A set S of nodes is said to block a path p if either

1. *p contains at least one arrow-emitting node that is in S, or*

2. *p contains at least one collision node that is outside S and has no descendant in S.*

If S blocks all *paths from set X to set Y, it is said to "d-separate X and Y," and then, variables X and Y are independent given S, written* $X \perp\!\!\!\perp Y|S$.[4]

D-separation implies conditional independencies for every distribution $P(v)$ that is compatible with the causal assumptions embedded in the diagram. To illustrate, the diagram in Fig. 2(a) implies $Z_1 \perp\!\!\!\perp Y|(X, Z_3, W_2)$, because the conditioning set $S = \{X, Z_3, W_2\}$ blocks all paths between Z_1 and Y. The set $S = \{X, Z_3, W_3\}$ however leaves the path (Z_1, Z_3, Z_2, W_2, Y) unblocked (by virtue of the collider at Z_3) and, so, the independence $Z_1 \perp\!\!\!\perp Y|(X, Z_3, W_3)$ is not implied by the diagram.

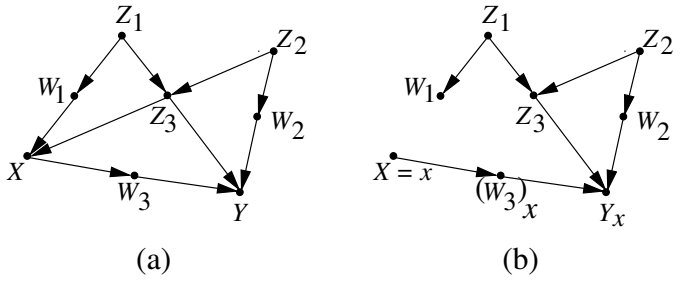

Figure 2: (a) Illustrating the intervention. (b) Submodel M_x, showing the intervention.

3 Intervention, identification, and causal calculus

A central question in causal analysis is that of predicting the results of interventions, such as those resulting from treatments or social programs, which we denote by the symbol $do(x)$ and define using the counterfactual Y_x as[5]

$$P(y|do(x)) \stackrel{\Delta}{=} P(Y_x = y) \tag{4}$$

Figure 3(b) illustrates the submodel M_x created by the atomic intervention $do(x)$; it sets the value of X to x and thus removes the influence of W_1 and Z_3 on X. We similarly define the result of *conditional interventions* by

$$P(y|do(x), z) \stackrel{\Delta}{=} P(y, z|do(x))/P(z|do(x) = P(Y_x = y|Z_x = z) \tag{5}$$

$P(y|do(x), z)$ captures the z-specific effect of X on Y, that is, the effect of setting X to x among those units only for which $Z = z$.

[4]By a "path" we mean a consecutive edges in the graph regardless of direction. See (Pearl, 2009b, p. 335) for a gentle introduction to d-separation and its proof. In linear models, the independencies implied by d-separation are valid for non-recursive models as well.

[5]An alternative definition of $do(x)$, invoking population averages only, is given in (Pearl, 2009b, p. 24).

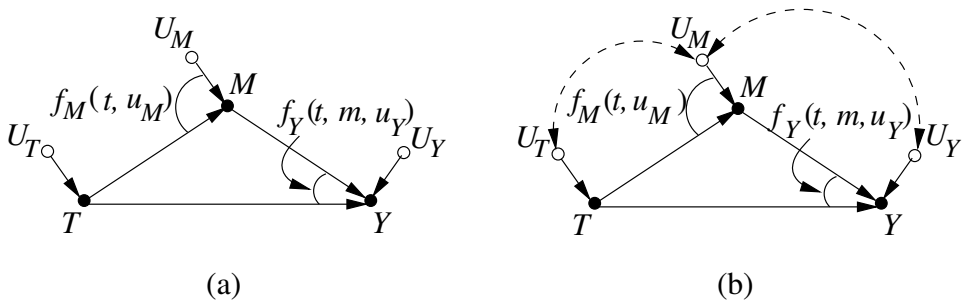

Figure 3: (a) The basic nonparametric mediation model, with no confounding. (b) A confounded mediation model in which dependence exists between $U_{_M}$ and $(U_{_T}, U_{_Y})$.

A second important question concerns *identification* in partially specified models: Given a set A of qualitative causal assumptions, as embodied in the structure of the causal graph, can the controlled (post-intervention) distribution, $P(y|do(x))$, be estimated from the available data which are governed by the pre-intervention distribution $P(z, x, y)$? In linear parametric settings, the question of identification reduces to asking whether some model parameter, β, has a unique solution in terms of the parameters of P (say the population covariance matrix). In the nonparametric formulation, the notion of "has a unique solution" does not directly apply since quantities such as $Q = P(y|do(x))$ have no parametric signature and are defined procedurally by a symbolic operation on the causal model M, as in Fig. 2(b). The following definition captures the requirement that Q be estimable from the data:

Definition 2 *(Identifiability) (Pearl, 2000a, p. 77)*
A causal query Q is identifiable from data compatible with a causal graph G, if for any two (fully specified) models M_1 and M_2 that satisfy the assumptions in G, we have

$$P_1(v) = P_2(v) \Rightarrow Q(M_1) = Q(M_2) \tag{6}$$

In words, equality in the probabilities $P_1(v)$ and $P_2(v)$ induced by models M_1 and M_2, respectively, entails equality in the answers that these two models give to query Q. When this happens, Q depends on P only and should therefore be expressible in terms of the parameters of P.

When a query Q is given in the form of a *do*-expression, for example $Q = P(y|do(x), z)$, its identifiability can be decided systematically using an algebraic procedure known as the *do*-calculus Pearl (1995). It consists of three inference rules that permit us to equate interventional and observational distributions whenever certain d-separation conditions hold in the causal diagram G.

3.1 The rules of *do*-calculus

Let X, Y, Z, and W be arbitrary disjoint sets of nodes in a causal DAG G. We denote by $G_{\overline{X}}$ the graph obtained by deleting from G all arrows pointing to nodes in X. Likewise, we denote by $G_{\underline{X}}$ the graph obtained by deleting from G all arrows emerging from nodes in X. To represent the deletion of both incoming and outgoing arrows, we use the notation $G_{\overline{X}\underline{Z}}$.

The following three rules are valid for every interventional distribution compatible with G.

Rule 1 (Insertion/deletion of observations):

$$P(y|do(x), z, w) = P(y|do(x), w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}} \tag{7}$$

Rule 2 (Action/observation exchange):

$$P(y|do(x), do(z), w) = P(y|do(x), z, w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}} \tag{8}$$

Rule 3 (Insertion/deletion of actions):

$$P(y|do(x), do(z), w) = P(y|do(x), w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\overline{Z(W)}}}, \tag{9}$$

where $Z(W)$ is the set of Z-nodes that are not ancestors of any W-node in $G_{\overline{X}}$.

To establish identifiability of a causal query Q, one needs to repeatedly apply the rules of *do*-calculus to Q, until an expression is obtained which no longer contains a *do*-operator[6]; this renders it estimable from nonexperimental data. The *do*-calculus was proven to be complete for queries in the form $Q = P(y|do(x), z)$ Huang and Valtorta (2006); Shpitser and Pearl (2006), which means that if Q cannot be reduced to probabilities of observables by repeated application of these three rules, such a reduction does not exist, i.e., the query is not estimable from observational studies without strengthening the assumptions.

3.2 Covariate selection: the back-door criterion

Consider an observational study where we wish to find the effect of treatment (X) on outcome (Y), and assume that the factors deemed relevant to the problem are structured as in Fig. 2(a); some are affecting the outcome, some are affecting the treatment, and some are affecting both treatment and response. Some of these factors may be unmeasurable, such as genetic trait or lifestyle, while others are measurable, such as gender, age, and salary level. Our problem is to select a subset of these factors for measurement and adjustment such that if we compare treated vs. untreated subjects having the same values of the selected factors, we get the correct treatment effect in that subpopulation of subjects. Such a set of factors is called a "sufficient set," "admissible set" or a set "appropriate for adjustment" (see Greenland et al. (1999); Pearl (2000b, 2009a)). The following criterion, named "back-door" Pearl (1993), provides a graphical method of selecting such a set of factors for adjustment.

Definition 3 *(admissible sets—the back-door criterion)*
A set S is admissible (or "sufficient") for estimating the causal effect of X on Y if two conditions hold:

1. *No element of S is a descendant of X.*
2. *The elements of S "block" all "back-door" paths from X to Y—namely, all paths that end with an arrow pointing to X.*

[6]Such derivations are illustrated in graphical details in (Pearl, 2009b, p. 87).

Based on this criterion we see, for example that, in Fig. 2, the sets $\{Z_1, Z_2, Z_3\}$, $\{Z_1, Z_3\}$, $\{W_1, Z_3\}$, and $\{W_2, Z_3\}$ are each sufficient for adjustment, because each blocks all back-door paths between X and Y. The set $\{Z_3\}$, however, is not sufficient for adjustment because it does not block the path $X \leftarrow W_1 \leftarrow Z_1 \rightarrow Z_3 \leftarrow Z_2 \rightarrow W_2 \rightarrow Y$.

The intuition behind the back-door criterion is as follows. The back-door paths in the diagram carry spurious associations from X to Y, while the paths directed along the arrows from X to Y carry causative associations. Blocking the former paths (by conditioning on S) ensures that the measured association between X and Y is purely causal, namely, it correctly represents the target quantity: the causal effect of X on Y. Conditions for relaxing restriction 1 are given in (Pearl, 2009b, p. 338)Shpitser et al. (2010); Pearl and Paz (2014)[7]).

The implication of finding a sufficient set, S, is that stratifying on S is guaranteed to remove all confounding bias relative to the causal effect of X on Y. In other words, it renders the causal effect of X on Y identifiable, via the *adjustment formula*[8]

$$P(Y = y|do(X = x)) = \sum_s P(Y = y|X = x, S = s)P(S = s) \tag{10}$$

Since all factors on the right-hand side of the equation are estimable (e.g., by regression) from pre-interventional data, the causal effect can likewise be estimated from such data without bias. Note that Eq. (2) is a special case of Eq. (10), where S chosen to include all variables preceding X in the causal order. Moreover, the back-door criterion implies the independence $X \perp\!\!\!\perp Y_x|S$, also known as "conditional ignorability" Rosenbaum and Rubin (1983) and, provides therefore the scientific basis for most inferences in the potential outcome framework.

The back-door criterion allows us to write Eq. (10) by inspection, after selecting a sufficient set, S, from the diagram. The selection criterion can be applied systematically to diagrams of any size and shape, thus freeing analysts from judging whether "X is conditionally ignorable given S," a formidable mental task required in the potential-response framework. The criterion also enables the analyst to search for an optimal set of covariates—namely, a set, S, that minimizes measurement cost or sampling variability Tian et al. (1998).

Theorem 1 *(Identification of Interventional Expressions) Given a causal graph G containing both measured and unmeasured variables, the consistent estimability of any expression of the form*

$$Q = P(y_1, y_2, \ldots, y_m|do(x_1, x_2, \ldots, x_n), z_1, z_2, \ldots, z_k)$$

can be decided in polynomial time. If Q is estimable, then its estimand can be derived in polynomial time. Furthermore, the algorithm is complete.

The results stated in Theorem 1 were developed in several stages over the past 20 years Pearl (1993, 1995); Tian and Pearl (2002); Shpitser and Pearl (2006). Bareinboim and Pearl

[7]In particular, the criterion devised by Pearl and Paz (2014) simply adds to Condition 2 of Definition 3 the requirement that X and its non-descendants (in Z) separate its descendants (in Z) from Y.

[8]Summations should be replaced by integration when applied to continuous variables, as in Imai et al. (2010).

(2012a) extended the identifiability of Q to combinations of observational and experimental studies.

4 Mediation analysis

Mediation analysis aims to uncover causal pathways along which changes are transmitted from causes to effects. Interest in mediation analysis stems from both scientific and practical considerations. Scientifically, mediation tells us "how nature works," and practically, it enables us to predict behavior under a rich variety of conditions and policy interventions. For example, in coping with the age-old problem of gender discrimination (Bickel et al., 1975; Goldberger, 1984) a policy maker may be interested in assessing the extent to which gender disparity inf hiring can be reduced by making hiring decisions gender-blind, compared with eliminating gender inequality in education or job qualifications. The former concerns the "direct effect" of gender on hiring while the latter concerns the "indirect effect" or the effect *mediated* via job qualification. .

The nonparametric structural model for a typical mediation problem takes the form:

$$t = f_{_T}(u_{_T}) \qquad m = f_{_M}(t, u_{_M}) \qquad y = f_{_Y}(t, m, u_{_Y}) \tag{11}$$

where T (treatment), M (mediator), and Y (outcome) are discrete or continuous random variables, $f_{_T}, f_{_M}$, and $f_{_Y}$ are arbitrary functions, and $U_{_T}, U_{_M}, U_{_Y}$ represent, respectively, omitted factors that influence T, M, and Y. In Fig. 3(a) the omitted factors are assumed to be arbitrarily distributed but mutually independent, written $U_{_T} \perp\!\!\!\perp U_{_M} \perp\!\!\!\perp U_{_Y}$. In Fig. 3(b) the dashed arcs connecting $U_{_T}$ and $U_{_M}$ (as well as $U_{_M}$ and $U_{_T}$) encode the understanding that the factors in question may be dependent.

4.1 Natural direct and indirect effects

Using the structural model of Eq. (11), four types of effects can be defined for the transition from $T = 0$ to $T = 1$[9]:

(a) Total effect –

$$\begin{aligned} TE &= E\{f_{_Y}[1, f_{_M}(1, u_{_M}), u_{_Y}] - f_{_Y}[0, f_{_M}(0, u_{_M}), u_{_Y}]\} \\ &= E[Y_1 - Y_0] \\ &= E[Y|do(T = 1)] - E[Y|do(T = 0)] \end{aligned} \tag{12}$$

TE measures the expected increase in Y as the treatment changes from $T = 0$ to $T = 1$, while the mediator is allowed to track the change in T as dictated by the function $f_{_M}$.

(b) Controlled direct effect –

$$\begin{aligned} CDE(m) &= E\{f_{_Y}[1, M = m, u_{_Y}] - f_{_Y}[0, M = m, u_{_Y}]\} \\ &= E[Y_{1,m} - Y_{0,m}] \\ &= E[Y|do(T = 1, M = m)] - E[Y|do(T = 0, M = m)] \end{aligned} \tag{13}$$

[9]Generalizations to arbitrary reference point, say from $T = t$ to $T = t'$, are straightforward. These definitions apply at the population levels; the unit-level effects are given by the expressions under the expectation. All expectations are taken over the factors $U_{_M}$ and $U_{_Y}$. Note that in this section we use parenthetical notation for counterfactuals, replacing the subscript notation used in Sections 2 and 3.

CDE measures the expected increase in Y as the treatment changes from $T = 0$ to $T = 1$, while the mediator is set to a specified level $M = m$ uniformly over the entire population.

(c) Natural direct effect[10] –

$$\begin{aligned} NDE &= E\{f_{_Y}[1, f_{_M}(0, u_{_M}), u_{_T}] - f_{_Y}[0, f_{_M}(0, u_{_M}), u_{_T}]\} \\ &= E[Y_{1,M_0} - Y_{0,M_0}] \end{aligned} \tag{14}$$

NDE measures the expected increase in Y as the treatment changes from $T = 0$ to $T = 1$, while the mediator is set to whatever value it *would have attained* (for each individual) prior to the change, i.e., under $T = 0$.

(d) Natural indirect effect –

$$\begin{aligned} NIE &= E\{f_{_Y}[0, f_{_M}(1, u_{_M}), u_{_Y}] - f_{_Y}[0, f_{_M}(0, u_{_M}), u_{_Y}]\} \\ &= E[Y_{0,M_1} - Y_{0,M_0}] \end{aligned} \tag{15}$$

NIE measures the expected increase in Y when the treatment is held constant, at $T = 0$, and M changes to whatever value it would have attained (for each individual) under $T = 1$. It captures, therefore, the portion of the effect which can be explained by mediation alone, while disabling the capacity of Y responds to X.

We note that, in general, the total effect can be decomposed as

$$TE = NDE - NIE_r \tag{16}$$

where NIE_r stands for the natural indirect effect under the reverse transition, from $T = 1$ to $T = 0$. This implies that NIE is identifiable whenever NDE and TE are identifiable. In linear systems, where reversal of transitions amounts to negating the signs of their effects, we have the standard additive formula, $TE = NDE + NIE$.

We further note that TE and $CDE(m)$ are *do*-expressions and can, therefore, be estimated from experimental data. Not so NDE and NIE; both are counterfactual expressions that cannot be reduced to *do*-expression. The reason is simple; there is no way to disable the direct effect by intervening on any variable in the model. The counterfactual language permits us to circumvent this difficulty by (figuratively) changing T to affect M while feeding Y the prior value of T.

Since Theorem 1 assures us that the identifiability of any *do*-expression can be determined by an effective algorithm, TE and $CDE(m)$ can be identified by those algorithms. NDE and NIE however require special analysis, given in the next section.

4.2 Sufficient conditions for identifying natural effects

The following is a set of assumptions or conditions, marked A-1 to A-4, that are sufficient for identifying both direct and indirect natural effects. Each condition is communicated using the causal diagram.

[10] Natural direct and indirect effects were conceptualized in Robins and Greenland (1992) and were formalized using Eqs. (14) and (15) in Pearl (2001).

4.2.1 Graphical conditions for identification

There exists a set W of measured covariates such that:

A-1 No member of W is a descendant of T.

A-2 W blocks all back-door paths from M to Y (not traversing $X \to M$ and $X \to Y$).

A-3 The W-specific effect of T on M is identifiable (using Summary Result 1 and possibly using experiments or auxiliary variables).

A-4 The W-specific joint effect of $\{T, M\}$ on Y is identifiable (using Theorem 1 and possibly using experiments or auxiliary variables).

Theorem 2 *(Identification of natural effects)*
When conditions A-1 and A-2 hold, the natural direct effect is experimentally identifiable and is given by

$$\begin{aligned} NDE = \sum_m \sum_w [E(Y|do(T=1, M=m)), W=w) \\ - E(Y|do(T=0, M=m), W=w)] \\ P(M=m|do(T=0), W=w) P(W=w) \end{aligned} \tag{17}$$

The identifiability of the do-expressions in Eq. (17) is guaranteed by conditions A-3 and A-4 and can be determined by Theorem 1.

In the non-confounding case (Fig. 3(a)), NDE reduces to : the mediation formula*:*

$$\begin{aligned} NDE = \sum_m [E(Y \mid T=1, M=m) - E(Y \mid T=0, M=m)] \\ P(M=m \mid T=0). \end{aligned} \tag{18}$$

which came to be known as the mediation formula (Pearl, 2012).

Shpitser (2013) further provides complete algorithms for identifying natural direct and indirect effects and extends these results to path-specific effects with multiple treatments and multiple outcomes.

5 External Validity and Transportability

In applications requiring identification, the role of the *do*-calculus is to remove the *do*-operator from the query expression. We now discuss a totally different application, to decide if experimental findings from environment π can be transported to a new, potentially different environment, π^*, in which only passive observations can be performed. This problem, labeled "transportability" in Pearl and Bareinboim (2011) is at the heart of every scientific investigation since, invariably, experiments performed in one environment (or population) are intended to be used elsewhere, where conditions may differ.

To formalize problems of this sort, a graphical representation called "selection diagrams" was devised (Fig. 4) which encodes knowledge about differences and commonalities between populations. A selection diagram is a causal diagram annotated with new variables, called S-nodes, which point to the mechanisms where discrepancies between the two populations are suspected to take place. The task of deciding if transportability is feasible now reduces

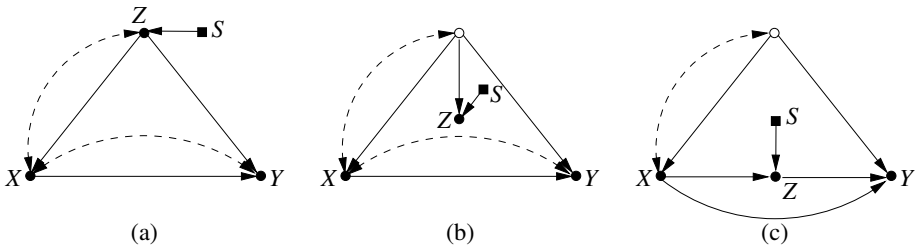

Figure 4: Selection diagrams depicting differences in populations. In (a), the two populations differ in age distributions. In (b), the populations differs in how reading skills (Z) depends on age (an unmeasured variable, represented by the hollow circle) and the age distributions are the same. In (c), the populations differ in how Z depends on X. Dashed arcs (e.g., $X \dashleftarrow\!\!\dashrightarrow Y$) represent the presence of latent variables affecting both X and Y.

to a syntactic problem of separating (using the *do*-calculus) the *do*-operator from a the S-variables in the query expression $P(y|do(x), z, s)$. In effect, this separation renders the disparities irrelevant to what we learn in the experimental setup.

Theorem 3 *Pearl and Bareinboim (2011)*
Let D be the selection diagram characterizing two populations, π and π^*, and S a set of selection variables in D. The relation $R = P^*(y|do(x), z)$ is transportable from π and π^* if and only if the expression $P(y|do(x), z, s)$ is reducible, using the rules of do-calculus, to an expression in which S appears only as a conditioning variable in do-free terms.

While Theorem 3 does not specify the sequence of rules leading to the needed reduction (if such exists), a complete and effective graphical procedure was devised by Bareinboim and Pearl (2012b), which also synthesizes a *transport formula* whenever possible. Each transport formula determines what information need to be extracted from the experimental and observational studies and how they ought to be combined to yield an unbiased estimate of the relation $R = P(y|do(x), s)$ in the target population π^*. For example, the transport formulas induced by the three models in Fig. 4 are given by:

(a) $P(y|do(x), s) = \sum_z P(y|do(x), z)P(z|s)$

(b) $P(y|do(x), s) = P(y|do(x))$

(c) $P(y|do(x), s) = \sum_z P(y|do(x), z)P(z|x, s)$

Each of these formulas satisfies Theorem 3, and each describes a different procedure of pooling information from π and π^*.

For example, (c) states that to estimate the causal effect of X on Y in the target population π^*, $P(y|do(x), z, s)$, we must estimate the z-specific effect $P(y|do(x), z)$ in the

source population π and average it over z, weighted by $P(z|x,s)$, i.e., the conditional probability $P(z|x)$ estimated at the target population π^*. The derivation of this formula follows by writing

$$P(y|do(x),s) = \sum_z P(y|do(x),z,s)P(z|do(x),s)$$

and noting that Rule 1 of *do*-calculus authorizes the removal of s from the first term (since $Y \perp\!\!\!\perp S|Z$ holds in $G_{\overline{X}}$) and Rule 2 authorizes the replacement of $do(x)$ with x in the second term (since the independence $Z \perp\!\!\!\perp X$ holds in $G_{\underline{X}}$).

A generalization of transportability theory to multi-environment has led to a method called "data fusion" (Bareinboim and Pearl, 2016) aimed at combining results from many experimental and observational studies, each conducted on a different population and under a different set of conditions, so as to synthesize an aggregate measure of effect size in yet another environment, different than the rest. This fusion problem has received enormous attention in the health and social sciences, where it is typically handled inadequately by a statistical method called "meta analysis" which "averages out" differences instead of rectifying them.

Using multiple "selection diagrams" to encode commonalities among studies, Bareinboim and Pearl (2013) "synthesized" an estimator that is guaranteed to provide unbiased estimate of the desired quantity based on information that each study share with the target environment. Remarkably, a consistent estimator may be constructed from multiple sources even in cases where it is not constructible from any one source in isolation.

Theorem 4 *Bareinboim and Pearl (2013)*

- *Nonparametric transportability of experimental findings from multiple environments can be determined in polynomial time, provided suspected differences are encoded in selection diagrams.*
- *When transportability is feasible, a transport formula can be derived in polynomial time which specifies what information needs to be extracted from each environment to synthesize a consistent estimate for the target environment.*
- *The algorithm is complete, i.e., when it fails, transportability is infeasible.*

Another problem that falls under the Data Fusion umbrella is that of "Selection Bias" (Bareinboim et al., 2014), which requires a generalization from a subpopulation selected for a study to the population at large, the target of the intended policy.

Selection bias is induced by preferential selection of units for data analysis, usually governed by unknown factors including treatment, outcome, and their consequences, and represents a major obstacle to valid causal and statistical inferences. It cannot be removed by randomized experiments and can rarely be detected in either experimental or observational studies.[11] For instance, in a typical study of the effect of training program on

[11]Remarkably, there are special situations in which selection bias can be detected even from observations, as in the form of a non-chordal undirected component (Zhang, 2008).

earnings, subjects achieving higher incomes tend to report their earnings more frequently than those who earn less. The data-gathering process in this case will reflect this distortion in the sample proportions and, since the sample is no longer a faithful representation of the population, biased estimates will be produced regardless of how many samples were collected. Our ability to eliminate such bias by analytical means thus provides a major opportunity to the empirical sciences.

Conclusions

One of the crowning achievements of modern work on causality has been to formalize counterfactual reasoning within a structural-based representation, the very representation researchers use to encode scientific knowledge. We showed that every structural equation model determines the truth value of every counterfactual sentence. Therefore, we can determine analytically if the probability of a counterfactual sentence is estimable from experimental or observational studies, or combination thereof.

This enables us to infer behavior of specific individuals, identified by a distinct set of characteristics, as well as average behavior of populations, identified by pre-intervention features or post-intervention response. Additionally, this formalization leads to a calculus of actions that resolves some of the most daunting problems in the empirical sciences. These include, among others, the control of confounding, the evaluation of interventional policies, the assessment of direct and indirect effect and the generalization of empirical results across heterogeneous environments.

Acknowledgement

This research was supported in parts by grants from International Business Machines Corporation (IBM) [#A1771928], National Science Foundation [#IIS-1527490 and #IIS-1704932], and Office of Naval Research [#N00014-17-S-B001].

References

Arlo-Costa, H. (2007). The logic of conditionals. In *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.). URL = ⟨http://plato.stanford.edu/entries/logic-conditionals/⟩.

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

Bareinboim, E. and Pearl, J. (2012a). Causal inference by surrogate experiments: z-identifiability. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence* (N. de Freitas and K. Murphy, eds.). AUAI Press, Corvallis, OR, 113–120.

Bareinboim, E. and Pearl, J. (2012b). Transportability of causal effects: Completeness results. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*. AAAI Press, 698–704.

Bareinboim, E. and Pearl, J. (2013). Meta-transportability of causal effects: A formal approach. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics* (C. M. Carvalho and P. Ravikumar, eds.), vol. 31 of *Proceedings of Machine Learning Research*. PMLR, Scottsdale, Arizona, USA, 135–143.

Bareinboim, E. and Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences* **113** 7345–7352.

Bareinboim, E., Tian, J. and Pearl, J. (2014). Recovering from selection bias in causal and statistical inference. In *Proceedings of the Twenty-eighth AAAI Conference on Artificial Intelligence* (C. E. Brodley and P. Stone, eds.). AAAI Press, Palo Alto, CA, 2410–2416. Best Paper Award, <http://ftp.cs.ucla.edu/pub/stat_ser/r425.pdf>.

Bickel, P., Hammel, E. and O'Connell, J. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science* **187** 398–404.

Cartwright, N. (1983). *How the Laws of Physics Lie*. Clarendon Press, Oxford.

Galles, D. and Pearl, J. (1998). An axiomatic characterization of causal counterfactuals. *Foundation of Science* **3** 151–182.

Goldberger, A. (1984). Reverse regression and salary discrimination. *The Journal of Human Resources* 293–318.

Greenland, S., Pearl, J. and Robins, J. (1999). Causal diagrams for epidemiologic research. *Epidemiology* **10** 37–48.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica* **11** 1–12. Reprinted in D.F. Hendry and M.S. Morgan (Eds.), *The Foundations of Econometric Analysis*, Cambridge University Press, 477–490, 1995.

Halpern, J. (1998). Axiomatizing causal reasoning. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence* (G. Cooper and S. Moral, eds.). Morgan Kaufmann, San Francisco, CA, 202–210. Also, *Journal of Artificial Intelligence Research* 12:3, 17–37, 2000.

Harper, W., Stalnaker, R. and Pearce, G. (1981). *Ifs*. D. Reidel, Dordrecht.

Huang, Y. and Valtorta, M. (2006). Pearl's calculus of intervention is complete. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence* (R. Dechter and T. Richardson, eds.). AUAI Press, Corvallis, OR, 217–224.

Imai, K., Keele, L. and Yamamoto, T. (2010). Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science* **25** 51–71.

Jeffrey, R. (1965). *The Logic of Decision*. McGraw-Hill, New York.

Joyce, J. (1999). *The Foundations of Causal Decision Theory*. Cambridge University Press, Cambridge, MA.

Lewis, D. (1973). Counterfactuals and comparative possibility. In W.L. Harper, R. Stalnaker, and G. Pearce (Eds.). *Ifs*, D. Reidel, Dordrecht, pages 57–85, 1981.

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

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika* **82** 669–710.

Pearl, J. (2000a). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York. 2nd edition, 2009.

Pearl, J. (2000b). Comment on A.P. Dawid's, Causal inference without counterfactuals. *Journal of the American Statistical Association* **95** 428–431.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, San Francisco, CA, 411–420.

Pearl, J. (2009a). Causal inference in statistics: An overview. *Statistics Surveys* **3** 96–146.

Pearl, J. (2009b). *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge University Press, New York.

Pearl, J. (2012). The causal mediation formula – a guide to the assessment of pathways and mechanisms. *Prevention Science* **13** 426–436.

Pearl, J. (2017). Physical and metaphysical counterfactuals: Evaluating disjunctive actions. *Journal of Causal Inference,* Causal, Casual, and Curious Section **5**. Published online: <https://doi.org/10.1515/jci-2017-0018>.

Pearl, J. and Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach. In *Proceedings of the Twenty-Fifth Conference on Artificial Intelligence (AAAI-11)* (W. Burgard and D. Roth, eds.). AAAI Press, Menlo Park, CA, 247–254. Available at: ⟨http://ftp.cs.ucla.edu/pub/stat_ser/r372a.pdf⟩.

Pearl, J. and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York.

Pearl, J. and Paz, A. (2014). Confounding equivalence in causal inference. *Journal of Causal Inference* **2** 75–93.

Robins, J. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3** 143–155.

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

Shpitser, I. (2013). Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive Science* **37** 1011–1035.

Shpitser, I. and Pearl, J. (2006). Identification of conditional interventional distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence* (R. Dechter and T. Richardson, eds.). AUAI Press, Corvallis, OR, 437–444.

Shpitser, I., VanderWeele, T. and Robins, J. (2010). On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*. AUAI, Corvallis, OR, 527–536.

Skyrms, B. (1980). *Causal Necessity*. Yale University Press, New Haven.

Spirtes, P., Glymour, C. and Scheines, R. (1993). *Causation, Prediction, and Search*. Springer-Verlag, New York.

Spirtes, P., Glymour, C. and Scheines, R. (2001). *Causation, Prediction, and Search*. 2nd ed. MIT Press, Cambridge, MA.

Stalnaker, R. (1972). Letter to David Lewis. In W.L. Harper, R. Stalnaker, and G. Pearce (Eds.), *Ifs*, D. Reidel, Dordrecht, pages 151–152, 1981.

Strotz, R. and Wold, H. (1960). Recursive versus nonrecursive systems: An attempt at synthesis. *Econometrica* **28** 417–427.

Tian, J., Paz, A. and Pearl, J. (1998). Finding minimal separating sets. Tech. Rep. R-254, University of California, Los Angeles, CA.

Tian, J. and Pearl, J. (2002). A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*. AAAI Press/The MIT Press, Menlo Park, CA, 567–573.

Weirich, P. (2008). Causal decision theory. In *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.). URL = ⟨http://plato.stanford.edu/archives/win2008/entries/decision-causal/⟩.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, Oxford.

Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence* **172** 1873–1986.