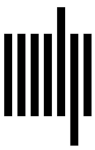
The Handbook of Rationality

Edited by: Markus Knauff, Wolfgang Spohn

Citation:

The Handbook of Rationality
Edited by: Markus Knauff, Wolfgang Spohn
DOI: 10.7551/mitpress/11252.001.0001
ISBN (electronic): 9780262366175
Publisher: The MIT Press
Published: 2021

Funding for the open access edition was provided by the MIT Libraries Open Monograph Fund.



The MIT Press

7.1 Causal and Counterfactual Inference

Judea Pearl

Summary

All accounts of rationality presuppose knowledge of how actions affect the state of the world and how the world would have changed had alternative actions been taken. The chapter presents a framework called the structural causal model (SCM), which operationalizes this knowledge and explicates how it can be derived from both theories and data. In particular, it shows how counterfactuals are computed and how they can be embedded in a calculus that solves critical problems in the empirical sciences.

1. 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, $do(X = x)$, where x is a value that variable X can take, then a rational agent is instructed to maximize the expected utility over all options x,

$$EU(x) = \sum_y P_x(y)U(y). \quad (1)$$

Here, $U(y)$ stands for the utility of outcome $Y = y$, and $P_x(y)$—the focus of this chapter—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, that is, $P_x(y) = P(y \mid x)$, is inappropriate for serving in equation (1), for it leads to paradoxical results of several kinds (see Pearl, 2000a, pp. 108–109; Skyrms, 1980). For example, patients would avoid going to the doctor to reduce the probability that they are 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 (Cartwright, 1983; Harper, Stalnaker, & Pearce, 1981; Jeffrey, 1965), seems to still baffle philosophers in the 21st century (Arló-Costa, 2007; Weirich, 2008/2016; Woodward, 2003). Modern discussions over evidential versus causal decision theory (chapter 8.2 by Peterson, this handbook) echo these debates.

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 equation (1). Relatively little has been said about the probability $P_x(y)$ that governs outcomes $Y = y$ when an action $do(X = x)$ is contemplated. 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, be it from empirical data or from world knowledge. We must also ask what mental representation and thought processes would permit a rational agent to combine world knowledge with empirical observations and compute $P_x(y)$.

Guided by ideas from structural econometrics (Haavelmo, 1943; Spirtes, Glymour, & Scheines, 1993; Strotz & Wold, 1960), 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 postintervention probability, $P(y \mid do(x))$, can be derived from preinterventional probabilities provided one possesses a diagrammatic representation of the processes

that govern variables in the domain (Pearl, 2000a; Spirtes, Glymour, & Scheines, 2001). Specifically, the post-intervention probability reads:[1]

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

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 graph-theoretical 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 preinterventional probability and a few qualitative features of the model (e.g., variable ordering) are sufficient for determining the postintervention probabilities as in equation (2).

The philosophical literature spawned a totally different perspective on the probability function $P_x(y)$ in equation (1). In a famous letter to David Lewis, Robert Stalnaker (1972/1981) suggested to replace conditional probabilities with probabilities of conditionals, that is, $P_x(y) = P(x > y)$, where "$x > y$" stands for the counterfactual conditional "Y would be y if X were x" (see chapter 6.1 by Starr, this handbook). 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 $P(y \mid x)$ transfers the entire probability mass from worlds excluded by $X = x$ to all remaining worlds, in proportion to the latter's prior probabilities $P(\cdot)$, imaging works differently: each excluded world w transfers its mass individually to a select set $S_x(w)$ of worlds that are considered "closest" to w among those satisfying $X = x$. Joyce (1999) used the "\"-symbol, as in "$P(y \setminus x)$," to denote the probability resulting from such an imaging process and derived a formula for $P(y \setminus x)$ in terms of the selection function $S_x(w)$.

In Pearl (2000a, p. 73), I have shown that the transformation defined by the *do*-operator, equation (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 redistribution 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 equation (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 (equation (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? Any credible theory of rationality must account for the astonishing ease with which humans comprehend, derive, and communicate counterfactual information. 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 & Mackenzie, 2018, for an extensive historical account). I call this theory the structural causal model (SCM).

In the rest of the chapter, 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 $\langle U, V, F \rangle$ defines an 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

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

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 nonrecursive 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, "$Y_x = y$ & $Z_{x'} = z$," 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 $P(Y_{x_1} = y_1 \mid Y = y_0, X = x_0)$, 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 $P(Y_x = y \mid e)$, where e is any propositional evidence, can be computed by the following three-step process (Pearl, 2009b, p. 207):

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

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 $Y = y$.

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 $X = x$, and 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 in figure 7.1.1.

To motivate the analysis, let X stand 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 the exam. The algebraic version of this model takes the form of the following equations:

$$x = \varepsilon_1,$$
$$z = \beta x + \varepsilon_2,$$
$$y = \alpha x + \gamma z + \varepsilon_3.$$

The coefficients α, β, and γ 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 mutually 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.

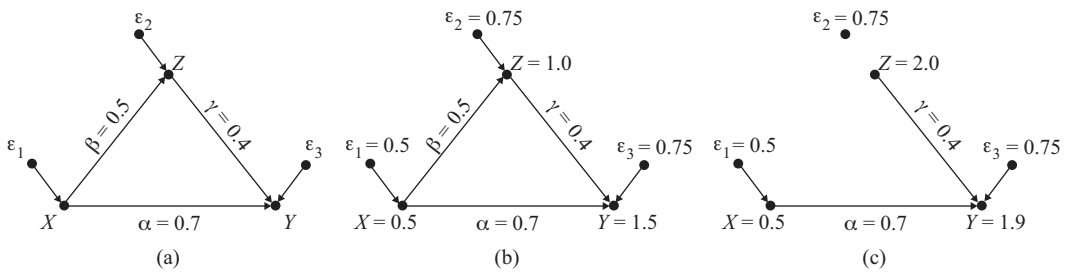


Figure 7.1.1

Structural models used for answering a counterfactual question about an individual $u = (\varepsilon_1, \varepsilon_2, \varepsilon_3)$. (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 have been 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, $u = (\varepsilon_1, \varepsilon_2, \varepsilon_3)$, 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 $\varepsilon_1, \varepsilon_2, \varepsilon_3$. (These are the invariant characteristics of unit u and do not change by interventions or counterfactual hypothesizing.) In our model, we get (see figure 7.1.1b):

   $\varepsilon_1 = 0.5$
   $\varepsilon_2 = 1 - 0.5 \times 0.5 = 0.75,$
   $\varepsilon_3 = 1.5 - 0.5 \times 0.7 - 1 \times 0.4 = 0.75.$

2. Modify the model, to form $M_{Z=2}$, in which Z is set to 2 and all arrows to Z are removed (figure 7.1.1c).
3. Compute the value of Y in the mutilated model formed in step 2, which gives

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

We can thus conclude that, had he doubled his study time, Joe's score would have been 1.90 instead of 1.5. This example illustrates the need to modify the original model (figure 7.1.1a), in which the combination $X = 1$, $\varepsilon_2 = 0.75$, $Z = 2.0$ 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 equation (3) and instructs us how to compute counterfactuals and their probabilities 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 latter 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 (Verma & Pearl, 1988). This "separation" condition, called "d-separation" (Geiger, Verma, & Pearl, 1990; 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 *collider* (i.e., a node obtaining head-to-head arrows) 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 $X \perp\!\!\!\perp Y \mid S$.[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 figure 7.1.2a implies $Z_1 \perp\!\!\!\perp Y \mid (X, Z_3, W_2)$, 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 $Z_1 \perp\!\!\!\perp Y \mid (X, Z_3, W_3)$ is not implied by the diagram.

3. Intervention, Identification, and Causal Calculus

To maximize the expectation defined by equation (1), a central problem for any rational agent is that of inferring the probability $P_x(y)$ from empirical data. In the context of social or medical policy making, this amounts

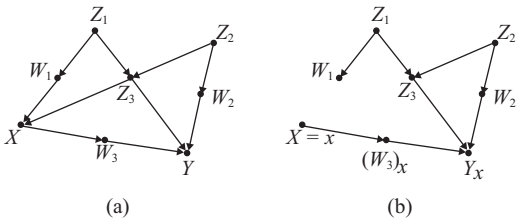


Figure 7.1.2
Illustrating the intervention $do(X = x)$. (a) The original model M. (b) The intervention submodel M_x and the counterfactual Y_x.

to estimating the interventional probability $P(y \mid do(x))$, which is defined, using the counterfactual Y_x, as[5]

$$P(y \mid do(x)) \triangleq P(Y_x = y). \tag{4}$$

Given a model M, the effect of an intervention $X = do(x)$ can be predicted from the submodel M_x as shown in figure 7.1.2. Figure 7.1.2b 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

$$\begin{aligned} P(y \mid do(x), z) &\triangleq P(y,z \mid do(x))/P(z \mid do(x)) \\ &= P(Y_x = y \mid Z_x = z). \end{aligned} \tag{5}$$

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

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 (postintervention) distribution, $P(y \mid do(x))$, be estimated from the available data that are governed by the preintervention 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 $Q = P(y \mid do(x))$ have no parametric signature and are defined procedurally by a symbolic operation on the causal model M, as in figure 7.1.2b. 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
>
> $$P_1(v) = P_2(v) \Longrightarrow Q(M_1) = Q(M_2). \tag{6}$$

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, $Q = P(y \mid do(x), z)$, 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 the Do-Calculus

Let X, Y, Z, and W be arbitrary disjoint sets of nodes in a causal directed acyclical graph (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 X-incoming and Z-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):

$$\begin{aligned} &P(y \mid do(x), z, w) = P(y \mid do(x), w) \\ &\quad \text{if } (Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}}}. \end{aligned} \tag{7}$$

Rule 2 (Action/observation exchange):

$$\begin{aligned} &P(y \mid do(x), do(z), w) = P(y \mid do(x), z, w) \\ &\quad \text{if } (Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}\underline{Z}}}. \end{aligned} \tag{8}$$

Rule 3 (Insertion/deletion of actions):

$$\begin{aligned} &P(y \mid do(x), do(z), w) = P(y \mid do(x), w) \\ &\quad \text{if } (Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}\overline{Z(W)}}}, \end{aligned} \tag{9}$$

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 the *do*-calculus to Q, until an expression is obtained that 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 $Q = P(y \mid do(x), z)$ (Huang & Valtorta, 2006; Shpitser & Pearl, 2006), which means that if Q cannot be reduced to probabilities of observables by repeated application of these three rules, then such a reduction does not exist, that is, the query is not estimable from observational studies without strengthening the assumptions.

3.2 Covariate Selection: The Backdoor Criterion

One of the most powerful results emerging from the *do*-calculus is a method of identifying a set of variables that, if measured, would permit us to predict the effect of action from passive observation. This set of variables coincides with the set Z of equation (2), which we called "past" in section 1 and will now receive a formal characterization in definition 3.

Consider an observational study in which we wish to find the effect of some treatment (X) on a certain outcome (Y), and assume that the factors deemed relevant to the problem are structured as in figure 7.1.2a; 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 versus 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" or a set "appropriate for adjustment" (see Greenland, Pearl, & Robins, 1999; Pearl, 2000b, 2009a). The following criterion, named "backdoor" (Pearl, 1993), provides a graphical method of selecting such a set of factors for adjustment:

Definition 3 (Admissible sets—the backdoor 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 "backdoor" paths from X to Y—namely, all paths that end with an arrow pointing to X.

Based on this criterion, we see, for example, that in figure 7.1.2, 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 backdoor 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 backdoor criterion is as follows. The backdoor 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), Pearl and Paz (2014), and Shpitser, VanderWeele, and Robins (2010).[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]

$$P(Y=y \mid do(X=x)) = \sum_s P(Y=y \mid X=x, S=s)P(S=s). \quad (10)$$

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

The backdoor criterion allows us to write equation (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, Paz, & Pearl, 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

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

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; Shpitser & Pearl, 2006; Tian & Pearl, 2002). Bareinboim and Pearl (2012a) extended the identifiability of Q to combinations of observational and experimental studies.

It is important to note at this point that the *do*-operator can be used not merely for fixing a variable at a predetermined value x but also for analyzing "soft interventions." For example, the effect of additive interventions, such as "administer 5 mg of insulin to a given patient," can be estimated using the *do*-calculus (Pearl, Glymour, & Jewell, 2016, p. 109). Likewise, the effects of stochastic interventions (e.g., "change the frequency with which this patient receives a drug") can be estimated by a method based on the *do*-operator (Pearl, 2009b, p. 113). The versatility of the *do*-operator is further discussed in Pearl (2009b, section 11.4).

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, Hammel, & O'Connell, 1975; Goldberger, 1984), a policy maker may be interested in assessing the extent to which gender disparity in 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 role that mediation analysis plays in rational decision making revolves around the richer set of options that emerges from understanding "how the world works." For example, the option of using mosquito nets was not considered by decision makers when malaria was believed to be caused by *mal'aria* ("bad air"). It was fairly rational in those days to use breathing masks in swampy areas. It is hardly rational today, given the overwhelming evidence about the mediating effect of the *Anopheles* mosquito. The logic of properly accounting for empirical data in one's belief system is an aspect of rational behavior that is gaining increased attention among researchers (Pearl, 2013).

The 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, and U_Y represent, respectively, omitted factors that influence T, M, and Y. In figure 7.1.3a, the omitted factors are assumed to be arbitrarily distributed but mutually independent. In figure 7.1.3b, 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 equation (11), four types of effects can be defined for the transition from $T = 0$ to $T = 1$:[9]

Total effect:

$$\begin{aligned} TE &= E\big[f_Y[1, f_M(1, u_M), u_Y] - f_Y[0, f_M(0, u_M), u_Y]\big] \\ &= E[Y_1 - Y_0] \\ &= E[Y \mid do(T = 1)] - E[Y \mid do(T = 0)]. \end{aligned} \tag{12}$$

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.

Controlled direct effect:

$$\begin{aligned} CDE(m) &= E\{f_Y[1, M = m, u_Y] - f_Y[0, M = m, u_Y]\} \\ &= E[Y_{1,m} - Y_{0,m}] \\ &= E[Y \mid do(T = 1, M = m)] - E[Y \mid do(T = 0, M = m)]. \end{aligned} \tag{13}$$

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.

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, that is, under $T = 0$.

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 that can be explained by mediation alone while disabling the capacity of Y to respond to X.

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

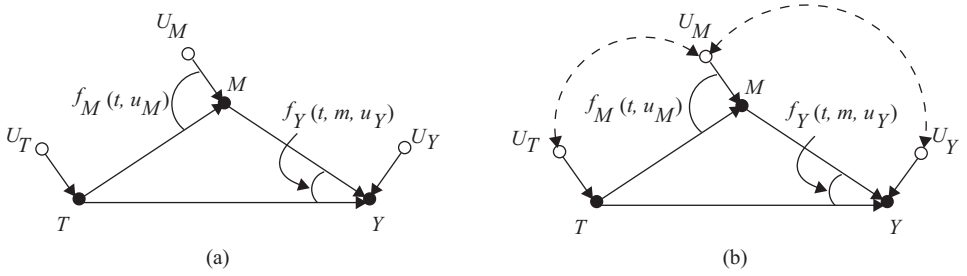


Figure 7.1.3

(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).

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

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*-expressions. 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 subsection.

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.

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 backdoor paths from M to Y (not traversing $X \rightarrow M$ and $X \rightarrow Y$).

A-3 The W-specific effect of T on M is identifiable (using theorem 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

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

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

In the nonconfounding case (figure 7.1.3a), *NDE* reduces to

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

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

To support the choice of optimal actions on the basis of nonexperimental data, 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 [a] "selection diagram" was devised (figure 7.1.4), 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 be located. The task of deciding if transportability is feasible now reduces to the syntactic problem of separating (using the *do*-calculus) the *do*-operator from the *S*-variables in the query expression $P(y \mid do(x), z, s)$. In effect, this separation renders the disparities irrelevant to what we learn in the experimental setup.

Theorem 3 (Pearl & 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 $R = P^*(y \mid do(x), z)$ is transportable from π and π^* if and only if the expression $P(y \mid do(x), z, s)$ is reducible, using the rules of the *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

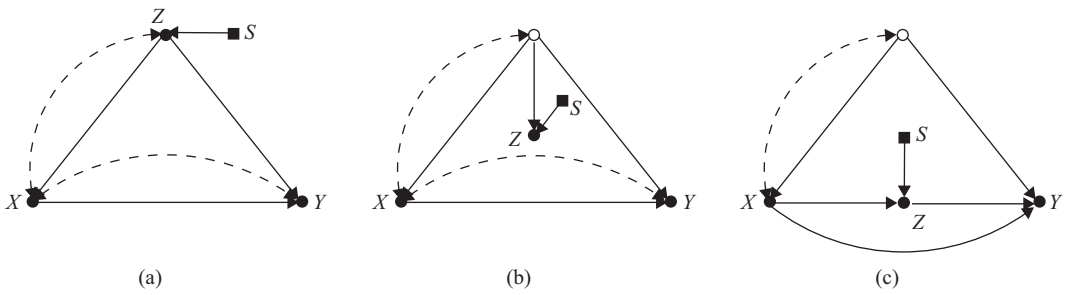


Figure 7.1.4
Selection diagrams depicting differences in populations. (a) The two populations differ in age distributions. (b) The populations differ in how reading skills (Z) depend on age (an unmeasured variable, represented by the hollow circle), and the age distributions are the same. (c) The populations differ in how Z depends on X. Dashed arcs (e.g., $X \leftarrow\!\!-\!\!-\!\!\rightarrow Y$) represent the presence of latent variables affecting both X and Y.

by Bareinboim and Pearl (2012b), which also synthesizes a *transport formula* whenever possible. Each transport formula determines what informations 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 $R = P(y \mid do(x), s)$ in the target population π^*. For example, the transport formulas induced by the three models in figure 7.1.4 are given by

$$P(y \mid do(x), s) = \sum_z P(y \mid do(x), z) P(z \mid s), \tag{a}$$

$$P(y \mid do(x), s) = P(y \mid do(x)), \tag{b}$$

$$P(y \mid do(x), s) = \sum_z P(y \mid do(x), z) P(z \mid x, s). \tag{c}$$

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^*, $P(y \mid do(x), z, s)$, we must estimate the z-specific effect $P(y \mid do(x), z)$ in the source population π and average it over z, weighted by $P(z \mid x, s)$, that is, the conditional probability $P(z \mid x)$ estimated at the target population π^*. The derivation of this formula follows by writing

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

and noting that rule 1 of the *do*-calculus authorizes the removal of s from the first term (since $Y \perp\!\!\!\perp S \mid Z$ 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 multiple environments has led to a method called "data fusion" (Bareinboim & 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 from 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 an unbiased estimate of the desired quantity based on information that each study shares 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 & 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 that specifies what information needs to be extracted from each environment to synthesize a consistent estimate for the target environment.
- The algorithm is complete, that is, when it fails, transportability is infeasible.

Another problem that falls under the data fusion umbrella is that of "selection bias" (Bareinboim, Tian, & Pearl, 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 observation, usually governed by unknown factors, thus rendering the data no longer representative of the environment (or population) of interest. Selection

bias 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 programs on 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.

6. Conclusions

Rational decisions demand rational assessments of the likely consequences of one's actions. This chapter offers a formal and normative account of how such assessments should be shaped by empirical observations and by prior world knowledge. The account is based on modern research in causal inference, which extends beyond probability and statistics and is becoming, in my opinion, an integral part of the theory of rationality.

One of the crowning achievements of modern work on causality has been to formalize counterfactual reasoning within a structure-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 a combination thereof.

This enables us to infer the behavior of specific individuals, identified by a distinct set of characteristics, as well as the average behavior of populations, identified by preintervention features or postintervention 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 effects, and the generalization of empirical results across heterogeneous environments. The same calculus can be leveraged to generate rational explanations for action recommended or actions taken in the past.

Acknowledgments

I am grateful to the coeditor, Wolfgang Spohn, for inviting me to participate in this handbook, for offering helpful comments on the first version of this chapter, and for pointing me to his earlier papers (Spohn, 1978, 1983), in which several key ideas on causal decision theory first appeared.

This research was supported in parts by grants from the International Business Machines Corporation (IBM) (#A1771928), the National Science Foundation (#IIS-1302448, #IIS-1527490, and #IIS-1704932), and the Office of Naval Research (#N00014-17-12091).

Notes

1. The relation between P_x and P takes a variety of equivalent forms, including the backdoor formula, truncated factorization, adjustment for direct causes, or the inverse probability weighting shown in equation (2) (Pearl, 2000a, pp. 72–73). The latter form is the easiest to describe without appealing to graphical notation. But see equation (10) in section 3.1 for a more general formula and definition 3 for a formal definition of the set Z.

2. I will use "future" and "past" figuratively; "affected" and "unaffected" (by X) are more accurate technically (i.e., descendants and nondescendants of X, in graph-theoretical terminology). The derivation of equation (2) requires that processes be organized recursively (avoiding feedback loops); more intricate formulas apply to nonrecursive models. See Pearl (2009b, pp. 72–73) or Spirtes, Glymour, & Scheines (2001) for a simple derivation of this and equivalent formulas. Equation (2) has also been anticipated in Spohn (1978, sections 3.3 and 5.2).

3. The structural definition of counterfactuals given in equation (3) was first introduced in Balke and Pearl (1995).

4. By a "path," we mean a sequence of 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 nonrecursive models as well.

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

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

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 nondescendants (in Z) separate its descendants (in Z) from Y.

8. Summations should be replaced by integration when applied to continuous variables, as in Imai, Keele, and Yamamoto (2010).

9. Generalizations to arbitrary reference points, 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.

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

11. Remarkably, selection bias can be detected by combining experimental and observational studies, if certain coherence inequalities are violated (Pearl, 2009b, p. 294).

References

Arló-Costa, H. (2007). The logic of conditionals. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from http://plato.stanford.edu/archives/sum2019/entries/logic-conditionals/

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

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

Bareinboim, E., & Pearl, J. (2012b). Transportability of causal effects: Completeness results. In J. Hoffman & B. Selman (Eds.), *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence* (pp. 698–704). Menlo Park, CA: AAAI Press.

Bareinboim, E., & Pearl, J. (2013). Meta-transportability of causal effects: A formal approach. In C. M. Carvalho & P. Ravikumar (Eds.), *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics* (JMLR Workshop and Conference Proceedings, Vol. 31, pp. 135–143). Scottsdale, AZ: PMLR.

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

Bareinboim, E., Tian, J., & Pearl, J. (2014). Recovering from selection bias in causal and statistical inference. In C. E. Brodley & P. Stone (Eds.), *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (pp. 2410–2416). Palo Alto, CA: AAAI Press.

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

Cartwright, N. (1983). *How the laws of physics lie*. Oxford, England: Clarendon Press.

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

Geiger, D., Verma, T., & Pearl, J. (1990). *d*-separation: From theorems to algorithms. In M. Henrion, R. D. Shachter, L. N. Kanal, & J. F. Lemmer (Eds.), *Uncertainty in AI* (Vol. 5, pp. 139–148). Amsterdam, Netherlands: North-Holland.

Goldberger, A. S. (1984). Reverse regression and salary discrimination. *Journal of Human Resources*, *19*(3), 293–318.

Greenland, S., Pearl, J., & Robins, J. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, *10*(1), 37–48.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica*, *11*, 1–12.

Halpern, J. (1998). Axiomatizing causal reasoning. In G. Cooper & S. Moral (Eds.), *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence* (pp. 202–210). San Francisco, CA: Morgan Kaufmann.

Harper, W. L., Stalnaker, R., & Pearce, G. (1981). *Ifs*. Dordrecht, Netherlands: Reidel.

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

Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science*, *25*(1), 51–71.

Jeffrey, R. C. (1965). *The logic of decision*. New York, NY: McGraw-Hill.

Joyce, J. (1999). *The foundations of causal decision theory*. Cambridge, England: Cambridge University Press.

Lewis, D. (1973). Counterfactuals and comparative possibility. *Journal of Philosophical Logic*, *2*(4), 418–446.

Pearl, J. (1993). Comment: Graphical models, causality, and intervention. *Statistical Science*, *8*(3), 266–269.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–710.

Pearl, J. (2000a). *Causality: Models, reasoning, and inference*. New York, NY: Cambridge University Press.

Pearl, J. (2000b). Comment on A. P. Dawid's "Causal inference without counterfactuals." *Journal of the American Statistical Association*, *95*(450), 428–431.

Pearl, J. (2001). Direct and indirect effects. In J. S. Breese & D. Koller (Eds.), *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (pp. 411–420). San Francisco, CA: Morgan Kaufmann.

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

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

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

Pearl, J. (2013). The curse of free-will and the paradox of inevitable regret. *Journal of Causal Inference*, *1*, 255–257.

Pearl, J. (2017). Physical and metaphysical counterfactuals: Evaluating disjunctive actions. *Journal of Causal Inference*, *5*(2), 20170018, eISSN 2193–3685.

Pearl, J., & Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach. In W. Burgard & D. Roth (Eds.), *Proceedings of the Twenty-Fifth Conference on Artificial Intelligence (AAAI-11)* (pp. 247–254). Menlo Park, CA: AAAI Press. Available at https://ftp.cs.ucla.edu/pub/stat_ser/r372-aaai-corrected-reprint.pdf

Pearl, J., Glymour, M., & Jewell, N. (2016). *Causal inference in statistics: A primer*. New York, NY: Wiley.

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. New York, NY: Basic Books.

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

Robins, J., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, *3*(2), 143–155.

Rosenbaum, P. R., & Rubin, D. B. (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*(6), 1011–1035.

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

Shpitser, I., VanderWeele, T., & Robins, J. M. (2010). On the validity of covariate adjustment for estimating causal effects. In P. Grünwald & P. Spirtes (Eds.), *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence* (pp. 527–536). Corvallis, OR: AUAI Press.

Skyrms, B. (1980). *Causal necessity*. New Haven, CT: Yale University Press.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. New York, NY: Springer.

Spirtes, P., Glymour, C., & Scheines, R. (2001). *Causation, prediction, and search* (2nd ed.). Cambridge, MA: MIT Press.

Spohn, W. (1978). *Grundlagen der Entscheidungstheorie* [Foundations of decision theory]. Kronberg/Taunus, Germany: Scriptor.

Spohn, W. (1983). *Eine Theorie der Kausalität* (unpublished Habilitationsschrift) [A theory of causality]. Munich, Germany: Ludwig Maximilian University.

Stalnaker, R. (1981). Letter to David Lewis. In W. Harper, R. Stalnaker, & G. Pearce, *Ifs* (pp. 151–152). Dordrecht, Netherlands: Reidel. (Original work published 1972)

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

Tian, J., Paz, A., & Pearl, J. (1998). *Finding minimal separating sets* (Tech. Rep. No. R-254). Los Angeles: University of California.

Tian, J., & Pearl, J. (2002). A general identification condition for causal effects. In R. Dechter, M. Kearns, & R. S. Sutton (Eds.), *Proceedings of the Eighteenth National Conference on Artificial Intelligence* (pp. 567–573). Menlo Park, CA: AAAI Press/MIT Press.

Verma, T., & Pearl, J. (1988). Causal networks: Semantics and expressiveness. In R. D. Shachter, T. S. Levitt, L. N. Kanal, & J. F. Lemmer (Eds.), *Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence* (pp. 352–359). Mountain View, CA: AUAI Press.

Weirich, P. (2016). Causal decision theory. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from http://plato.stanford.edu/archives/win2016/entries/decision-causal/

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford, England: Oxford University Press.

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Knauff, Markus, editor. | Spohn, Wolfgang, editor.
Title: The handbook of rationality / edited by Markus Knauff and Wolfgang Spohn.
Description: Cambridge : The MIT Press, 2021. | Includes bibliographical references and index.
Identifiers: LCCN 2020048455 | ISBN 9780262045070 (hardcover)
Subjects: LCSH: Reasoning (Psychology) | Reason. | Cognitive psychology. | Logic. | Philosophy of mind.
Classification: LCC BF442 .H36 2021 | DDC 153.4/3—dc23
LC record available at https://lccn.loc.gov/2020048455