In H. Geffner, R. Dechter, and J.Y. Halpern (Eds.), Probabilistic and Causal Inference: The Works
of Judea Pearl, Association for Computing Machinery, 2021.

27

On Pearl's Hierarchy and the Foundations of Causal Inference

Elias Bareinboim (Columbia University),
Juan D. Correa (Columbia University),
Duligur Ibeling (Stanford University),
Thomas Icard (Stanford University)

Abstract

Cause-and-effect relationships play a central role in how we perceive and make sense of the world around us, how we act upon it, and ultimately, how we understand ourselves. Almost two decades ago, computer scientist Judea Pearl made a breakthrough in understanding causality by discovering and systematically studying the "Ladder of Causation," a framework that highlights the distinct roles of seeing, doing, and imagining. In honor of this landmark discovery, we name this the Pearl Causal Hierarchy (PCH). In this chapter, we develop a novel and comprehensive treatment of the PCH through two complementary lenses: one logical-probabilistic and another inferential-graphical. Following Pearl's own presentation of the hierarchy, we begin by showing how the PCH organically emerges from a well-specified collection of causal mechanisms (a structural causal model, or SCM). We then turn to the logical lens. Our first result, the Causal Hierarchy Theorem (CHT), demonstrates that the three layers of the hierarchy almost always separate in a measure-theoretic sense. Roughly speaking, the CHT says that data at one layer virtually always underdetermines information at higher layers. As in most practical settings the scientist does not have access to the precise form of the underlying causal mechanisms—only to data generated by them with respect to some of the PCH's layers—this motivates us to study inferences within the PCH through the graphical lens. Specifically, we explore a set of methods known as causal inference that enable inferences bridging the PCH's layers given a partial

specification of the SCM. For instance, one may want to infer what would happen had an intervention been performed in the environment (second-layer statement) when only passive observations (first-layer data) are available. We introduce a family of graphical models that allows the scientist to represent such a partial specification of the SCM in a cognitively meaningful and parsimonious way. Finally, we investigate an inferential system known as do-calculus, showing how it can be sufficient, and in many cases necessary, to allow inferences across the PCH's layers. We believe that connecting with the essential dimensions of human experience as delineated by the PCH is a critical step toward creating the next generation of artificial intelligence (AI) systems that will be safe, robust, human-compatible, and aligned with the social good.

27.1 Introduction

Causal information is deemed highly valuable and desirable along many dimensions of the human endeavor, including science, engineering, business, and law. The ability to learn, process, and leverage causal information is arguably a distinctive feature of *Homo sapiens* when compared to other species, perhaps one of the hallmarks of human intelligence [Penn and Povinelli 2007]. Pearl argued for the centrality of causal reasoning eloquently in his most recent book [Pearl and Mackenzie 2018, p. 1], for instance: "Some tens of thousands of years ago, humans began to realize that certain things cause other things and that tinkering with the former can change the latter... From this discovery came organized societies, then towns and cities, and eventually the science and technology-based civilization we enjoy today. All because we asked a simple question: Why?"

Given the centrality of causation throughout so many aspects of human experience, we would naturally like to have a formal framework for encoding and reasoning with cause-and-effect relationships. Interestingly, the 20th century saw other instances in which an intuitive, ordinary concept underwent mathematical formalization before entering engineering practice. As an especially notable example, it may be surprising to readers outside computer science and related disciplines to learn that the notion of *computation* itself was only semi-formally understood up until the 1920s. Following the seminal work of mathematician and philosopher Alan Turing, among others, multiple breakthroughs ensued, including the very emergence of the modern computer, passing through the theory and foundations of computer science, and culminating in the rich and varied technological advances we enjoy today.

We feel it is appropriate in this special edition dedicated to Judea Pearl, a Turing awardee himself, to recognize a similar historical development in the discipline of causality. The subject was studied in a semi-formal way for centuries

[Hume 1739, 1748, von Wright 1971, Mackie 1980], to cite a few prominent references, and Pearl, his collaborators, and many others helped to understand and formalize this notion. Following this precise mathematization, we now see a blossoming of developments and rapid expansion toward applications.

What was the crucial development that spawned such dramatic progress on this centuries-old problem? One critical insight, tracing back at least to the British empiricist philosophers, is that the causal mechanisms behind a system under investigation are not generally observable, but they do produce observable traces ("data," in modern terminology).[1] That is, "reality" and the data generated by it are fundamentally distinct. This dichotomy has been prominent at least since Pearl's seminal *Biometrika* paper [Pearl 1995], and received central status and comprehensive treatment in his longer treatise [Pearl 2000]. This insight naturally leads to two practical desiderata for any proper framework for causal inference, namely:

1. The causal mechanisms underlying the phenomenon under investigation should be accounted for—indeed, formalized—in the analysis.
2. This collection of mechanisms (even if mostly unobservable) should be formally tied to its output: the generated phenomena and corresponding datasets.

This intuitive picture is illustrated in Figure 27.1(a). One of the main goals of this chapter is to make this distinction crisp and unambiguous, translating these two desiderata into a formal framework, and uncovering its consequences for the practice of causal inference.

Regarding the first requirement, the underlying reality ("ground truth") that is our putative target can be naturally represented as a collection of causal mechanisms in the form of a mathematical object called a *structural causal model* (SCM) [Pearl 1995, 2000], to be introduced in Section 27.2. In many practical settings, it may be challenging, even impossible, to determine the specific form of the underlying causal mechanisms, especially when high-dimensional, complex phenomena are involved and humans are present in the loop.[2] Nevertheless, we ordinarily

[1] For instance, Locke famously argued that when we observe data, we cannot "so much as guess, much less know, their manner of production" [Locke 1690, Essay IV]. Hume maintained a similarly skeptical stance, stating that "nature has kept us at a great distance from all her secrets, and has afforded only the knowledge of a few superficial qualities of objects; while she conceals from us those powers and principles, on which the influence of these objects entirely depends" [Hume 1748, section 4.16]. See de Pierris [2015] for a discussion.

[2] At the same time, many of the natural sciences, most prominently physics and chemistry, will often purport to determine the underlying causal mechanisms quite precisely, under strict experimental conditions.

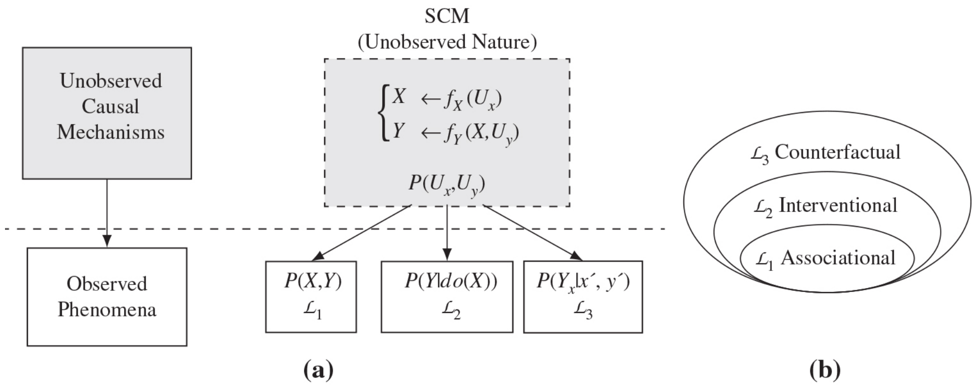

SCM
(Unobserved Nature)
Unobserved Causal Mechanisms
$\begin{cases} X \leftarrow f_X(U_x) \\ Y \leftarrow f_Y(X, U_y) \end{cases}$
$P(U_x, U_y)$
Observed Phenomena
$P(X,Y)$ $\mathcal{L}_1$
$P(Y|do(X))$ $\mathcal{L}_2$
$P(Y_x|x', y')$ $\mathcal{L}_3$
$\mathcal{L}_3$ Counterfactual
$\mathcal{L}_2$ Interventional
$\mathcal{L}_1$ Associational
(a)
(b)

Figure 27.1 (a) Collection of causal mechanisms (or SCM) generating certain observed phenomena (qualitatively different probability distributions). (b) PCH's containment structure.

presume that these causal mechanisms are there regardless of our practical ability to discover their form, shape, and specific details.

Regarding the second requirement, Pearl further noted something very basic and fundamental, namely, that each collection of causal mechanisms (i.e., SCM) induces a causal hierarchy (or "ladder of causation"), which highlights qualitatively different aspects of the underlying reality. We fondly name this the Pearl Causal Hierarchy (PCH, for short), for he was the first to identify and study it systematically [Pearl 1995, 2000, Pearl and Mackenzie 2018]. The hierarchy consists of three layers (or "rungs") encoding different concepts: the associational, the interventional, and the counterfactual, corresponding roughly to the ordinary human activities of seeing, doing, and imagining, respectively [Pearl and Mackenzie 2018, chapter 27]. Knowledge at each layer allows reasoning about different classes of causal concepts, or "queries." Layer 1 deals with purely "observational," factual information. Layer 2 encodes information about what would happen, hypothetically speaking, were some intervention to be performed, namely, effects of actions. Finally, Layer 3 involves queries about what would have happened, counterfactually speaking, had some intervention been performed, given that something else in fact occurred (possibly conflicting with the hypothetical intervention). The hierarchy establishes a useful classification of concepts that might be relevant for a given task, thereby also classifying formal frameworks in terms of the questions that they are able to represent and, ideally, answer.

27.1.1 Roadmap of the Chapter

Against this background, we start in Section 27.2 by showing how the PCH naturally emerges from an SCM, formally characterizing the layers by means of symbolic logical languages, each of which receives a straightforward interpretation in an SCM. Thus, as soon as one admits that a domain of interest can be represented by an SCM (whether or not we, as an epistemological matter, know much about it), the hierarchy of causal concepts already exists.[3] In Section 27.3, we prove that the PCH is strict for almost-all SCMs (Theorem 27.1), in a technical sense of "almost-all" (Figure 27.1(b)).[4] It follows (Corollary 27.1) that it is *generically impossible* to draw higher-layer inferences using only lower-layer information, a result known informally in the field under the familiar adage: "no causes-in, no causes-out" [Cartwright 1989].

In the second part of the chapter (Section 27.4), we acknowledge that in many practical settings our ability to interact with (observe and experiment on) the phenomenon of interest is modest at best, and inducing a reasonable, fully specified SCM is essentially hopeless.[5] Virtually all approaches to causal inference, therefore, set for themselves a more restricted target, operating under the less-stringent condition that only partial knowledge of the underlying SCM is available. The problem of causal inference is thus to perform inferences across layers of the hierarchy from a partial understanding of the SCM. Technically speaking, if one has Layer-1 type of data, for example, collected through random sampling, and aims to infer the effect of a new intervention (Layer-2 type of query), we show that the problem is not always solvable.

Departing from these impossibility results, we develop a framework that can parsimoniously and efficiently encode knowledge (viz., structural constraints) necessary to perform this general class of inferences. In particular, we move beyond Layer-1 type constraints (conditional independences) and investigate structural constraints that live in Layer 2. We use these constraints to define a new family

[3] This is despite skepticism that has been expressed in the literature about meaningfulness of one layer of the hierarchy or another; cf., for example, Maudlin [2019] on Layer 2, and Dawid [2000] on Layer 3.

[4] Hierarchies abound in logic and computer science, particularly those pertaining to computational resources, with prominent examples being the Chomsky–Schützenberger hierarchy [Chomsky 1959] and its probabilistic variant (see Icard [2020]), or the polynomial time complexity hierarchy [Stockmeyer 1977]. Such hierarchies delimit what can be computed given various bounds on computational resources. Perhaps surprisingly, the Pearl hierarchy is orthogonal to these hierarchies. If one's representation language is only capable of encoding queries at a given layer, no amount of time or space for computation—and no amount of data either—will allow making inferences at higher layers.

[5] Of course, if we have been able to induce the structural mechanisms themselves—as may be feasible in some of the sciences, for example, molecular biology or Newtonian physics—we can simply "read off" any causal information we like by computing it directly or, for instance, by simulating the corresponding mechanisms.

of graphical models called *causal Bayesian networks* (CBNs), which are composed of a pair, a graphical model, and a collection of observational and interventional distributions. Against this backdrop, we provide a novel proof of *do-calculus* [Pearl 1995] based strictly on Layer 2 semantics. We then show how the graphical structure bridges the layers of the PCH; one may be able to draw inferences at a higher layer from a combination of partial knowledge of the underlying structural model, in the form of a causal graph, and data at lower layers. We conclude and summarize this chapter in Section 27.5.

27.1.2 Notation

We now introduce the notation used throughout this chapter. Single random variables are denoted by (non-boldface) uppercase letters X and the range (or possible values) of X is written as $\text{Val}(X)$. Lowercase x denotes a particular element in this range, $x \in \text{Val}(X)$. Boldfaced uppercase $\mathbf{X}$ denotes a collection of variables, $\text{Val}(\mathbf{X})$ their possible joint values, and boldfaced lowercase $\mathbf{x}$ a particular joint realization $\mathbf{x} \in \text{Val}(\mathbf{X})$. For example, two independent fair coin flips are represented by $\mathbf{X} = \{X_1, X_2\}$, $\text{Val}(X_1) = \text{Val}(X_2) = \{0, 1\}$, $\text{Val}(\mathbf{X}) = \{(0, 0), \ldots, (1, 1)\}$, with $P(x_1) = P(x_2) = \sum_{x_2} P(x_1, x_2) = \sum_{\mathbf{x}(X_1)=x_1} P(\mathbf{x}) = 1/2$.

27.2 Structural Causal Models and the Causal Hierarchy

We build on the language of SCMs to describe the collection of mechanisms underpinning a phenomenon of interest. Essentially, any causal inference can be seen as an inquiry about these mechanisms or their properties, in some way or another. We will generally dispense with the distinction between the underlying system and its SCM.

Each SCM naturally defines a qualitative hierarchy of concepts, described as the "ladder of causation" in Pearl and Mackenzie [2018], which we have been calling the PCH (Figure 27.1). Following Pearl's presentation, we label the layers (or rungs, or levels) of the hierarchy *associational*, *interventional*, and *counterfactual*. The concepts of each layer can be described in a formal language and correspond roughly to distinct notions within human cognition. Each of these allows one to articulate, with mathematical precision, qualitatively different types of questions regarding the observed variables of the underlying system; for some examples, see Table 27.1.

SCMs provide a flexible formalism for data-generating models, subsuming virtually all of the previous frameworks in the literature. In the sequel, we formally define SCMs and then show how a fully specified model underpins the concepts in the PCH.

Table 27.1 **Pearl's Causal Hierarchy**

	Layer (Symbolic)	Typical Activity	Typical Question	Example	Machine Learning
$\mathcal{L}_1$	Associational $P(y \mid x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell us about the disease?	Supervised/ Unsupervised Learning
$\mathcal{L}_2$	Interventional $P(y \mid do(x), c)$	Doing	What if? What if I do X?	What if I take aspirin, will my headache be cured?	Reinforcement Learning
$\mathcal{L}_3$	Counterfactual $P(y_x \mid x', y')$	Imagining	Why? What if I had acted differently?	Was it the aspirin that stopped my headache?	

Definition 27.1 **Structural Causal Model (SCM)**

An SCM $\mathcal{M}$ is a 4-tuple $\langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$, where

- $\mathbf{U}$ is a set of background variables, also called exogenous variables, that are determined by factors outside the model;
- $\mathbf{V}$ is a set $\{V_1, V_2, \ldots, V_n\}$ of variables, called endogenous, that are determined by other variables in the model—that is, variables in $\mathbf{U} \cup \mathbf{V}$;
- $\mathcal{F}$ is a set of functions $\{f_1, f_2, \ldots, f_n\}$ such that each f_i is a mapping from (the respective domains of) $U_i \cup Pa_i$ to V_i, where $U_i \subseteq \mathbf{U}$, $Pa_i \subseteq \mathbf{V} \backslash V_i$, and the entire set $\mathcal{F}$ forms a mapping from $\mathbf{U}$ to $\mathbf{V}$. That is, for $i = 1, \ldots, n$, each $f_i \in \mathcal{F}$ is such that

$$v_i \leftarrow f_i(pa_i, u_i), \tag{27.1}$$

  that is, it assigns a value to V_i that depends on (the values of) a select set of variables in $\mathbf{U} \cup \mathbf{V}$; and
- $P(\mathbf{U})$ is a probability function defined over the domain of $\mathbf{U}$. ■

Each SCM can be seen as partitioning the variables involved in the phenomenon into sets of exogenous (unobserved) and endogenous (observed) variables, respectively, $\mathbf{U}$ and $\mathbf{V}$. The exogenous ones are determined "outside" of the model and their associated probability distribution, $P(\mathbf{U})$, represents a summary of the state of the world outside the phenomenon of interest. In many settings, these variables represent the *units* involved in the phenomenon, which correspond to elements of the population under study, for instance, patients, students, and customers. Naturally, their randomness (encoded in $P(\mathbf{U})$) induces variations in the endogenous set $\mathbf{V}$.

Inside the model, the value of each endogenous variable V_i is determined by a causal process, $v_i \leftarrow f_i(pa_i, u_i)$, that maps the exogenous factors U_i and a set of endogenous variables Pa_i (so-called parents) to V_i. These causal processes—or mechanisms—are assumed to be invariant unless explicitly intervened on (as defined later in the section).[6] Together with the background factors, they represent the data-generating process according to which Nature assigns values to the endogenous variables in the study.

Henceforth, we assume that $\mathbf{V}$ and its domain are finite,[7] and that the model is acyclic (sometimes known as *recursive*).[8] A structural model is *Markovian* if the exogenous parent sets U_i, U_j are independent whenever $i \neq j$. Here, we will allow for the sharing of exogenous parents and for arbitrary dependences among the exogenous variables, which means that, in general, the SCM need not be Markovian. This wider class of models is called *semi-Markovian*. For concreteness, we provide a simple SCM next.

Example 27.1 Consider a game of chance described through the SCM $\mathcal{M}^1 = \langle \mathbf{U} = \{U_1, U_2\}, \mathbf{V} = \{X, Y\}, \mathcal{F}, P(U_1, U_2)\rangle$, where

$$\mathcal{F} = \begin{cases} X & \leftarrow U_1 + U_2 \\ Y & \leftarrow U_1 - U_2 \end{cases}, \tag{27.2}$$

and $P(U_i = k) = 1/6$, $i = 1, 2$, $k = 1, \ldots, 6$. In other words, this structural model represents the setting in which two dice are rolled but only the sum (X) and the difference (Y) of their values are observed. Here, $\text{Val}(X) = \{2, \ldots, 12\}$ and $\text{Val}(Y) = \{-5, \ldots, 0, \ldots, 5\}$. ■

[6] It is possible to conceive an SCM as "a high-level abstraction of an underlying system of differential equations" [Schölkopf 2019], which under relatively mild conditions is attainable [Rubenstein et al. 2017].

[7] Much of the theory of SCMs extends straightforwardly to the infinitary setting [Ibeling and Icard 2019].

[8] An SCM $\mathcal{M}$ is said to be recursive if there exists a "temporal" order over the functions in $\mathcal{F}$ such that for every pair $f_i, f_j \in \mathcal{F}$, if $f_i < f_j$ in the order, we have that f_i does not have V_j as an argument. In particular, this implies that choosing a unit $\mathbf{u}$ uniquely fixes the values of all variables in $\mathbf{V}$. For $\mathbf{Y} \subseteq \mathbf{V}$, we write $\mathbf{Y}(\mathbf{u})$ to denote the solution of $\mathbf{Y}$ given unit $\mathbf{u}$. For a more comprehensive discussion, see Galles and Pearl [1998] and Halpern [1998, 2000].

27.2.1 Pearl Hierarchy, Layer 1—Seeing

Layer 1 of the hierarchy (Table 27.1) captures the notion of "seeing," that is, observing a certain phenomenon unfold, and perhaps making inferences about it. For instance, if we observe a certain symptom, how will this change our belief in the disease? An SCM gives natural valuations for quantities of this kind (cf. equation (7.2) in Pearl [2000]), as shown next.

Definition 27.2 **Layer 1 Valuation—"Observing"**

An SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$ defines a joint probability distribution $P^{\mathcal{M}}(\mathbf{V})$ such that for each $\mathbf{Y} \subseteq \mathbf{V}$:[9]

$$P^{\mathcal{M}}(\mathbf{y}) = \sum_{\{\mathbf{u} \mid \mathbf{Y}(\mathbf{u})=\mathbf{y}\}} P(\mathbf{u}), \tag{27.3}$$

where $\mathbf{Y}(\mathbf{u})$ is the solution for $\mathbf{Y}$ after evaluating $\mathcal{F}$ with $\mathbf{U} = \mathbf{u}$. ■

This evaluation is graphically depicted in Figure 27.2(i), which represents a mapping from the external and unobserved state of the system (distributed as $P(\mathbf{U})$), to an observable state (distributed as $P(\mathbf{V})$). For concreteness, let us consider Example 27.1 again. Let the dice (exogenous variables) be $\langle U_1 = 1, U_2 = 1 \rangle$, then $\mathbf{V} = \{X, Y\}$ attain their values through $\mathcal{F}$ as $X = 1 + 1 = 2$ and $Y = 1 - 1 = 0$. As $P(U_1 = 1, U_2 = 1) = 1/36$ and $\langle U_1 = 1, U_2 = 1 \rangle$ is the only configuration capable of producing the observed behavior $\langle X = 2, Y = 0 \rangle$, it follows that $P(X = 2, Y = 0) = 1/36$. More interestingly, consider the different dice (exogenous) configurations $\langle U_1, U_2 \rangle = \{\langle 1, 1 \rangle, \langle 2, 2 \rangle, \langle 3, 3 \rangle, \langle 4, 4 \rangle, \langle 5, 5 \rangle, \langle 6, 6 \rangle\}$, which are all compatible with $\langle Y = 0 \rangle$. As each of the $\mathbf{U}$'s realization happens with probability 1/36, the event of the difference between the first and second dice being zero ($Y = 0$) occurs with probability 1/6. Finally, what is the probability of the difference of the two dice being zero ($Y = 0$) if we know that their sum is two, that is, $P(Y = 0 \mid X = 2)$? The answer is one as the only event compatible with $\langle X = 2, Y = 0 \rangle$ is $\langle U_1 = 1, U_2 = 1 \rangle$. Without any evidence, the event ($Y = 0$) happens with probability 1/6, yet if we know that $X = 2$, the event becomes certain (probability 1).

Many tasks throughout data sciences can be seen as evaluating the probability of certain events occurring. In the context of modern machine learning, for example, one could observe a certain collection of pixels, or features, with the goal of predicting whether it contains a dog or a cat. Consider a slightly more involved example that appears in the context of medical decision-making.

[9] We will typically omit the superscript on $P^{\mathcal{M}}$ whenever there is no room for confusion, thus using P for both the distribution $P(\mathbf{U})$ on exogenous variables and the distributions $P(\mathbf{Y})$ on endogenous variables induced by the SCM.

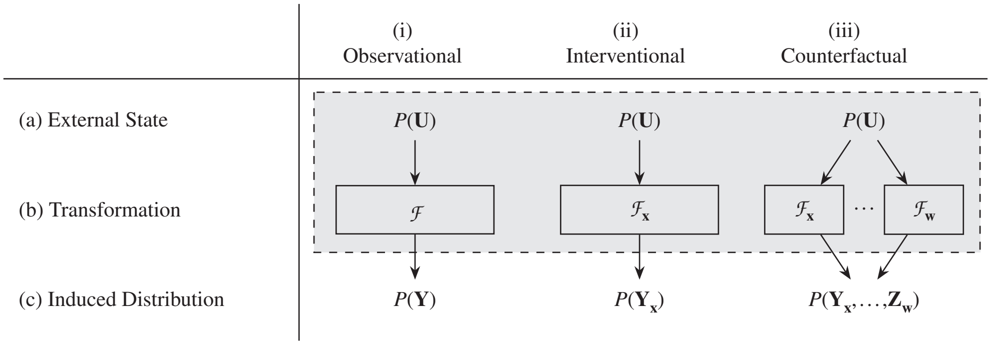


Figure 27.2 Given an SCM's initial state (i.e., population) (a), we show the different functional transformations (b) and the corresponding induced distribution (c) of each layer of the hierarchy. (i) represents the transformation (i.e., $\mathcal{F}$) from the natural state of the system ($P(\mathbf{U})$) to an observational world, (ii) to an interventional world (i.e., with modified mechanisms $\mathcal{F}_x$), and (iii) to multiple counterfactual worlds (i.e., with multiple modified mechanisms).

Example 27.2 The SCM $\mathcal{M}^2 = \langle \mathbf{V} = \{X, Y, Z\}, \mathbf{U} = \{U_r, U_x, U_y, U_z\}, \mathcal{F} = \{f_x, f_y, f_z\}, P(U_r, U_x, U_y, U_z)\rangle$, where $\mathcal{F}$ will be specified below. The endogenous variables $\mathbf{V}$ represent, respectively, a certain treatment X (e.g., drug), an outcome Y (survival), and the presence or not of a symptom Z (hypertension). The exogenous variable U_r represents whether the person has a certain natural resistance to the disease, and U_x, U_y, U_z are sources of variations outside the model affecting X, Y, Z, respectively. In this population, units with resistance ($U_r = 1$) are likely to survive ($Y = 1$) regardless of the treatment received. Whenever the symptom is present ($Z = 1$), physicians try to counter it by prescribing this drug ($X = 1$). While the treatment ($X = 1$) helps resistant patients (with $U_r = 1$), it worsens the situation for those who are not resistant ($U_r = 0$). The form of the underlying causal mechanisms is:

$$\mathcal{F} = \begin{cases} Z & \leftarrow \mathbb{1}_{\{U_r=1, U_z=1\}} \\ X & \leftarrow \mathbb{1}_{\{Z=1, U_x=1\}} + \mathbb{1}_{\{Z=0, U_x=0\}} \\ Y & \leftarrow \mathbb{1}_{\{X=1, U_r=1\}} + \mathbb{1}_{\{X=0, U_r=1, U_y=1\}} + \mathbb{1}_{\{X=0, U_r=0, U_y=0\}} \end{cases} . \tag{27.4}$$

Finally, all the exogenous variables are binary with $P(U_r = 1) = 0.25$, $P(U_z = 1) = 0.95$, $P(U_x = 1) = 0.9$, and $P(U_y = 1) = 0.7$.

Recall that Definition 27.2 (Equation 27.3) induces a mapping between $P(\mathbf{U})$ and $P(\mathbf{V})$, such that a query $P(Y = 1 \mid X = 1)$ can be evaluated from $\mathcal{M}$ as:

$$P(Y = 1 \mid X = 1) = \frac{P(Y = 1, X = 1)}{P(X = 1)} = \frac{\sum_{\{\mathbf{u} \mid Y(\mathbf{u})=1, X(\mathbf{u})=1\}} P(\mathbf{u})}{\sum_{\{\mathbf{u} \mid X(\mathbf{u})=1\}} P(\mathbf{u})} = \frac{0.215}{0.29} = 0.7414, \tag{27.5}$$

which is just the ratio between the sum of the probabilities of the events in the space of $\mathbf{U}$ consistent with the events $\langle Y = 1, X = 1\rangle$ and $\langle X = 1\rangle$. This means that the probability of survival given that one took the drug is higher than chance. Similarly, one could obtain other probabilistic expressions such as $P(Y = 1|X = 0) = 0.3197$ or $P(Z = 1) = 0.2375$. One may be tempted to believe at this point that the drug has a positive effect upon comparing the probabilities $P(Y = 1|X = 0)$ and $P(Y = 1|X = 1)$. We shall discuss this issue next. ■

27.2.2 Pearl Hierarchy, Layer 2—Doing

Layer 2 of the hierarchy (Table 27.1) allows one to represent the notion of "doing," that is, intervening (acting) in the world to bring about some state of affairs. For instance, if a physician gives a drug to her patient, would the headache be cured? A modification of an SCM gives natural valuations for quantities of this kind, as defined next.

Definition 27.3 **Submodel—"Interventional SCM"**

Let $\mathcal{M}$ be a causal model, $\mathbf{X}$ a set of variables in $\mathbf{V}$, and $\mathbf{x}$ a particular realization of $\mathbf{X}$. A submodel $\mathcal{M}_x$ of $\mathcal{M}$ is the causal model

$$\mathcal{M}_{\mathbf{x}} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}_{\mathbf{x}}, P(\mathbf{U})\rangle, \quad \text{where } \mathcal{F}_{\mathbf{x}} = \{f_i : V_i \notin \mathbf{X}\} \cup \{\mathbf{X} \leftarrow \mathbf{x}\}. \tag{27.6}$$

■

In other words, performing an external intervention (or action) is modeled through the replacement of the original (natural) mechanisms associated with some variables $\mathbf{X}$ with a constant $\mathbf{x}$, which is represented by the *do*-operator.[10,11] The impact of the intervention on an outcome variable Y is called *potential response* (cf. definition (7.1.4) in Pearl [2000]).

Definition 27.4 **Potential Response**

Let $\mathbf{X}$ and $\mathbf{Y}$ be two sets of variables in $\mathbf{V}$, and $\mathbf{u}$ be a unit. The potential response

[10] The idea of representing intervention through the modification of equations in a structural system appears to have first emerged in the context of Econometrics, see Haavelmo [1943], Marschak [1950], and Simon [1953]. It was then made more explicit and called "wiping out" by Strotz and Wold [1960].

[11] Pearl credits his realization on the connection of this operation with graphical models to a lecture of Peter Spirtes at the International Congress on Logic, Methodology and Philosophy of Science (Uppsala, Sweden, 1991), in his words [Pearl 2000, p. 104]: "In one of his slides, Peter illustrated how a causal diagram would change when a variable is manipulated. To me, that slide of Spirtes's—when combined with the deterministic structural equations—was the key to unfolding the manipulative account of causation (...)."

$\mathbf{Y_x}(\mathbf{u})$ is defined as the solution for $\mathbf{Y}$ of the set of equations $\mathcal{F}_{\mathbf{x}}$ with respect to SCM $\mathcal{M}$ (for short, $\mathbf{Y}_{\mathcal{M}_\mathrm{x}}(\mathrm{u})$). That is, $\mathbf{Y_x}(\mathbf{u}) = \mathbf{Y}_{\mathcal{M}_\mathrm{x}}(\mathrm{u})$. ■

An SCM gives valuation for interventional quantities (equation 7.3 Pearl [2000]) as follows:

Definition 27.5 **Layer 2 Valuation—"Intervening"**
An SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U})\rangle$ induces a family of joint distributions over $\mathbf{V}$, one for each intervention $\mathbf{x}$. For each $\mathbf{Y} \subseteq \mathbf{V}$:

$$P^{\mathcal{M}}(\mathbf{y_x}) = \sum_{\{\mathbf{u}\,|\,\mathbf{Y_x}(\mathbf{u})=\mathbf{y}\}} P(\mathbf{u}). \tag{27.7}$$

■

The *potential response* expresses causal effects, and over a probabilistic setting it induces random variables. Specifically, Y_x denotes a random variable induced by averaging the potential response $Y_x(\mathbf{u})$ over all $\mathbf{u}$ according to $P(\mathbf{U})$.[12] Further, note that this procedure disconnects X from any other source of "natural" variation when it follows the original function f_x (e.g., the observed (Pa_x) or unobserved (U_x) parents). This means that the variations of Y in this world would be due to changes in X (say, from 0 to 1) that occurred externally, from outside the modeled system.[13] This, in turn, guarantees that they will be *causal*. To see why, note that all variations of X that may have an effect on Y can only be realized through variables of which X is an argument, as X itself is a constant, not affected by other variables. Indeed, the notion of an *average causal effect* can be formally written as $E(Y_{X=1}) - E(Y_{X=0})$.[14]

The distribution $P(\mathbf{Y_x})$ defined in Equation (27.7) is often written $P(\mathbf{Y} \,|\, do(\mathbf{x}))$, and we henceforth adopt this notation in the context of PCH's second layer.[15]

[12] The notation $Y_x(u)$ is borrowed from the potential-outcome framework of Neyman [1923] and Rubin [1974]. See Pearl [2000, section 7.4.4] for a more detailed comparison; see also Pearl and Bareinboim [2019].

[13] For a discussion of what it means for these changes to arise "from outside" the system, see, for example, Woodward [2003]. Of course, in many settings this simply means the intervention is performed deliberately by an *agent* outside the system, for example, in typical reinforcement learning applications [Sutton and Barto 2018].

[14] This difference and the corresponding expected values are sometimes taken as the definition of "causal effect," see Rosenbaum and Rubin [1983]. In the structural account of causation pursued here, this quantity is not a primitive but derivable from the SCM, as all others within the PCH. To witness, note $Y_{X=1} \longleftarrow f_Y(1, \varepsilon_Y)$ when $do(X = 1)$.

[15] This allows researchers to use the syntax to immediately distinguish statements that are amenable to some sort of experimentation, at least in principle, from other counterfactuals that may be empirically unrealizable.

Example 27.3 **Example 27.1 continued**

Let us consider the same dice game but now the observer decides to misreport the sum of the two dice as 2, which can be written as submodel $\mathcal{M}_{X=2}$:

$$\mathcal{F}_{X=2} = \begin{cases} X & \leftarrow 2 \\ Y & \leftarrow U_1 - U_2, \end{cases} , \tag{27.8}$$

while $P(\mathbf{U})$ remains invariant. It can be immediately seen that $Y_{X=2}(u_1, u_2)$ is the same as $Y(u_1, u_2)$; in other words, misreporting the sum of the two dice will of course not change their difference. This, in turn, entails the following probabilistic invariance,

$$P(Y = 0 \,|\, do(X = 2)) = P(Y = 0). \tag{27.9}$$

In fact, the distribution of Y when X is fixed to two remains the same as before (i.e., $P(Y = 0 \,|\, do(X = 2)) = 1/6$). We saw in the first part of the example that knowing that the sum was two meant that, with probability one, their difference had to be zero (i.e., $P(Y = 0 \,|\, X = 2) = 1$). On the other hand, intervening on X will not change Y's distribution (Equation 27.9); as we say, X does not have a *causal effect* on Y. ■

Example 27.4 **Example 27.2 continued**

Consider now that a public health official performs an intervention by giving the treatment to all patients regardless of the symptom (Z). This means that the function f_X would be replaced by the constant 1. In other words, patients do not have an option of deciding their own treatment but are compelled to take the specific drug.[16] This is represented through the new modified set of mechanisms,

$$\mathcal{F}_{X=1} = \begin{cases} Z & \leftarrow \mathbb{1}_{\{U_r=1, U_z=1\}} \\ X & \leftarrow \mathbb{1} \\ Y & \leftarrow \mathbb{1}_{\{X=1, U_r=1\}} + \mathbb{1}_{\{X=0, U_r=1, U_y=1\}} + \mathbb{1}_{\{X=0, U_r=0, U_y=0\}} \end{cases} , \tag{27.10}$$

and where the distribution of exogenous variables remains the same. Note that the potential response $Y_{X=1}(\mathbf{u})$ represents the survival of patient $\mathbf{u}$ had they been treated, while the random variable $Y_{X=1}$ describes the average population survival

[16] This physical procedure is the very basis for the discipline of experimental design [Fisher 1936], which is realized through randomization of the treatment assignment in a sample of the population. In practice, the function of X, f_X, is replaced with an alternative source of randomness that is uncorrelated with any other variable in the system.

had everyone been given the treatment. Notice that for those patients who naturally received treatment ($X \leftarrow f_x(\mathbf{U}) = 1$), the natural outcome $Y(\mathbf{u})$ is equal to $Y_{X=1}(\mathbf{u})$. For this intervened model, $Y_{X=1}(\mathbf{u})$ is equal to 1 in every event where $U_r = 1$, regardless of U_z, U_x, and U_y. Then

$$P(Y = 1 \mid do(X = 1)) = \sum_{\{\mathbf{u} \mid Y_{X=1}(\mathbf{u})=1\}} P(\mathbf{u}) = \sum_{\{u_r \mid Y_{X=1}(u_r)=1\}} P(u_r) = P(U_r = 1) = 0.25. \tag{27.11}$$

Similarly, one can evaluate $P(Y = 1 \mid do(X = 0))$, which is equal to 0.4. This may be surprising as from the perspective of Layer 1, $P(Y = 1 \mid X = 1) - P(Y = 1 \mid X = 0) = 0.43 > 0$, which appears to suggest that taking the drug is helpful, having a positive effect on recovery. On the other hand, interventionally speaking, $P(Y = 1 \mid do(X = 1)) - P(Y = 1 \mid do(X = 0)) = -0.15 < 0$, which means that the drug has a negative (average) effect in the population. ■

The evaluation of an interventional distribution is a function of the modified system $\mathcal{M}_\mathbf{x}$ that reflects $\mathcal{F}_\mathbf{x}$, which follows from the replacement of $\mathbf{X}$, as illustrated in Figure 27.2(ii). Even though computing observational and interventional distributions is immediate from a fully specified SCM, the distinction between Layer 1 (seeing) and Layer 2 (doing) is a central topic in causal inference, as discussed more substantively in Section 27.4.

27.2.3 Pearl Hierarchy, Layer 3—Imagining Counterfactual Worlds

Layer 3 of the hierarchy (Table 27.1) allows operationalizing the notion of "imagination" (and the closely related activities of retrospection, prospection, and other forms of "modal" reasoning), that is, thinking about alternative ways the world could be, including ways that might conflict with how the world, in fact, currently is. For instance, if the patient took the aspirin and the headache was cured, would the headache still be gone had they not taken the drug? Or, if an individual ended up getting a great promotion, would this still be the case had they not earned a PhD? What if they had a different gender? Obviously, in this world, the person has a particular gender, has a PhD, and ended up getting the promotion, so we would need a way of conceiving and grounding these alternative possibilities to evaluate such scenarios. In fact, no experiment in the world (Layer 2) will be sufficient to answer this type of question in general, despite their ubiquity in human discourse, cognition, and decision-making. Fortunately, the meaning of every term in the counterfactual layer ($\mathcal{L}_3$) can be directly determined from a fully specified SCM, as described in the sequel:

Definition 27.6 **Layer 3 Valuation**

An SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$ induces a family of joint distributions over counterfactual events $\mathbf{Y}_\mathbf{x}, \dots, \mathbf{Z}_\mathbf{w}$, for any $\mathbf{Y}, \mathbf{Z}, \dots, \mathbf{X}, \mathbf{W} \subseteq \mathbf{V}$:

$$P^{\mathcal{M}}(\mathbf{y}_\mathbf{x}, \dots, \mathbf{z}_\mathbf{w}) = \sum_{\substack{\{\mathbf{u} \,|\, \mathbf{Y}_\mathbf{x}(\mathbf{u})=\mathbf{y}, \\ \dots,\, \mathbf{Z}_\mathbf{w}(\mathbf{u})=\mathbf{z}\}}} P(\mathbf{u}). \tag{27.12}$$

■

Note that the left-hand side (LHS) of Equation (27.12) contains variables with different subscripts, which, syntactically, encode different counterfactual "worlds."

Example 27.5 **Example 27.2 continued**

As there is a group of patients who did not receive the treatment and died ($X = 0$, $Y = 0$), one may wonder whether these patients would have been alive ($Y = 1$) had they been given the treatment ($X = 1$). In the language of Layer 3, this question is written as $P(Y_{X=1} = 1 | X = 0, Y = 0)$. This is a non-trivial question as these individuals did not take the drug and are already deceased in the actual world (as displayed after the conditioning bar, $X = 0, Y = 0$); the question is about an unrealized world and how these patients would have reacted had they been submitted to a different course of action (formally written before the conditioning bar, $Y_{X=1} = 1$). In other words, did they die because of the lack of treatment? Or would this fatal unfolding of events happen regardless of the treatment? Unfortunately, there is no conceivable experiment in which we could draw samples from $P(Y_{X=1} = 1 | X = 0, Y = 0)$, as these patients cannot be resuscitated and submitted to the alternative condition. This is the very essence of counterfactuals.

For simplicity, note that $P(Y_{X=1} = 1 | X = 0, Y = 0)$ can be written as the ratio $P(Y_{X=1} = 1, X = 0, Y = 0)/P(X = 0, Y = 0)$, where the denominator is trivially obtainable as it only involves observational probabilities (about one specific world, the factual one). The numerator, $P(Y_{X=1} = 1, X = 0, Y = 0)$, refers to two different worlds, which requires us to climb up to the third layer in order to formally specify the quantity of interest. Using the procedure dictated in Equation (27.12), we obtain

$$\begin{aligned} P(Y_{X=1} = 1 | X = 0, Y = 0) &= \frac{P(Y_{X=1} = 1, X = 0, Y = 0)}{P(X = 0, Y = 0)} \\ &= \frac{\sum_{\{\mathbf{u} \,|\, Y_{X=1}(\mathbf{u})=1, X(\mathbf{u})=0, Y(\mathbf{u})=0\}} P(\mathbf{u})}{\sum_{\{\mathbf{u} \,|\, X(\mathbf{u})=0, Y(\mathbf{u})=0\}} P(\mathbf{u})} = 0.0217. \end{aligned}$$

This evaluation is shown step by step in Bareinboim et al. [2020, appendix D], but we emphasize here that the expression in the numerator involves evaluating multiple worlds simultaneously (in this case, one factual and one related to intervention $do(X = 1)$), as illustrated in Figure 27.2(iii). The conclusion following from this counterfactual analysis is clear: even if we had given the treatment to everyone who did not survive, only around 2% would have survived. In other words, the drug would not have prevented their deaths. Another aspect of this situation worth examining is whether the treatment would have been harmful for those who did not get it and still survived, formally written in Layer 3 language as $P(Y_{X=1} = 1 | X = 0, Y = 1)$. Following the same procedure, we find that this quantity is 0.1079, which means that about 90% of such people would have died had they been given the treatment. While a uniform policy over the entire population would be catastrophic (as shown in Example 27.4), the physicians knew what they were doing in this case and were effective in choosing the treatment for the patients who could benefit more from it. ■

There are many other counterfactual quantities implied by a structural model, for example, the previous two quantities can be combined to form the *probability of necessity and sufficiency* (PNS) [Pearl 2000, chapter 9], written as $P(y_x, y'_{x'})$. The PNS encodes the extent to which a certain treatment to a particular outcome would be both necessary and sufficient. This quantity addresses a quintessential "why" question, where one wants to understand what caused a given event. Still in the purview of Layer 3, some critical applications demand that counterfactuals be nested inside other counterfactuals. For instance, consider the quantity $Y_{x,M_{x'}}$ that represents the counterfactual value of Y had X been x, and M had whatever value it would have taken had X been x'. In other words, for Y the value of X is x, while for M the value of X is x'. This type of nested counterfactuals allow us to write contrasts such as $P[Y_{x,M_x} - Y_{x,M_{x'}}]$, the so-called *indirect effect* on Y when X changes from x' to x [Pearl 2001]. The use of nested counterfactuals led to a very natural and general treatment of direct, indirect, and spurious effects, including a precise understanding of their relationship in non-linear systems [Pearl 2012, VanderWeele 2015, Zhang and Bareinboim 2018].

27.3 Pearl Hierarchy—A Logical Perspective

We have seen that each layer of the PCH corresponds to a different intuitive notion in human cognition: seeing, acting, and imagining. Table 27.1 presents characteristic questions associated with each of the layers. Layer 1 concerns questions like, "How likely is Y given that I observe X?" Layer 2 asks hypothetical ("conditional") questions such as, "How likely *would* Y be if one were to make X happen?" Layer 3 takes us further, allowing questions like, "Given that I observed X and Y, how likely would Y have been if X' had been true instead of X?"

What does the difference among these questions amount to, given that an SCM answers all of them? Implicit in our presentation was a series of increasingly complex symbolic languages (Definitions 27.2, 27.5, and 27.6). Each type of question above can be phrased in one of these languages, the analysis of which reveals a logical perspective on PCH. We begin our analysis by isolating the syntax of these systems. We define languages $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_3$, each based on polynomials built over basic probability terms $P(\alpha)$. The only differences among them are the terms $P(\alpha)$ allowed: as we go up in the PCH, increasingly complex expressions α are allowed in the probability terms. In particular, $\mathcal{L}_1$ is just a familiar probabilistic logic (see Fagin et al. [1990]).

Definition 27.7 **Symbolic Languages $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_3$**

Let variables $\mathbf{V}$ be given and $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$. Each language $\mathcal{L}_i$, $i = 1, 2, 3$, consists of (Boolean combinations of) inequalities between polynomials over terms $P(\alpha)$, where $P(\alpha)$ is an $\mathcal{L}_i$ term, defined as follows:

$\mathcal{L}_1$ terms are those of the form $P(\mathbf{Y} = \mathbf{y})$, encoding the probability that $\mathbf{Y}$ take on values $\mathbf{y}$;

$\mathcal{L}_2$ terms additionally include probabilities of *conditional* expressions, $P(\mathbf{Y}_{\mathbf{x}} = \mathbf{y})$, giving the probability that variables $\mathbf{Y}$ *would* take on values $\mathbf{y}$, were $\mathbf{X}$ to have values $\mathbf{x}$;

$\mathcal{L}_3$ terms encode probabilities over *conjunctions* of conditional (that is, $\mathcal{L}_2$) expressions, $\mathrm{P}(\mathbf{Y}_{\mathbf{x}} = \mathbf{y}, \dots, \mathbf{Z}_{\mathbf{w}} = \mathbf{z})$, symbolizing the joint probability that all of these conditional statements hold simultaneously. ■

For concreteness, a typical $\mathcal{L}_1$ sentence might be $P(X = 1, Y = 1) = P(X = 1)P(Y = 1)$. The $\mathcal{L}_1$ conjunction over all such combinations

$$\begin{aligned} &P(X=1,Y=1)=P(X=1)P(Y=1)\wedge P(X=1,Y=0)=P(X=1)P(Y=0)\\ &\quad\wedge P(X=0,Y=1)=P(X=0)P(Y=1)\wedge P(X=0,Y=0)=P(X=0)P(Y=0) \end{aligned} \tag{27.13}$$

would express that X and Y are probabilistically independent if X and Y are binary variables. Of course, we would ordinarily write this simply as $P(X, Y) = P(X)P(Y)$.

In $\mathcal{L}_2$ we have sentences like $P(Y_{X=1} = 1) = 3/4$, which intuitively expresses that the probability of Y taking on value 1 were X to take on value 1 is 3/4.[17] As before, we could also write this as $P(Y = 1 \mid do(X = 1)) = 3/4$. Finally, $\mathcal{L}_3$ allows

[17] These “conditional” expressions such as $Y_{X=1} = 1$ are familiar from the literature in conditional logic. In David Lewis’s early work on counterfactual conditionals, $Y_{X=1} = 1$ would have been written $X = 1 \;\Box\!\!\rightarrow\; Y = 1$ (see Lewis [1973]). More recently, some authors have used notation from dynamic logic, $[X = 1]Y = 1$, with the same interpretation over SCMs (see, e.g., Halpern [2000]). For more discussion on the connection between the present SCM-based interpretation

statements about joint probabilities over conditional terms with possibly inconsistent subscripts (also known as antecedents in logic). For instance, $P(y_x, y'_{x'}) \geq P(y \,|\, x) - P(y \,|\, x')$ is a statement expressing a lower bound on the PNS. [18]

Definition 27.7 gives the formal structure (*syntax*) of $\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3$, but not their interpretation or meaning (*semantics*). In fact, we have already specified their meaning in SCMs via Definitions 27.2, 27.5, and 27.6. Specifically, let Ω denote the set of all SCMs over endogenous variables $\mathbf{V}$. Then each $\mathcal{M} \in \Omega$ assigns a real number to $P(\alpha)$ for all α at each layer, namely the value $P^{\mathcal{M}}(\alpha) \in [0, 1]$. Given such numbers, arithmetic and logic suffice to finish evaluating these languages. Thus, in each SCM $\mathcal{M}$, every sentence of our languages, such as Equation (27.13), comes out true or false.[19] At this stage, we are ready to formally define the PCH:

Definition 27.8 **Pearl Causal Hierarchy (PCH)**

Let $\mathcal{M}^*$ be a fully specified SCM. The collection of observational, interventional, and counterfactual distributions induced by $\mathcal{M}^*$, as delineated by languages $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_3$ (syntax) and following Definitions 27.2, 27.5, and 27.6 (semantics), is called the Pearl Causal Hierarchy. ■

In summary, as soon as we have an SCM, the PCH is thereby well defined, in the sense that this SCM provides valuations for any conceivable quantity in these languages $\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3$ (of associations, interventions, and counterfactuals, respectively). It therefore makes sense to ask about properties of the hierarchy for any given SCM, as well as for the class Ω of all SCMs. One substantive question is whether the PCH can be shown strict.

If we take $\mathcal{L}_1$ terms to involve a tacit empty intervention, that is, that $P(\mathbf{y})$ means $P(\mathbf{y}_{\varnothing})$, then the formal syntax of this series of languages clearly forms a strict hierarchy $\mathcal{L}_1 \subsetneq \mathcal{L}_2 \subsetneq \mathcal{L}_3$: there are patently $\mathcal{L}_2$ terms that do not appear in $\mathcal{L}_1$ (e.g.,

and Lewis's "system-of-spheres" interpretation, we refer readers to Pearl [2000, sections 7.4.1–7.4.3] and Briggs [2012], Halpern [2013], and Zhang [2013]. A third interpretation is over (probabilistic) "simulation" programs, which under suitable conditions are equivalent to SCMs—see Ibeling and Icard [2018, 2019, 2020].

[18] For details of this bound and the assumptions guaranteeing it, see Pearl [2000, theorem 9.2.10]. Formally speaking, statements such as this one involving conditional probabilities are shorthand for polynomial inequalities; in this case the polynomial inequality is $P(y_x, y'_{x'})P(x)P(x') + P(x', y)P(x) \geq P(x, y)P(x')$.

[19] Building on the classic axiomatization for (finite) *deterministic* SCMs [Galles and Pearl 1998, Halpern 2000], the probabilistic logical languages $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_3$ were axiomatized over probabilistic SCMs in Ibeling and Icard [2020]. The work presented in this chapter—including Definition 27.8 and Theorem 27.1 (below)—can be cast in axiomatic terms, although these results do not depend in any direct way on questions of axiomatization.

$P(y_x)$), and $\mathcal{L}_3$ terms that do not appear in $\mathcal{L}_2$ (e.g., $P(y_x, y'_{x'})$). One has the impression that each layer of the Pearl hierarchy is somehow richer or more expressive than those below it, capable of encoding information about an underlying ground truth that surpasses what lower layers can possibly express. Is this an illusion, the mere appearance of complexity, or are the concepts expressed by the layers in some way fundamentally distinct?[20] The sense of strictness that we would like to understand concerns the fundamental issue of logical *expressiveness*. If each language did not expressively exceed its predecessors, then in some sense our talk of causation and imagination would be no more than mere figures of speech, being fully reducible to lower layers.

What would it mean for the layers of the hierarchy *not* to be distinct? Toward clarifying this, let us call the set of all layer i ($\mathcal{L}_i$) statements that come out true according to some $\mathcal{M} \in \Omega$ the $\mathcal{L}_i$*-theory of* $\mathcal{M}$. We shall write $\mathcal{M} \sim_i \mathcal{M}'$ for $\mathcal{M}, \mathcal{M}' \in \Omega$ to mean that their $\mathcal{L}_i$-theories coincide, that is, that $\mathcal{M}, \mathcal{M}'$ agree on all layer i statements. Intuitively, $\mathcal{M} \sim_i \mathcal{M}'$ says that $\mathcal{M}$ and $\mathcal{M}'$ are indistinguishable given knowledge only of $\mathcal{L}_i$.

For the remainder of this section assume that the true data-generating process $\mathcal{M}^*$ is fixed. Suppose we had that $\mathcal{M}^* \sim_2 \mathcal{M}$ implies $\mathcal{M}^* \sim_3 \mathcal{M}$ for any other SCM $\mathcal{M} \in \Omega$; that is, any SCM $\mathcal{M}$ which agrees with $\mathcal{M}^*$ on all $\mathcal{L}_2$ valuations also agrees on all of the $\mathcal{L}_3$ valuations.[21] This would mean that the collection of $\mathcal{L}_2$ facts *fully determines* all of the $\mathcal{L}_3$ facts. More colloquially, if this happens, it means that we can answer any $\mathcal{L}_3$ question—including any counterfactual question, for example, the exact value of $P(y_x \,|\, y'_{x'})$—merely from $\mathcal{L}_2$ information. For instance, simply construct any SCM $\mathcal{M}$ with the right $\mathcal{L}_2$ valuation (i.e., such that $\mathcal{M} \sim_2 \mathcal{M}^*$) and read off the $\mathcal{L}_3$ facts from $\mathcal{M}$.[22] In this case it would not matter that $\mathcal{M}$ is not the true data-generating process, as any differences would not be visible even at $\mathcal{L}_3$. This can

[20] As a rough analogy, consider the ordinary concepts of "cardinality of the integers," "cardinality of the rational numbers," and "cardinality of the real numbers." One's first intuition may be that these are three distinct notions, and moreover that they form a kind of hierarchy: there are *strictly more* rational numbers than integers, and strictly more real numbers than rational numbers. Of course, in this instance the intuition can be vindicated in the second case but dismissed as an illusion in the first. (See, e.g., Cantorian arguments from any basic textbook in logic or CS.)

[21] For readers familiar with causal inference, this can be seen as a generalization of the notion of identifiability (e.g., see Pearl [2000, definition 3.2.3]), where P represents all quantities in layer i, Q all quantities in layer j, and the set of features F_M is left unrestricted (all in the notation of Pearl [2000]). This more relaxed notion has a long history in mathematical logic, namely, Padoa's method in the theory of definability [Beth 1956].

[22] Alternatively, given the completeness results in Ibeling and Icard [2020], one could axiomatically derive any $\mathcal{L}_3$ statement from appropriate $\mathcal{L}_2$ statements.

happen in exceptional circumstances, for instance, if the functional relationship is deterministic.

An additional motivation for understanding when layers of the PCH might collapse comes from the observation that, at least in some notable cases, adding syntactic complexity does not genuinely increase expressivity. As an example, we could extend the language $\mathcal{L}_3$ to allow more complex expressions. We discussed nested counterfactuals earlier in this chapter (Section 27.2), namely, statements such as $P(Y_{x,Z_{x'}})$, which can also be given a natural interpretation in SCMs. Such notions are of significant interest, but it can be shown that any such statement is systematically reducible to a Layer 3 statement. (See Bareinboim et al. [2020, appendix B] for details.) That is, for any statement φ involving nested counterfactual expressions, there is an $\mathcal{L}_3$ statement ψ such that φ and ψ hold in exactly the same models.[23] Such a result shows that adding nested counterfactuals, while providing a useful shorthand, would not allow us to say anything about the world above and beyond what we can say in $\mathcal{L}_3$. Does something similar happen with Layers 1, 2, and 3 themselves? How often might an $\mathcal{L}_3$-theory completely reduce to an $\mathcal{L}_2$-theory, or an $\mathcal{L}_2$-theory reduce to an $\mathcal{L}_1$-theory?

In light of the foregoing, we can say exactly what it means for the PCH to collapse in a given SCM $\mathcal{M}^*$. Note that the quantification here is over the class of all SCMs in Ω, that is, all SCMs with the same set of endogenous (i.e., observable) variables as $\mathcal{M}^*$:

Definition 27.9 **Collapse relative to $\mathcal{M}^*$**

Layer j of the causal hierarchy *collapses* to Layer i, with $i < j$, relative to $\mathcal{M}^* \in \Omega$ if $\mathcal{M}^* \sim_i \mathcal{M}$ implies that $\mathcal{M}^* \sim_j \mathcal{M}$ for all $\mathcal{M} \in \Omega$.[24] ■

The significance of the possibility of collapse cannot be overstated. To the extent that Layer 2 collapses to Layer 1, this would imply that we can draw all possible causal conclusions from mere correlations. Likewise, if Layer 3 collapses to Layer 2, this means that we could make statements about any counterfactual merely by conducting controlled experiments.

Our main result can then be stated (first, informally) as:

Theorem 27.1 **Causal Hierarchy Theorem (CHT), informal version**

The PCH almost never collapses. That is, for almost any SCM, the layers of the hierarchy remain distinct. ■

[23] In logic, we would say that nested counterfactuals are thus *definable* in $\mathcal{L}_3$ (see, e.g., Beth [1956]).

[24] Equivalently, there does not exist $\mathcal{M} \in \Omega$ such that $\mathcal{M}^* \sim_i \mathcal{M}$ but $\mathcal{M}^* \not\sim_j \mathcal{M}$. In other words, every layer j query can be answered with suitable layer i data.

What does *almost-never* mean? Here is an analogy. Suppose (fully specified) SCMs are drawn at random from Ω. Then, the probability that we draw an SCM relative to which PCH collapses is 0. This holds regardless of the distribution on SCMs, so long as it is smooth.

The CHT says that there will typically be causal questions that one cannot answer with knowledge and/or data restricted to a lower layer.[25] This can be seen as the formal grounding for the intuition behind the PCH discussed in Pearl and Mackenzie [2018, chapter 27]:

Corollary 27.1 To answer questions at Layer i, one needs knowledge at Layer i or higher.

With this intuitive understanding of the CHT, we now state the formal version and offer an outline of the main arguments used in the proof. In order to state the theorem, note that $\sim_3$ is an equivalence relation on Ω, inducing $\mathcal{L}_3$-equivalence classes of SCMs. Under a suitable encoding, this space of equivalence classes can be seen as a convex subset of $[0, 1]^K$, for $K \in \mathbb{N}$. This means we can put a natural (uniform) *measure* on the space of (equivalence classes) of SCMs. The theorem then states (for the complete proof and further details, we refer readers to Bareinboim et al. [2020, appendix A]):

Theorem 27.1 **CHT, formal version**

With respect to the Lebesgue measure over (a suitable encoding of $\mathcal{L}_3$ -equivalence classes of) SCMs, the subset in which any PCH collapse occurs is measure zero. ■

It bears emphasis that the CHT is a theory-neutral result in the sense that it makes only minimal assumptions and only presupposes the existence of a temporal ordering of the structural mechanisms—an assumption made to obtain unique valuations via Definitions 27.2, 27.5, and 27.6.

In the remainder of this section, we would like to discuss the basic idea behind the CHT proof. There are essentially two parts to the argument: one showing that $\mathcal{L}_2$ almost never collapses to $\mathcal{L}_1$, and the second showing that $\mathcal{L}_3$ almost never collapses to $\mathcal{L}_2$. In both parts it suffices to identify some simple property of SCMs that we can show is *typical*, and moreover sufficient to ensure non-collapse.

In fact, Layer 2 never collapses to Layer 1: for any SCM $\mathcal{M}^*$ there is always another SCM $\mathcal{M}$ with the same $\mathcal{L}_1$-theory but a different $\mathcal{L}_2$-theory. In case there

[25] The investigation of the next section will be on conditions that could allow causal inferences from lower-level data combined with graphical assumptions of the underlying SCM; see, for example, Bareinboim and Pearl [2016]. Another common thread in the literature is structural learning: adopting arguably mild assumptions of minimality (e.g., faithfulness) one can often discover fragments of the underlying causal diagram (Layer 2) from observational data (Layer 1) [Spirtes et al. 2001, Peters et al. 2017].

is any non-trivial dependence in $\mathcal{M}^*$, we can construct a second model $\mathcal{M}$ with a single exogenous variable U and all endogenous variables depending only on U, such that $\mathcal{M}^* \sim_1 \mathcal{M}$ (cf. Suppes and Zanotti [1981]). On the other hand, if $\mathcal{M}^*$ has no variable depending on any other, it is possible to induce such a dependence that, nonetheless, does not show up at Layer 1. (For full details of the argument, see Bareinboim et al. [2020, appendix A]).

The case of Layers 2 and 3 is slightly more subtle. The reason is that adding or removing arguments in the underlying functional relationships usually changes the corresponding causal effect. Here we need to show that the equations of the true $\mathcal{M}^*$ can be perturbed in a way that it does not affect any $\mathcal{L}_2$ facts but does change some joint probabilities over combinations of potential responses. It turns out there are many ways to accomplish this goal; however, for the CHT we need a systematic method. One possibility—again, informally speaking—is to take two exogenous variable settings that witness two different values for some potential response, and swap these values with some sufficiently small probability (see Bareinboim et al. [2020, appendix A]). For this to work, essentially all we need is for there to be at least some non-trivial probabilistic relationship between variables. This property is quite obviously typical of SCMs. We illustrate this method with our running Example 27.2 (Example 27.7 below).

Turning now to these examples, we start with a variation of a classic construction presented by Pearl himself [Pearl 2000, section 1.4.4]. The example has been used to demonstrate the inadequacy of (causal) Bayesian networks (discussed further in the next section) for encoding counterfactual information. Here we use it to illustrate a more abstract lesson, namely, that knowing the values of higher-layer expressions generically requires knowing progressively more about the underlying SCM (Corollary 27.1).

Example 27.6 Let $\mathcal{M}^* = \langle \mathbf{U} = \{U_1, U_2\}, \mathbf{V} = \{X, Y\}, \mathcal{F}^*, P(U)\rangle$, where

$$\mathcal{F}^* = \begin{cases} X & \leftarrow U_1 \\ Y & \leftarrow U_2 \end{cases}. \tag{27.14}$$

and U_1, U_2 are binary with $P(U_1 = 1) = P(U_2 = 1) = 1/2$. Let the variable X represent whether the patient received treatment and Y whether they recovered. Evidently, $P^{\mathcal{M}^*}(x, y) = 1/4$ for all values of X, Y. In particular X, Y are independent. Now, suppose that we just observed samples from $P^{\mathcal{M}^*}$ and were confident, statistically speaking, that X, Y are probabilistically independent. Would we be justified in concluding that X has no causal effect on Y? If the actual mechanism happened to be $\mathcal{M}^*$, then this would certainly be the case. However, this Layer 1 data is equally consistent with other SCMs in which Y depends strongly on X. Let $\mathcal{M}$ be just like $\mathcal{M}^*$, except with mechanisms:

$$\mathcal{F} = \begin{cases} X & \leftarrow \mathbb{1}_{U_1=U_2} \\ Y & \leftarrow U_1 + \mathbb{1}_{X=1,U_1=0,U_2=1} \end{cases} . \tag{27.15}$$

Then $P^{\mathcal{M}^*}(X,Y) = P^{\mathcal{M}}(X,Y)$, yet $P^{\mathcal{M}^*}(Y=1\,|\,do(X=1)) = 1/2$ as X does not affect Y in $\mathcal{M}^*$, while $P^{\mathcal{M}}(Y=1\,|\,do(X=1)) = 3/4$. If $\mathcal{M}$ were the actual mechanisms, assigning the treatment would actually improve the chance of survival. Thus, just as one cannot infer causation from correlation, one cannot always expect to infer correlation from causation.

Having internalized this lesson that correlation and causation are distinct, one might perform a randomized controlled trial and discover that all causal effects in this setting trivialize; in particular, $P(Y\,|\,do(X)) = P(Y)$—the treatment does not affect the chance of survival at all. Suppose we observe patient S, who took the treatment and died. We might well like to know whether S's death occurred *because of* the treatment, *in spite of* the treatment, or *regardless of* the treatment. This is a quintessentially counterfactual question: given that S took the treatment and died, what is the probability that S *would have* survived had they not been treated? We write this as $P(Y_{X=0} = 1\,|\,X=1, Y=0)$, as discussed in Example 27.4. Can we infer anything about this expression from Layer 2 information (in this case, that all causal effects trivialize)? We cannot, as shown by other variations of $\mathcal{M}^*$, say $\mathcal{M}'$ such that

$$\mathcal{F}' = \begin{cases} X & \leftarrow U_1 \\ Y & \leftarrow XU_2 + (1-X)(1-U_2) \end{cases} . \tag{27.16}$$

Like $\mathcal{M}$, this model reveals a dependence of Y on X. However, this is not at all visible at Layer 1 or at Layer 2; all causal effects trivialize in $\mathcal{M}'$ as well. The dependence only becomes visible at Layer 3. In $\mathcal{M}^*$, we have $P^{\mathcal{M}^*}(Y_{X=0} = 1\,|\,X=1, Y=0) = 0$, whereas in $\mathcal{M}'$ we have the exact opposite pattern, $P^{\mathcal{M}'}(Y_{X=0} = 1\,|\,X=1, Y=0) = 1$. These two models thus make diametrically opposed predictions about whether S *would have* survived had they not taken the treatment. In other words, the best *explanation* for S's death may be completely different depending on whether the world is like $\mathcal{M}^*$ or $\mathcal{M}'$. In $\mathcal{M}^*$, S would have died anyway, while in $\mathcal{M}'$, S would actually have survived, if only they had not been given the treatment. Needless to say, such matters can be of fundamental importance for critical practical questions, such as determining who or what is to blame for S's death. ■

The CHT tells us that the failure of collapse witnessed in Example 27.6 is typical. However, it is worth seeing further examples to appreciate the many ways we can take an SCM $\mathcal{M}^*$ and find an alternative SCM $\mathcal{M}$ that agrees at all lower layers but disagrees at higher layers.

We discuss two quite different strategies in the next example. To show that Layer 2 does not collapse to Layer 1, we actually *eliminate* the functional dependence of one variable on another—all probabilistic dependence patterns are due to common causes. More interestingly, we employ a very general method to show that Layer 3 does not collapse to Layer 2, whose efficacy is proven systematically in Bareinboim et al. [2020, lemma 2].

Example 27.7 **Example 27.2 continued**

For the SCM $\mathcal{M}^* = \mathcal{M}^2$ of Example 27.2, consider another model $\mathcal{M}$ with the equation for Y replaced by a new equation $Y \leftarrow \mathbb{1}_{\{U_r=1,U_x=1,U_z=1\}} + \mathbb{1}_{\{U_r=1,U_x=0,U_z=0\}} + \mathbb{1}_{\{U_r=1,U_x=0,U_y=1,U_z=1\}} + \mathbb{1}_{\{U_r=1,U_x=1,U_y=1,U_z=0\}} + \mathbb{1}_{\{U_r=1,U_x=1,U_y=0\}}$, and everything else unchanged. It is then easy to check that $\mathcal{M}^* \sim_1 \mathcal{M}$. However, Y now no longer shows a functional dependence on X: the probabilistic dependence of Y on X is due to the common causes U_x, U_z, U_r. While in Example 27.4 we saw that $P^{\mathcal{M}^*}(Y \,|\, X) \neq P^{\mathcal{M}^*}(Y \,|\, do(X))$, here we have $P^{\mathcal{M}}(Y \,|\, X) = P^{\mathcal{M}}(Y \,|\, do(X))$. In other words, even though X does exert a causal influence on Y (assuming $\mathcal{M}^*$ is the true data-generating process), we would not be able to infer this from observational data alone.

To show that Layer 3 does not collapse to Layer 2, consider a third model $\mathcal{M}'$, in which X, Y, Z all share one exogenous parent U, with $\text{Val}(U) = \{0,1\}^4 \cup \{u_1^*, u_2^*\}$. The probability of a quadruple $\langle u_r, u_z, u_x, u_y \rangle$ in this model is simply given by the product from model $\mathcal{M}^*$—$P(U_r = u_r) \cdot P(U_z = u_z) \cdot P(U_x = u_x) \cdot P(U_y = u_y)$—with one exception: for the two quadruples, $\langle 1,1,1,0 \rangle$ and $\langle 1,1,0,0 \rangle$, we subtract $\varepsilon = .005$ from these probabilities, and redistribute the remaining mass so that u_1^* and u_2^* each receive probability ε. This produces a proper distribution $P'(U)$. We will continue to write, for example, $U_r = u$ simply to mean that $U \neq u_1^*, u_2^*$ and the first coordinate of U is u, and similarly for U_z, U_x, U_y. The mechanisms are now:

$$\mathcal{F}' = \begin{cases} Z \leftarrow \mathbb{1}_{\{U_r=1,U_z=1\}} + \mathbb{1}_{U\in\{u_1^*,u_2^*\}} \\ X \leftarrow \mathbb{1}_{\{Z=1,U_x=1\}} + \mathbb{1}_{\{Z=0,U_x=0\}} + \mathbb{1}_{U=u_2^*} \\ Y \leftarrow \mathbb{1}_{\{X=1,U_r=1\}} + \mathbb{1}_{\{X=0,U_r=1,U_y=1\}} + \mathbb{1}_{\{X=0,U_r=0,U_y=0\}} + \mathbb{1}_{\{X=1,U\in\{u_1^*,u_2^*\}\}} \end{cases} . \tag{27.17}$$

To check that the joint distributions $P^{\mathcal{M}^*}(X,Y,Z)$ and $P^{\mathcal{M}'}(X,Y,Z)$ are the same, note that the two models coincide at all exogenous settings with the exception of the two quadruples $\langle 1,1,1,0 \rangle$ and $\langle 1,1,0,0 \rangle$. In the first we have $Z = X = Y = 1$, and the ε-loss in probability for this possibility is corrected by the fact that $X(u_2^*) = Y(u_2^*) = Z(u_2^*) = 1$ and $P'(u_2^*) = \varepsilon$. Similarly for $\langle 1,1,0,0 \rangle$ and the state $Z = 1, X = Y = 0$, which results when $U = u_1^*$. To show that $\mathcal{M}^* \sim_2 \mathcal{M}'$ is also straightforward.

However, consider the $\mathcal{L}_3$ expression $Y_{Z=1} = 1, Y_{Z=0} = 1$, which says that the patient would survive no matter whether hypertension was induced or prevented. For both exogenous settings $\langle 1, 1, 1, 0\rangle$ and $\langle 1, 1, 0, 0\rangle$, this expression is false, yet in setting u_2^* the expression is true. Hence, $P^{\mathcal{M}'}(Y_{Z=1} = 1, Y_{Z=0} = 1) = P^{\mathcal{M}^*}(Y_{Z=1} = 1, Y_{Z=0} = 1) + \varepsilon$. ■

While collapse of the layers is possible if $\mathcal{M}^*$ is exceptional, the CHT shows that this is the exception indeed. Typical cases are similar to Examples 27.6 and 27.7, each showing a different way of perturbing an SCM to obtain a second SCM revealing non-collapse. In fact, a typical data-generating process $\mathcal{M}^*$ encodes rich information at all three layers, and even small changes to the mechanisms in $\mathcal{M}^*$ can have substantial impact on quantities across the hierarchy. Critically, such differences will often be visible only at higher layers in the PCH.

The lesson learned from the CHT is clear—as the layers of PCH come apart in the generic case and one cannot make inferences at one layer given knowledge at lower layers (e.g., using observational data to make interventional claims), some additional assumptions are logically necessary if one wants in general to do *causal inference*.

27.4 Pearl Hierarchy—A Graphical Perspective

All conceivable quantities from any layer of the PCH—associational, interventional, and counterfactual—are immediately computable once the fully specified SCM is known. Unfortunately, in most practical settings, it's usually hard to determine the structural model at this level of precision, and the CHT severely curtails the ability to "climb up" the PCH via lower-level data. Learning about cause-and-effect relationships is arguably one of the main goals found throughout the sciences. After all, how could causal inferences be performed?

The recognition that there are mechanisms underlying the phenomena of interest, but that we usually cannot determine them precisely, gives rise to the discipline of *causal inference* [Pearl 2000]. Virtually every approach to causal inference works under the stringent condition that only partial knowledge of the underlying SCM is available. One pervasive task is to determine the effect of an intervention—what would happen with Y were X to be intervened on and set to x, $P(Y \mid do(X = x))$—from observational data, $P(X, Y)$. This constitutes a cross-layer inference where the goal is to use data from layer $\mathcal{L}_1$ to try to make an inference about an $\mathcal{L}_2$ quantity, given a partial specification of the underlying SCM (see Figure 27.3 [a–d]).

In this section, we investigate the question of what type of causal knowledge could be (1) intuitively meaningful, (2) possibly available, and (3) powerful enough to encode constraints that would allow cross-layer inferences, *as if* the SCM were

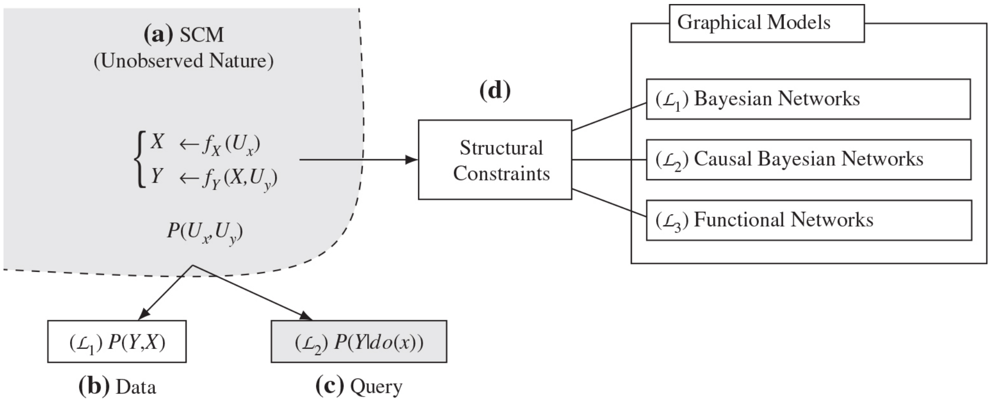

(a) SCM
(Unobserved Nature)
$X \leftarrow f_X(U_x)$
$Y \leftarrow f_Y(X, U_y)$
$P(U_x, U_y)$
(d)
Structural Constraints
Graphical Models
($\mathcal{L}_1$) Bayesian Networks
($\mathcal{L}_2$) Causal Bayesian Networks
($\mathcal{L}_3$) Functional Networks
($\mathcal{L}_1$) $P(Y, X)$
(b) Data
($\mathcal{L}_2$) $P(Y|do(x))$
(c) Query

Figure 27.3 Example of Prototypical Causal Inference—on top the SCM itself, representing the unobserved collection of mechanisms and corresponding uncertainty (a); at the bottom, the different probability distributions entailed by the model (b, c); on the right side, the graphical model representing the specific constraints of the SCM (d).

itself available. A key observation useful to answer this question is that each SCM imprints specific "marks" on the distributions it generates, depicted generically in the schema in Figure 27.3(d) as *structural constraints*.

One first attempt to solve this task could be to leverage $\mathcal{L}_1$-constraints, those imprinted on the observed $\mathcal{L}_1$ data by the unknown SCM, to make inferences about the target $\mathcal{L}_2$-quantity. This is especially appealing considering that $\mathcal{L}_1$ data is often readily available. The signature type of constraint for $\mathcal{L}_1$ distributions is known as *conditional independence*, and *Bayesian Networks* (BNs) are among the most prominent formal models used to encode this type of knowledge. The example below shows that $\mathcal{L}_1$ constraints (and BNs) alone are insufficient to support causal reasoning in general.

Example 27.8 Let $\mathcal{M}^1$ and $\mathcal{M}^2$ be two SCMs such that $\mathbf{V} = \{X, Z, Y\}, \mathbf{U} = \{U_x, U_z, U_y\}$, and the structural mechanisms are, respectively,

$$\mathcal{F}_1 = \begin{cases} X & \leftarrow U_x \\ Z & \leftarrow X \oplus U_z \\ Y & \leftarrow Z \oplus U_y \end{cases}, \qquad \mathcal{F}_2 = \begin{cases} X & \leftarrow Z \oplus U_x \\ Z & \leftarrow Y \oplus U_z \\ Y & \leftarrow U_y \end{cases}, \tag{27.18}$$

where $\oplus$ is the logical *xor* operator. Further, the distributions of the exogenous variables are $P^1(U_x = 1) = P^2(U_y = 1) = 1/2$, $P^1(U_z = 1) = P^2(U_x = 1) = a$, and $P^1(U_y = 1) = P^2(U_z = 1) = b$, for some $a, b \in (0, 1)$. It can immediately be seen (via Definition 27.2 and Equation (27.3)) that both models generate the same observational distribution,

$$\begin{aligned}
P^{1,2}(X=0,Z=0,Y=0) &= P^{1,2}(X=1,Z=1,Y=1) = (1-a)(1-b)/2,\\
P^{1,2}(X=0,Z=0,Y=1) &= P^{1,2}(X=1,Z=1,Y=0) = (1-a)b/2,\\
P^{1,2}(X=0,Z=1,Y=1) &= P^{1,2}(X=1,Z=0,Y=0) = a(1-b)/2,\\
P^{1,2}(X=0,Z=1,Y=0) &= P^{1,2}(X=1,Z=0,Y=1) = ab/2. \qquad (27.19)
\end{aligned}$$

We further compute the effect of the intervention $do(x)$ (via Definition 27.5 and Equation 27.7),

$$P^1(Y=1\,|\,do(X=1)) = ab + (1-a)(1-b), \quad P^2(Y=1\,|\,do(X=1)) = 1/2, \qquad (27.20)$$

which are different for most values a, b. The models $\mathcal{M}^1$ and $\mathcal{M}^2$ naturally induce BNs $\mathcal{G}^1$ and $\mathcal{G}^2$, respectively; see Figure 27.4(a) and (b).[26] In terms of $\mathcal{L}_1$-constraints, $\mathcal{G}^1$ and $\mathcal{G}^2$ both imply that X is independent of Y given Z (for short, $X \perp Y\,|\,Z$) and nothing more.[27] This means that $\mathcal{G}^1$ and $\mathcal{G}^2$ are equivalent through the lens of $\mathcal{L}_1$, while the original $\mathcal{M}^1$ and $\mathcal{M}^2$ generate different answers to $\mathcal{L}_2$ queries, as shown in Equation (27.20). ■

The main takeaway from the example is that from only the distribution $P(\mathbf{V})$ and the qualitative (conditional independence) constraints implied by it, it is impossible to tell whether the underlying reality corresponds to $\mathcal{M}^1$, $\mathcal{M}^2$, or any other SCM inducing the same $P(\mathbf{V})$, while each such model could entail a different causal effect. This suggests that, in general, causal inference cannot be carried out with mere $\mathcal{L}_1$ objects—the observational distribution, its constraints, and corresponding models (BNs). This result can be seen as a graphical instantiation of Corollary 27.1 and is schematically summarized in Figure 27.4.

27.4.1 Causal Inference via $\mathcal{L}_2$-constraints—Markovian Causal Bayesian Networks

Having witnessed the impossibility of performing causal inference from $\mathcal{L}_1$ constraints, we come back to the original question—what kind of structural constraints (Figure 27.3(d)) imprinted by the underlying SCM could license causal

[26] This construction follows from the order in which the functions are determined in the SCM, systematized in Definition 24 [Bareinboim et al. 2020, appendix C]. This procedure is guaranteed to produce BNs that are compatible with the independence constraints implied by the SCM in $\mathcal{L}_1$ [Bareinboim et al. 2020, theorem 8, appendix C].

[27] We refer readers to Bareinboim et al. [2020, appendix C], for more details on a criterion called *d-separation* [Pearl 1988], which is the tool used for reading these constraints off from the graphical model.

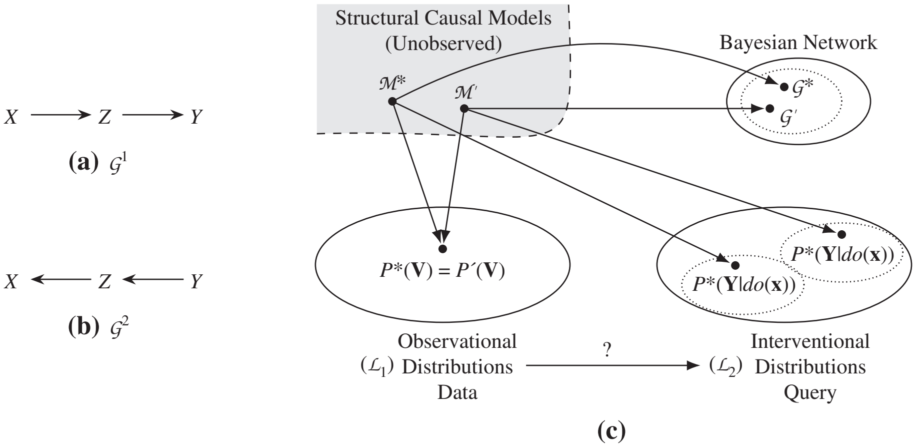


Figure 27.4 Two causal diagrams encoding knowledge about the causal mechanisms governing three observable variables X, Z, and Y. In (a) X is an argument to f_Z, and Z an argument to f_Y. In (b) the opposite is true. In (c), schema representing the impossibility of identifying causal queries from $\mathcal{L}_1$ data, constraints, and graphical models.

inferences? To answer this question, it is instructive to compare more closely the effect of an intervention $X = 1$ in the two SCMs from Example 27.8. First, note that the function f_Y does not depend on X in the submodel $\mathcal{M}^2_{X=1}$ (constructed following Definition 27.3); so, probabilistically, Y will not depend on X. This implies the following relationship between distributions,

$$P^2(Y = 1 \,|\, do(X = 1)) = P(Y = 1), \tag{27.21}$$

In contrast, note that (i) f_Y does take into account the value of X in $\mathcal{M}^1_{X=1}$, and (ii) Y responds (or varies) in the same way when X takes a particular value, be it naturally (as in $\mathcal{M}^1$) or due to an intervention (as in $\mathcal{M}^1_{X=1}$). These facts can be formally written as

$$P^1(Y = 1 \,|\, do(X = 1)) = P(Y = 1 \,|\, X = 1). \tag{27.22}$$

The exact computation of Equations (27.21) and (27.22) follows immediately from Definitions 27.2 and 27.5. Remarkably, the intuition behind these equalities does not arise from the particular form of the underlying functions, the exogenous variables, or their distribution, but from structural properties of the model. In particular, they are determined by qualitative functional dependences among the variables: what variable is an argument to the function of the other.

Technically, these equalities can be seen as constraints (not conditional independences) and can be pieced together and given a graphical interpretation.

Consider again Equation (27.21) as an example, which says that variable X does not have an effect on Y (doing X does not change the marginal distribution of Y), which graphically would entail that X is not an ancestor of Y in $\mathcal{G}^2$. While true in $\mathcal{M}^2$, it certainly does not hold in $\mathcal{M}^1$, nor, consequently, in $\mathcal{G}^1$. Even though $\mathcal{G}^1$ and $\mathcal{G}^2$ are graphically equivalent with respect to $\mathcal{L}_1$, and could be used interchangeably for probabilistic reasoning, they are, interventionally speaking, very distinct objects.

These constraints encode one of the fundamental intuitions we have about causality, namely, the asymmetry that a cause may change its effect but not the other way around. Our goal henceforth will be to systematically incorporate these constraints into a new family of graphical models with arrows carrying causal meaning and supporting $\mathcal{L}_2$-types of inferences. First, we introduce a procedure that returns a new graphical model following the intuition behind the constraints discussed so far, and then show how it relates to the collection of interventional distributions ($\mathcal{L}_2$-valuations) entailed by the SCM.

Definition 27.10 **Causal Diagram (Markovian Models)**

Consider a Markovian SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$. Then, $\mathcal{G}$ is said to be a *causal diagram* (of $\mathcal{M}$) if constructed as follows:

1. add a vertex for every endogenous variable in the set $\mathbf{V}$,
2. add an edge ($V_j \longrightarrow V_i$) for every $V_i \in \mathbf{V}$ if V_j appears as an argument of $f_i \in \mathcal{F}$.

■

The procedure encapsulated in Definition 27.10 is central to the elicitation of the knowledge necessary to perform causal inference (Figure 27.3(d)). Intuitively, $\mathcal{G}$ has an arrow from A to B ($A \longrightarrow B$) if B "listens" to the value of A; functionally, A appears as an argument of the mechanism of B. The importance of this notion has been emphasized in the literature by Pearl: "This listening metaphor encapsulates the entire knowledge that a causal network conveys; the rest can be derived, sometimes by leveraging data" [Pearl and Mackenzie 2018, p. 129]. This construction produces a coarsening of the underlying SCM such that the arguments of the functions are preserved while their particular forms are discarded.[28]

The assumptions that the causal diagram encodes about the SCM impose constraints not only over the $\mathcal{L}_1$-distribution P but also over all the interventional ($\mathcal{L}_2$) distributions as encapsulated in the following definition [Bareinboim et al. 2012].

Definition 27.11 **Causal Bayesian Network (CBN)-Markovian**

Let $\mathbf{P}_*$ be the collection of all interventional distributions $P(\mathbf{V} \mid do(\mathbf{x}))$, $\mathbf{X} \subseteq \mathbf{V}$,

[28] Given the lack of constraints over the form and shape of the underlying functions and distribution of the exogenous variables, these models are usually called *non-parametric* in the causal inference literature.

$\mathbf{x} \in \mathrm{Val}(\mathbf{X})$, including the null intervention, $P(\mathbf{V})$, where $\mathbf{V}$ is the set of observed variables. A directed acyclic graph $\mathcal{G}$ is called a CBN for $\mathbf{P}_*$ if for all $\mathbf{X} \subseteq \mathbf{V}$, the following conditions hold:

(i) [Markovian] $P(\mathbf{V} \mid do(\mathbf{x}))$ is Markov relative to $\mathcal{G}$.

(ii) [Missing-link] For every $V_i \in \mathbf{V}$, $V_i \notin \mathbf{X}$ such that there is no arrow from $\mathbf{X}$ to V_i in $\mathcal{G}$:

$$P(v_i \mid do(pa_i), do(\mathbf{x})) = P(v_i \mid \mathrm{do}(pa_i)). \tag{27.23}$$

(iii) [Parents do/see] For every $V_i \in \mathbf{V}$, $V_i \notin \mathbf{X}$:

$$P(v_i \mid do(\mathbf{x}), do(pa_i)) = P(v_i \mid do(\mathbf{x}), pa_i). \tag{27.24}$$

■

The first condition requires the graph to be *Markov relative*[29] to every interventional distribution $P(\mathbf{V} \mid do(\mathbf{X} = \mathbf{x}))$, which holds if every variable is independent of its non-descendants given its parents.[30] The second condition, missing-link, encapsulates the type of constraint exemplified by Equation (27.21): after fixing the parents of a variable by intervention, the corresponding function should be insensitive to any other intervention elsewhere in the system. In other words, the parents Pa_i *interventionally* shield V_i from interventions ($do(\mathbf{X})$) on other variables. Finally, the third condition, parents do/see, encodes the intuition behind Equation (27.22): whether the function f_i takes the value of its arguments following an intervention ($do(Pa_i = pa_i)$) or by observation (conditioned on $Pa_i = pa_i$), the same behavior for V_i is observed.

Some observations follow immediately from these conditions. First, and perhaps not surprisingly, a CBN encodes stronger assumptions about the world than a BN. In fact, all the content of a BN is encapsulated in condition (i) of a CBN (Definition 27.11) with respect to the observational (null intervention) distribution $P(\mathbf{V})$ ($\mathcal{L}_1$). A CBN encodes additional constraints on interventional distributions ($\mathcal{L}_2$) beyond conditional independence, involving different interventions such as those represented in conditions (ii) and (iii).

[29] This notion is also known in the literature as *compatibility* or *i-mapness* [Pearl 1988, Koller and Friedman 2009], which is usually encoded in the decomposition of $P(\mathbf{v})$ as $\prod_i P(v_i \mid pa_i)$ in the Markovian case.

[30] In some accounts of causation, this condition is known as the *causal Markov condition* (CMC), and is usually phrased in terms of "causal" parents. We invite the reader to check that conditions (ii) and (iii) are in no way implied by (i). One could in fact see Definition 27.11 as offering a precise characterization of what CMC formally means.

Second, readers familiar with graphical models will be quick to point out that the knowledge encoded in these models is not in the presence but in the absence of the arrows; each missing arrow makes a claim about a certain type of invariance. In the context of BNs ($\mathcal{L}_1$), each missing arrow corresponds to a conditional independence, a probabilistic type of invariance.[31] On the other hand, each missing arrow in a CBN represents an $\mathcal{L}_2$-type constraint, for example, the lack of a direct effect, as encoded in Definition 27.11 through condition (ii). This new family of constraints closes a long-standing semantic gap, from a graphical model's perspective, rendering the causal interpretation of the graphical model totally unambiguous.

Before proving that this graphical model encapsulates all the probabilistic and causal constraints required for reasoning in $\mathcal{L}_2$, we show next that the $\mathcal{L}_2$-empirical content of an SCM—that is, the collection of observational and interventional distributions (Definition 27.5)—indeed matches the content of the CBN (Definition 27.10), as defined above.

Theorem 27.2 **$\mathcal{L}_2$-Connection—SCM-CBN (Markovian)**

The causal diagram $\mathcal{G}$ induced by the SCM $\mathcal{M}$ (following the constructive procedure in Definition 27.10) is a CBN for $\mathbf{P}_*^{\mathcal{M}}$—the collection of observational and experimental distributions induced by $\mathcal{M}$. ■

For the complete proof, see Bareinboim et al. [2020, appendix D]. As this result demonstrates, CBNs serve as proxies for SCMs in terms of the observed $\mathcal{L}_2$ distributions. In practice, whenever the SCM is not fully known and the collection of interventional distributions is not available, this duality suggests that a CBN can act as a basis for causal reasoning. To ground this point, we go back to our task of inferring the interventional distribution, $P(\mathbf{Y} \mid do(\mathbf{X} = \mathbf{x}))$, from a combination of the observational distribution, $P(\mathbf{V})$, and the qualitative knowledge of the SCM encoded in the causal diagram $\mathcal{G}$. A remarkable result that holds in Markovian models is that causal inference is always possible, that is, any interventional distribution is computable from $\mathcal{L}_1$-data.

Theorem 27.3 **Truncated Factorization Product (Markovian)**

Let the graphical model $\mathcal{G}$ be a CBN for the set of interventional distributions $\mathbf{P}_*$. For any $\mathbf{X} \subseteq \mathbf{V}$, the interventional ($\mathcal{L}_2$) distribution $P(\mathbf{V} \mid do(\mathbf{x}))$ is identifiable through the truncated factorization product, namely,

$$P(\mathbf{v} \mid do(\mathbf{x})) = \prod_{\{i \mid V_i \notin \mathbf{X}\}} P(v_i \mid pa_i)\Big|_{\mathbf{X}=\mathbf{x}}. \tag{27.25}$$

■

[31] One can show that there always exists a separator, in the *d-separation* sense, between nonadjacent nodes.

In other words, the interventional distribution in the LHS of Equation (27.25) can be expressed as the product given in the right-hand side (RHS) involving only $\mathcal{L}_1$-quantities, where the factors relative to the intervened variables are removed, hence the name *truncated factorization product* (see Pearl [2000, equation 1.37]).[32] Obviously, any marginal distribution of interest can be obtained by summing out the irrelevant factors, including the causal effect of X on Y.

27.4.2 Causal Inference via $\mathcal{L}_2$-constraints—Semi-Markovian Causal Bayes Networks

The treatment provided for the Markovian case turned out to be simple and elegant, yet surprisingly powerful. The causal graph is a perfect surrogate for the SCM in the sense that all $\mathcal{L}_2$ quantities (causal effects) are computable from $\mathcal{L}_1$-type of data (observational) and the constraints in $\mathcal{G}$. A "model-theoretic" way of understanding this result is that all the SCMs that induce the same causal diagram and generate the same observational distribution will also generate the same set of experimental distributions, immediately computable via the truncated product (Theorem 27.3). This is a quite remarkable result as we moved from a model based on $\mathcal{L}_1$-structural constraints (e.g., a Bayes net) such that no causal inference was permitted, to a model encoding $\mathcal{L}_2$-constraints (a causal Bayes net) such that any conceivable cross-layer inference is immediately allowed.

In light of these results, one may be tempted to surmise that causal inference is a solved problem. This could not be farther from the truth, unfortunately. The assumption that all the relevant factors about the phenomenon under investigation are measured and represented in the causal diagram (i.e., Markovianity holds) is often too stringent, and violated in most real-world scenarios. This means that the aforementioned results are usually not applicable in practice. Departing from this observation, our goal is to understand the principles that allow cross-layer inferences when the Markov condition does not hold, which entails incorporating unobserved confounders as a building block of $\mathcal{L}_2$-graphical models. We start by investigating the reasons the machinery developed so far is insufficient to accommodate such cases.

Example 27.9 **Example 27.1 revisited**

Recall the two-dice game where the endogenous variables X and Y (the sum and difference of two dice, respectively) do not functionally depend on each other, despite their strong association. One could attempt to model such a setting with

[32] The truncated formula is also known as the "manipulation theorem" [Spirtes et al. 2001] or G-computation formula [Robins 1986, p. 1423]. For further details, we refer readers to Pearl [2000, section 3.6.4].

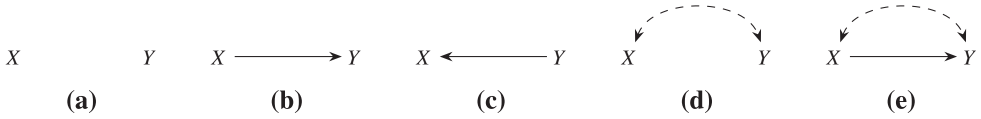


Figure 27.5 The diagram in (a) implies that neither X nor Y is an argument to the function of the other. In (b, c) one endogenous variable causes the other. In (d) there is no causal relationship yet the functions share exogenous arguments, as encoded through the bidirected arrow. In (e) both types of influence are encoded.

the graphical structure shown in Figure 27.5(a), somewhat naively, trying to avoid a directed arrow between X and Y. As previously noted, if the sum of the dice is equal to two ($X = 2$), one could, with probability one, infer that the two dice obtained the same value ($Y = 0$). The hypothesized graphical model, however, forces the two variables to be independent, which would rule out the possibility of performing such an inference.

Upon recognition of such impropriety, one could reconsider adding an arrow from X to Y (or Y to X) so as to leverage the valuable information shared across the observed variables, as shown in Figure 27.5(b). We previously learned, on the other hand, that reporting that the sum of the dice is 2 does not change their difference, formally, $P(Y \mid do(X = 2)) = P(Y)$ must hold in this setting (Equation 27.9). Obviously, this would be violated were the world to mirror this graphical structure. To witness, consider the alternative SCM $\mathcal{M}'$ where the function for X is identical and $Y \leftarrow (X - 2U_2)$. We can verify that $P(X, Y)$ is the same as in $\mathcal{M}^1$, while the causal effect of X on Y is non-zero. ■

The recognition that certain dependencies among endogenous variables cannot be *explained* by other variables inside the model (but also cannot be ignored) led Pearl to introduce a new type of arrow to account for these relationships. The new arrows are dashed and bidirected. In the example above, variables X and Y are correlated due to the existence of two common exogenous variables, $\{U_1, U_2\}$, which are arguments of both f_X and f_Y. We will usually refer to these variables as U_{xy} since, *a priori*, we will neither know, nor want to assume, their particular form, dimensionality, or distribution. This new type of arrow will allow for the probabilistic dependence between them, $(X \perp\!\!\!\perp Y)$, while being neutral with respect to their interventional invariance. That is, it would accept constraints such as $P(Y \mid do(X)) = P(Y)$ and $P(X \mid do(Y)) = P(X)$. See Figure 27.5(d) for a graphical example.

In practice, some variables may be related through both sources of variations—one exogenous, not explained by the variables in the model, and another endogenous, causally explained by the relationships between the variables in the model, as shown in Figure 27.5(e). Due to the unobserved confounder U_{xy}, the equality $P(Y \mid do(x)) = P(Y \mid x)$ will not, in general, hold. In other words, Y's distribution will be different depending on whether we observe $X = x$ or intervene and $do(X = x)$. Fundamentally, this will translate into a violation of the constraint encoded in Equation (27.22) and, more generally, in condition (iii) of the definition of CBNs (Definition 27.11).

Our goal, henceforth, will be to cope with the complexity arising due to violations of Markovianity. One particular implication of these violations is the widening of the empirical content carried by the CBN versus its underlying SCM, as shown in the next example.

Example 27.10 Consider two SCMs $\mathcal{M}^*$ and $\mathcal{M}'$ such that $\mathbf{V} = \{X, Y\}$, $\mathbf{U} = \{U_{xy}, U_y\}$, the structural mechanisms are $\mathcal{F} = \{X \leftarrow U_{xy}, Y \leftarrow (X \oplus U_y)$ if $X = U_{xy}$, δ otherwise$\}$, where $\delta = 0$ for $\mathcal{M}^*$ and $\delta = 1$ for $\mathcal{M}'$. The exogenous distributions of both models, $P^*(\mathbf{U})$ and $P'(\mathbf{U})$, are the same and given by $P(U_{xy} = 1) = 1/2$, $P(U_y = 1) = 3/4$, and they both follow the diagram shown in Figure 27.5(e). It is easy to verify that both models induce the same $P(\mathbf{V})$, while $P^*(Y = 1 \mid do(X = 1)) = 1/8 \neq 5/8 = P'(Y = 1 \mid do(X = 1))$.

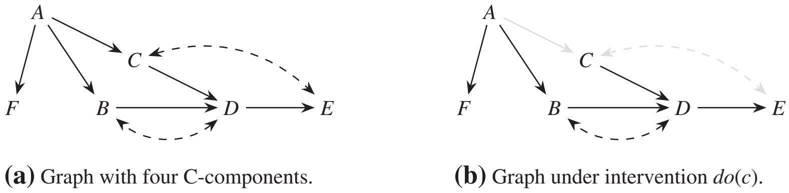
■

Remarkably, this is our first encounter with a situation in which a causal diagram—encoding all the $\mathcal{L}_2$-structural invariances of the underlying SCM $\mathcal{M}^*$—is too weak, incapable of answering the intended cross-layer inference—computing $P(Y \mid do(x))$ from the corresponding $\mathcal{L}_1$-distribution, $P(X, Y)$. There exists at least one other SCM $\mathcal{M}'$ that shares the same set of structural features, in the form of the constraints encoded in the causal diagram, but generates a different answer for the causal effect. In other words, one cannot commit and make a claim about the target effect as there are multiple, unobserved SCMs compatible with the given diagram and observational data.

Whenever the causal effect is not uniquely computable from the constraints embedded in the graphical model, we say that it is non-identifiable from $\mathcal{G}$ (to be formally defined later on). More generally, we would like to understand under what conditions an interventional distribution can be computed from the observational one, given the structural constraints encoded in the causal diagram. First, we supplement the Markovian construction of CBNs, given in Definition 27.10, to formally account for the existence of unobserved confounders.

Definition 27.12 **Causal Diagram (Semi-Markovian Models)**
Consider an SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$. Then, $\mathcal{G}$ is said to be a *causal diagram* (of $\mathcal{M}$) if constructed as follows:

(1) add a vertex for every endogenous variable in the set $\mathbf{V}$,



(a) Graph with four C-components.

(b) Graph under intervention *do*(*c*).

Figure 27.6 Causal diagram with bidirected arrows and its mutilated counterpart under $do(c)$.

(2) add an edge ($V_j \longrightarrow V_i$) for every $V_i, V_j \in \mathbf{V}$ if V_j appears as an argument of $f_i \in \mathcal{F}$.

(3) add a bidirected edge ($V_j \leftarrow\!\!-\!-\!-\!\!\rightarrow V_i$) for every $V_i, V_j \in \mathbf{V}$ if the corresponding $U_i, U_j \subset \mathbf{U}$ are correlated or the corresponding functions f_i,f_j share some $U \in \mathbf{U}$ as an argument. ■

Following this procedure, each SCM $\mathcal{M}$ induces a unique causal diagram. Furthermore, each bidirected arrow encodes unobserved confounding in $\mathcal{G}$. They indicate correlation between the unobserved parents of the endogenous variables at the endpoints of such edges.

27.4.2.1 Revisiting Locality in Semi-Markovian Models

Graphical models provide a transparent and systematic way of encoding structural constraints about the underlying SCM (Figure 27.3(d)). In practice, these constraints follow from the autonomy of the structural mechanisms [Aldrich 1989, Pearl 2000], which materializes as local relationships in the causal diagram. In Markovian models, these local constraints appear in the form of family relationships, for example, (1) each variable V_i is independent of its non-descendants given its parents Pa_i, or (2) each variable is invariant to interventions in other variables once its parents are held constant (following Definition 27.11). The local nature of these relations leads to a parsimonious factorization of the joint probability distribution, and translates into desirable sample and computational complexity properties.

On the other hand, the family relations in semi-Markovian models are less well-behaved and the boundaries of influence among the variables are usually less local. To witness, consider Figure 27.6(a), and note that, where $Pa_d = \{B, C\}$ and the remaining $NDesc_d = \{A, F\}$, $D \perp\!\!\!\perp NDesc_d \,|\, Pa_d$ does not hold as D and A are connected through the open path $D \leftarrow\!\!-\!-\!-\!\!\rightarrow B \longleftarrow A$. We introduce below a construct called *confounded component* [Tian and Pearl 2002b] to restore and help to make sense of modularity in these models.

Definition 27.13 **Confounded Component**

Let $\{\mathbf{C}_1, \mathbf{C}_2, \ldots \mathbf{C}_k\}$ be a partition over the set $\mathbf{V}$. $\mathbf{C}_i$ is said to be a confounded component (C-component) of $\mathcal{G}$ if there exists a path made of bidirected edges between V_i and V_j, for every $V_i, V_j \in \mathbf{C}_i$ in $\mathcal{G}$, and $\mathbf{C}_i$ is maximal. ■

This construct represents clusters of variables that share the same exogenous variations regardless of their directed connections. The causal diagram in Figure 27.6(a) has two bidirected edges indicating the presence of unobserved confounders affecting the pairs (B, D) and (C, E) and contains four C-components, namely, $\mathbf{C}_1 = \{A\}$, $\mathbf{C}_2 = \{B, D\}$, $\mathbf{C}_3 = \{C, E\}$, and $\mathbf{C}_4 = \{F\}$. Similarly, each causal diagram in Figure 27.5(a–c) contains two C-components, $\mathbf{C}_1 = \{X\}$ and $\mathbf{C}_2 = \{Y\}$, while each in Figure 27.5(d, e) contains one C-component, $\mathbf{C}_1 = \{X, Y\}$.

Our goal is to understand the boundaries of influence among variables in semi-Markovian models as the parents of a node no longer shield it from its non-descendants, and this condition is a basic building block in the construction of Markovian models. Consider again the graph in Figure 27.6(a) and the node E and its only parent D. If we condition on D, E will not be independent of its non-descendants in the graph. Obviously, E is automatically connected to its bidirected neighbors, so it cannot be separated from C. Further, upon conditioning on the parent D, the collider through C is opened up as D is its descendant (i.e., $E \dashleftarrow\dashrightarrow C \leftarrow A$ carries correlation given D). In this case, the ancestors and descendants of C also become correlated with E, which is now connected to every other variable in the graph (A, F, B). Further, note that by conditioning on C itself, its descendants will be independent of E but its ancestors and ancestors' descendants will still be connected. In this graph, E is connected to all other nodes upon conditioning on its observed parent D and C-component neighbor C, that is, A, B, F. Then, we also need to condition on the parents of C (i.e., A) to render its other ancestors and their descendants (i.e., F) independent of E.

Putting these observations together, for each endogenous variable V_i, we need to condition on its parents, the variables in the same C-component that precede it, and the parents of the latter so as to shield V_i from the other non-descendants in the graph. Such a maximal set is formally defined as Pa_i^+ as follows. Let $<$ be a topological order $V_1, \ldots, V_n$ of the variables $\mathbf{V}$ in $\mathcal{G}$,[33] and let $\mathcal{G}(V_i)$ be the subgraph of $\mathcal{G}$ composed only of variables in $V_1, \ldots, V_i$. Given $\mathbf{X} \subseteq \mathbf{V}$, let $Pa^1(\mathbf{X}) = \mathbf{X} \cup \{Pa(X) : X \in \mathbf{X}\}$; further, let $\mathbf{C}(V_i)$ be the C-component of V_i in $\mathcal{G}(V_i)$. Then define $Pa_i^+ = Pa^1(\{V \in \mathbf{C}(V_i) : V \leq V_i\}) \setminus \{V_i\}$. For instance, in Figure 27.6(a), $Pa_e^+ = \{D, C, A\}$ and $Pa_d^+ = \{B, C, A\}$.

[33] That is, an order on the nodes (endogenous variables) $\mathbf{V}$ such that if $V_j \longrightarrow V_i \in \mathcal{G}$, then $V_j < V_i$.

Akin to the concept of *Markov relative*, a causal diagram also imposes factorization constraints over the observational distribution in semi-Markovian CBNs, as shown next.

Definition 27.14 **Semi-Markov Relative**

A distribution P is said to be *semi-Markov relative* to a graph $\mathcal{G}$ if for any topological order $<$ of $\mathcal{G}$, P factorizes as

$$P(\mathbf{v}) = \prod_{V_i \in \mathbf{V}} P(v_i \,|\, pa_i^{+}), \tag{27.26}$$

where Pa_i^{+} is defined using $<$. ■

Not only is the joint observational distribution related to a causal graph, but so are the $\mathcal{L}_2$-distributions $P(\cdot \,|\, do(\mathbf{x}))$ under an intervention $do(\mathbf{X} = \mathbf{x})$. The corresponding graph is $\mathcal{G}_{\overline{\mathbf{X}}}$, where the incoming arrows toward $\mathbf{X}$ are cut, and the semi-Markovian factorization is

$$P_{\mathbf{x}}(\mathbf{v}) = \prod_{V_i \in \mathbf{V}} P_{\mathbf{x}}(v_i \,|\, pa_i^{\mathbf{x}+}), \tag{27.27}$$

where $Pa_i^{\mathbf{x}+}$ is constructed as $Pa_i^{\mathbf{x}+}$ but according to $\mathcal{G}_{\overline{\mathbf{X}}}$.

Example 27.11 **Factorization implied by the semi-Markov condition**

Let $P(A, B, C, D, E, F)$ be a distribution semi-Markov relative to the diagram $\mathcal{G}$ in Figure 27.6(a). One topological order of $\mathcal{G}$ is $A < B < C < D < E < F$, which implies that $P(a, b, c, d, e, f) = P(a)P(b \,|\, a)P(c \,|\, a)P(d \,|\, b, c, a)P(e \,|\, d, c, a)P(f \,|\, a)$. In contrast, an application of the chain rule yields: $P(a, b, c, d, e, f) = P(a)P(b \,|\, a)P(c \,|\, b, a)P(d \,|\, b, c, a)P(e \,|\, d, c, b, a)P(f \,|\, e, d, c, b, a)$.

A comparison of the two previous factorizations highlights some of the independence constraints implied by the semi-Markov condition, for instance, $(C \perp\!\!\!\perp B \,|\, A)$, $(E \perp\!\!\!\perp B \,|\, D, C, A)$, and $(F \perp\!\!\!\perp E, D, C, B \,|\, A)$. The same applies to interventional distributions. First, let $P_c(A, B, C, D, E, F)$ be semi-Markov relative to $\mathcal{G}_{\overline{C}}$ (Figure 27.6(b)). Then, note that $P_c(A, B, C, D, E, F)$ factorizes as $P_c(a)\ P_c(b \,|\, a)P_c(c)\ P_c(d \,|\, b, c, a)P_c(e \,|\, d)P_c(f \,|\, a)$. This distribution satisfies the same conditional independence constraints as $P(A, B, C, D, E, F)$, but also additional ones such as $(E \perp\!\!\!\perp A \,|\, D)$. This constraint holds true as $(C \dashleftarrow\!\dashrightarrow E)$ is absent in $\mathcal{G}_{\overline{C}}$. The extended parents in both distributions are $Pa_e^{+} = \{A, C, D\}$ and $Pa_e^{C+} = \{D\}$. ■

27.4.2.2 CBNs with Latent Variables—Putting All the Pieces Together

The constructive procedure described in Definition 27.12 produces a coarsening of the underlying SCM such that (1) the arguments of the functions are preserved while their particular forms are discarded, and (2) the relationships between the exogenous variables are preserved while their precise distribution is discarded.[34] The pair $(\mathcal{G}, \mathbf{P}_*)$ consisting of a causal diagram $\mathcal{G}$, constructed through such a procedure, and the collection of interventional ($\mathcal{L}_2$) distributions, $\mathbf{P}_*$, will be called a CBN if it satisfies the definition below. This substitutes for Definition 27.11 in semi-Markovian models, and is similar to the way that constraints on a (observational) probability distribution (viz., conditional independencies) are captured by graphical constraints in a BN and the additional missing-link and do-see constraints are encoded in the Markov-CBNs (Definition 27.11).

Definition 27.15 **Causal Bayesian Network (CBN)-Semi-Markovian**

Let $\mathbf{P}_*$ be the collection of all interventional distributions $P(\mathbf{V} \mid do(\mathbf{x}))$, $\mathbf{X} \subseteq \mathbf{V}$, $\mathbf{x} \in \text{Val}(\mathbf{X})$, including the null intervention, $P(\mathbf{V})$, where $\mathbf{V}$ is the set of observed variables. A graphical model with directed and bidirected edges $\mathcal{G}$ is a CBN for $\mathbf{P}_*$ if for every intervention $do(\mathbf{X} = \mathbf{x})$, $\mathbf{X} \subseteq \mathbf{V}$, the following conditions hold:

(i) [Semi-Markovian] $P(\mathbf{V} \mid do(\mathbf{x}))$ is semi-Markov relative to $\mathcal{G}_{\overline{\mathbf{X}}}$.

(ii) [Missing directed-link] For every $V_i \in \mathbf{V} \backslash \mathbf{X}$, $\mathbf{W} \subseteq \mathbf{V} \backslash (Pa_i^{\mathbf{x}+} \cup \mathbf{X} \cup \{V_i\})$:

$$P(v_i \mid do(\mathbf{x}), pa_i^{\mathbf{x}+}, do(\mathbf{w})) = P(v_i \mid do(\mathbf{x}), pa_i^{\mathbf{x}+}), \tag{27.28}$$

(iii) [Missing bidirected-link] For every $V_i \in \mathbf{V} \backslash \mathbf{X}$, let $Pa_i^{\mathbf{x}+}$ be partitioned into two sets of confounded and unconfounded parents, Pa_i^c and Pa_i^u in $\mathcal{G}_{\overline{\mathbf{X}}}$. Then

$$P(v_i \mid do(\mathbf{x}), pa_i^c, do(pa_i^u)) = P(v_i \mid do(\mathbf{x}), pa_i^c, pa_i^u). \tag{27.29}$$

■

The first condition requires each interventional distribution to factorize in a semi-Markovian fashion relative to the corresponding interventional graph $\mathcal{G}_{\overline{\mathbf{X}}}$, as discussed in Example 27.11. The remaining conditions give semantics for the missing directed and bidirected links in the model, which encode the lack of direct effect and of unobserved confounders between the corresponding variables, respectively. Specifically, the missing directed-link condition (ii) states that under any intervention $do(\mathbf{X} = \mathbf{x})$, conditioning on the set of augmented parents $Pa_i^{\mathbf{x}+}$ renders V_i invariant to an intervention on other variables $\mathbf{W}$—in other words, $\mathbf{W}$ has no direct effect on V_i. For instance, note that for $V_i = D$ in Figure 27.6(a), $P(d \mid do(f, e), b, c, a) = P(d \mid b, c, a)$ as well as $P(d \mid do(b, c), do(a, f, e)) = P(d \mid do(b, c))$.

[34] Given the lack of constraints over the form and shape of the underlying functions and distribution of the exogenous variables, it is possible to non-parametrically write one in terms of the other.

Further, the missing bidirected-link condition relaxes the stringent parents do/see condition in Markovian CBNs (Definition 27.11(iii)). Note that the do/see condition does not hold due to the unobserved correlation between certain endogenous variables, for instance, both $P(d \mid do(b)) = P(d \mid b)$ and $P(e \mid do(d)) = P(e \mid d)$ do not hold in Figure 27.6(a).[35] Still, given the set of extended parents of V_i, observations and interventions on parents not connected via a bidirected path (i.e., Pa_i^u) yield the same distribution. For instance, $P(e \mid do(a, d), c) = P(e \mid a, d, c)$, where $Pa_e^u = \{A, D\}, Pa_e^c = \{C\}$; also, $P(d \mid do(b, a, c)) = P(d \mid do(b), a, c)$, where $Pa_d^u = \{A, C\}, Pa_d^c = \{B\}$. There exists no unobserved confounding in Markovian models, so $Pa_i^u = Pa_i$, which means that the condition is enforced for all parents.

Finally, the causal diagram $\mathcal{G}$ constructed from the SCM and the set of interventional distributions $\mathbf{P}_*$ can be formally connected through the following result:

Theorem 27.4 **$\mathcal{L}_2$-Connection—SCM-CBN (Semi-Markovian)**

The causal diagram $\mathcal{G}$ induced by the SCM $\mathcal{M}$ (following the constructive procedure in Definition 27.12) is a CBN for $\mathbf{P}_*^{\mathcal{M}}$. ■

One could take an axiomatic view of CBNs and consider alternative constructions that satisfy their conditions, detached from the structural semantics (similarly to the Markovian case). We provide in Bareinboim et al. [2020, appendix D] a procedure called ConstructCBN (see Theorem 10) that constitutes such an alternative. It can be seen as the experimental-stochastic counterpart of the SCM-functional Definition 27.12. We show in the next section that CBNs can act as a basis for causal inference regardless of their underlying generating model.

27.4.2.3 Cross-layer Inferences through CBNs with Latent Variables

The causal diagram associated with a CBN will sometimes be a proper surrogate for the SCM, and allow one to compute the effect of interventions *as if* the fully specified SCM were available. Unfortunately, in some other cases, it will be insufficient, as evident from the discussion in Example 27.10. We introduce next the notion of identifiability [Pearl 2000, p. 77] to more visibly capture each of these instances.

Definition 27.16 **Effect Identifiability**

The causal effect of an action $do(\mathbf{X} = \mathbf{x})$ on a set of variables $\mathbf{Y}$ given a set of observations on variables $\mathbf{Z} = \mathbf{z}$, $P(\mathbf{Y} \mid do(\mathbf{x}), \mathbf{z})$, is said to be identifiable from P and $\mathcal{G}$ if for every two models $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$ with causal diagram $\mathcal{G}$, $P^{(1)}(\mathbf{v}) = P^{(2)}(\mathbf{v}) > 0$ implies $P^{(1)}(\mathbf{Y} \mid do(\mathbf{x}), \mathbf{z}) = P^{(2)}(\mathbf{Y} \mid do(\mathbf{x}), \mathbf{z})$. ■

[35] To see why this is the case in the last expression, first let U_d be any exogenous argument to f_D. Now note that $P(e \mid do(d))$ does not depend on U_d, while $P(e \mid d)$ does due to the path $U_d \longrightarrow D \longleftarrow C \dashleftarrow\dashrightarrow E$.

This formalizes the very natural type of cross-layer inference we have discussed in Figure 27.3, namely: given qualitative assumptions encoded in the causal diagram $\mathcal{G}$, one would like to establish whether the interventional distribution ($\mathcal{L}_2$-quantity) $P(\mathbf{Y} \mid do(\mathbf{x}), \mathbf{z})$ is inferable from the observational one ($\mathcal{L}_1$-data). We introduce next a set of inference rules known as *do-calculus* [Pearl 1995] developed to answer this question.[36,37]

Theorem 27.5 **Do-Calculus**

Let $\mathcal{G}$ be a CBN for $\mathbf{P}_*$, *then* $\mathbf{P}_*$ satisfies the Do-Calculus rules according to $\mathcal{G}$. Namely, for any disjoint sets $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$ the following three rules hold:

$$\textbf{\textit{Rule 1}} \quad P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{z}, \mathbf{w}) = P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{w}) \qquad \textit{if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{X}, \mathbf{W}) \textit{ in } \mathcal{G}_{\overline{\mathbf{X}}}. \tag{27.30}$$

$$\textbf{\textit{Rule 2}} \quad P(\mathbf{y} \mid do(\mathbf{x}), do(\mathbf{z}), \mathbf{w}) = P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{z}, \mathbf{w}) \qquad \textit{if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{X}, \mathbf{W}) \textit{ in } \mathcal{G}_{\overline{\mathbf{X}}\underline{\mathbf{Z}}}. \tag{27.31}$$

$$\textbf{\textit{Rule 3}} \quad P(\mathbf{y} \mid do(\mathbf{x}), do(\mathbf{z}), \mathbf{w}) = P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{w}) \qquad \textit{if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{X}, \mathbf{W}) \textit{ in } \mathcal{G}_{\overline{\mathbf{X}\mathbf{Z}(\mathbf{W})}}, \tag{27.32}$$

where a graph $\mathcal{G}_{\overline{\mathbf{X}}\underline{\mathbf{Z}}}$ is obtained from $\mathcal{G}$ by removing the arrows incoming to $\mathbf{X}$ and outgoing from $\mathbf{Z}$, *and* $\mathbf{Z}(\mathbf{W})$ is the set of $\mathbf{Z}$-nodes non-ancestors of $\mathbf{W}$ in the corresponding graph. ■

These rules can be seen as a tool that allows one to navigate in the space of interventional distributions, jumping across unrealized worlds, and licensed by the invariances encoded in the causal graph. Specifically, rule 1 can be seen as an extension of the d-separation criterion for reading conditional independences under a fixed intervention $do(\mathbf{X} = \mathbf{x})$ from the graph denoted $\mathcal{G}_{\overline{\mathbf{X}}}$. Furthermore, rules 2 and 3 entail constraints among distributions under different interventions. Rule 2 permits the *exchange* of a $do(\mathbf{z})$ operator with an observation of $\mathbf{Z} = \mathbf{z}$, capturing situations when intervening and observing $\mathbf{Z}$ influence the set of variables $\mathbf{Y}$ indistinguishably. Rule 3 licenses the *removal* or *addition* of an intervention from

[36] The do-calculus can be seen as an inference engine that allows the local constraints encoded in the CBN, in terms of the family relationships, to be translated and combined to generate (global) constraints involving other variables.

[37] The duality between local and global constraints is a central theme in probabilistic reasoning, where the family factorization dictated by the graphical model is local while d-separation is global, allowing one to read off non-trivial constraints implied by the model [Pearl 1988, Lauritzen 1996]. The graphical model could be seen as a basis, that is, a parsimonious encoder of exponentially many conditional independences. In causal inference, do-calculus can be seen as a generalization of d-separation to generate global, interventional-type of constraints.

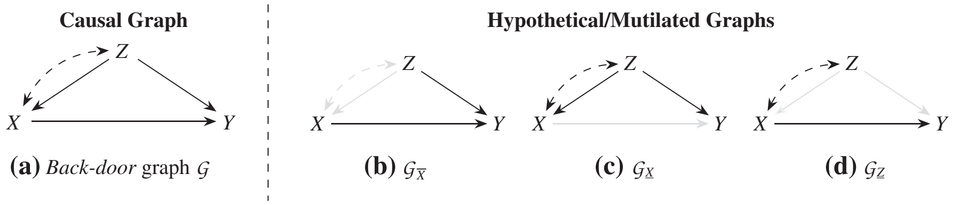

Causal Graph
Hypothetical/Mutilated Graphs
Z
X
Y
(a) Back-door graph $\mathcal{G}$
(b) $\mathcal{G}_{\overline{X}}$
(c) $\mathcal{G}_{\underline{X}}$
(d) $\mathcal{G}_{\underline{Z}}$

Figure 27.7 (a) Graph representing a model where the query $P(y \mid do(x))$ is identifiable. The query can be derived using do-calculus rules licensed by graphs (b), (c), and (d).

a probability expression, recognizing situations where $do(\mathbf{z})$ has no effect whatsoever on $\mathbf{Y}$. A more detailed discussion of do-calculus can be found in Pearl [2000, chapter 3].[38]

We have previously shown that in simple settings causal inference is unattainable with only $\mathcal{L}_1$-data, and that knowledge conveniently encoded in the form of a causal diagram is required. Next, we show how the knowledge from the diagram together with the inference rules of do-calculus allows for the identification of the query $P(y \mid do(x))$ in the context of the model represented in Figure 27.7(a). First, we start with the target query and then apply do-calculus:

$$P(y \mid do(x)) = \sum_z P(y \mid do(x), z) P(z \mid do(x)) \qquad \text{Summing over } Z \qquad (27.33)$$

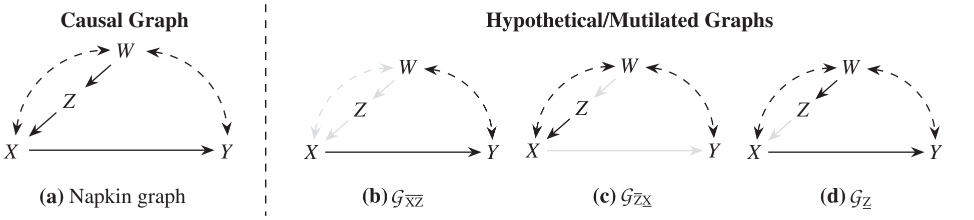
$$= \sum_z P(y \mid do(x), z) P(z) \qquad \text{Rule 3: } (Z \perp\!\!\!\perp X)_{\mathcal{G}_{\overline{X}}} \qquad (27.34)$$

$$= \sum_z P(y \mid x, z) P(z) \qquad \text{Rule 2: } (Y \perp\!\!\!\perp X \mid Z)_{\mathcal{G}_{\underline{X}}}. \qquad (27.35)$$

Each step above is accompanied by the corresponding probability axiom or rule, supported by the licensing graphs $\mathcal{G}_{\overline{X}}$ and $\mathcal{G}_{\underline{X}}$ (Figure 27.7(b) and (c), respectively). As desired, the RHS of Equation (27.35) is a function of $P(\mathbf{V})$, hence, estimable from $\mathcal{L}_1$-data. This means that no matter the functional form of the endogenous variables or the distribution over the exogenous ones, for all SCMs compatible with the graph in Figure 27.7(a), the causal effect of X on Y will always be equal to Equation (27.35). This can be seen as an instance of the back-door criterion [Pearl 1993], and the particular function in Equation (27.35) is known as adjustment (for $\mathbf{Z}$).

The importance of the back-door criterion stems from the fact that adjustment is a very common technique used to identify causal effects in the sciences. While the adjustment expression has been used since much earlier than the discovery of

[38] Interestingly, the do-calculus theorem (Theorem 27.5) as stated here was derived entirely within the domain of CBNs and Layer 2 constraints, which contrasts with the traditional proposition ([Pearl 1995, theorem 27.3]) based on Layer 3 facts.



(a) Napkin graph

(b) $\mathcal{G}_{\overline{XZ}}$

(c) $\mathcal{G}_{\overline{Z}\underline{X}}$

(d) $\mathcal{G}_{\underline{Z}}$

Figure 27.8 Napkin graph (a) and derived graphs used to identify $P(y \mid do(x))$.

the criterion itself [Pearl 1993], the back-door is the first to provide a transparent way one could judge the plausibility of the assumptions required to map $\mathcal{L}_1$-data to an $\mathcal{L}_2$-quantity based on a model of the world.[39]

For the effect of Z on X, $P(X \mid do(z))$, in the same graph (Figure 27.7(a)), there exists no set $\mathbf{Z}$ that can be used to identify the effect by adjustment. Note that in the graph where the arrows outgoing from Z are cut (Figure 27.7(d)), Z and X cannot be separated due to the existence of the latent path, $Z \dashleftarrow\dashrightarrow X$. More strongly, $P(X \mid do(z))$ is not identifiable from the observational distribution by any other means. We leave as an exercise the construction of a counter-example based on Example 27.10's proof. Broadly, the effect of a certain intervention may or may not be identifiable, depending on the particular causal diagram and the topological relations between treatment, outcome, and latent variables.

Finally, there are involved scenarios that are somewhat surprising as they go beyond some of the intuitions discussed in the examples above; see diagram in Figure 27.8(a). The task is to identify the effect of X on Y, $P(Y \mid do(x))$, from $P(W, Z, X, Y)$. It is obvious that the effect cannot be identified by the back-door criterion, and in $\mathcal{G}_{\underline{X}}$, conditioning on $\{Z\}, \{W\}, \{Z, W\}$ leaves the back-door path $X \dashleftarrow\dashrightarrow W \dashleftarrow\dashrightarrow Y$ opened. After all, one may be tempted to believe that the effect of X on Y is not identifiable in this case. Contrary to this intuition, consider the following derivation in do-calculus:

$$
\begin{aligned}
P(y \mid do(x)) &= P(y \mid do(x), do(z)) && \text{Rule 3: } (Y \perp\!\!\!\perp Z \mid X)_{\mathcal{G}_{\overline{XZ}}} && (27.36)\\
&= P(y \mid do(z), x) && \text{Rule 2: } (Y \perp\!\!\!\perp X)_{\mathcal{G}_{\overline{Z}\underline{X}}} && (27.37)\\
&= \frac{P(y, x \mid do(z))}{P(x \mid do(z))} && \text{Def. of cond. probability.} && (27.38)
\end{aligned}
$$

The rules used in each step and the licensing graphs are shown in Figure 27.8(b)–(c). At this point, the back-door adjustment (similar to Equations (27.33)–(27.35))

[39] The back-door criterion provides a formal and transparent condition to judge the validity of a condition called *conditional ignorability* [Imbens and Rubin 2015]; see further details in Pearl [2000, section 11.3.2].

can be applied to solve for both factors in Equation (27.38). To witness, note that in the numerator, $P(y, x \mid do(z))$, $\{W\}$ is back-door admissible with respect to $(Z, \{Y, X\})$, as $(Y, X \perp\!\!\!\perp Z \mid W)_{\mathcal{G}_{\underline{Z}}}$, as shown in Figure 27.8(d). The denominator follows by marginalizing Y out. Putting these two results together and replacing it back into Equation (27.38) lead to:

$$P(y \mid do(x)) = \frac{\sum_w P(y, x \mid z, w)P(w)}{\sum_w P(x \mid z, w)P(w)}. \tag{27.39}$$

The RHS of Equation (27.39) is expressible in terms of $P(\mathbf{V})$, which means that for any SCM compatible with the graph, the causal effect will always be the same, regardless of the details of the underlying mechanisms and distribution over the exogenous variables. The expression shown in Equation (27.39) is a ratio following from the application of the back-door criterion twice.

The problem of deciding identifiability, also known as non-parametric identification, has been extensively studied in the literature. There are a number of conditions that have been proposed to solve this problem, including Galles and Pearl [1995], Pearl and Robins [1995], Kuroki and Miyakawa [1999], and Spirtes et al. [2001]. The do-calculus provides a general mathematical treatment for non-parametric identification [Pearl 1995]. It has been made systematic and shown to be complete for the task of identification from a combination of observations and experiments [Tian and Pearl 2002a, Huang and Valtorta 2006, Shpitser and Pearl 2006, Bareinboim and Pearl 2012, Lee et al. 2019]. In other words, given a causal diagram $\mathcal{G}$ and a collection of observational and experimental distributions, the target effect of $\mathbf{X}$ on $\mathbf{Y}$ given a set of covariates $\mathbf{Z}$, $P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{z})$, is identifiable if and only if there exists a sequence of application of the rules of do-calculus that reaches an estimand in terms of the available distributions.

27.5 Conclusions

We investigated a mathematical structure called the PCH, which was discovered by Judea Pearl when studying the conditions under which some types of causal explanations can be inferred from data [Pearl 2000, Pearl and Mackenzie 2018]. The PCH is certainly one of the most productive conceptual breakthroughs in the science of causal inference over the last decades. It highlights and formalizes the distinct roles of some basic human capabilities—*seeing*, *doing*, and *imagining*—spanning cognition, AI, and scientific discovery. The structure is pervasive in the empirical world: as long as a complex system can be described as a collection of causal mechanisms—that is, an SCM (Definition 27.1)—the hierarchy relative to the modeled phenomena emerges (Definition 27.8).

The main contribution of this chapter is a detailed analysis of the PCH through different perspectives: one semantical (Section 27.2), another logical-probabilistic (Section 27.3), and another inferential-graphical (Section 27.4). These complementary approaches elucidate the PCH from different angles, ranging from when one knows everything about a specific SCM (semantical), to talking about classes of SCMs in general (probabilistic), and ending with one SCM that is particular to the environment of interest but which is not fully observed (graphical). We hope these distinct angles provide a powerful tool for studying causation across different research communities, with far-reaching implications for scientific practice in a wide range of data-driven fields. For instance, we expect these results to underpin the next generation of AI systems, which should be data-efficient, explainable, and aligned with society's goals.

Acknowledgments

This work has benefited immensely from conversations with David Blei, Carlos Cinelli, Paul Daniell, Philip Dawid, Heyang Gong, David Kinney, Sanghack Lee, Judea Pearl, Jin Tian, Yuhao Wang, and Junzhe Zhang. We are grateful to the Editors, Professors Rina Dechter, Hector Geffner, and Joe Halpern, for the opportunity to contribute to this volume. Elias Bareinboim and Juan Correa were partially supported by grants from NSF IIS-1704352 and IIS-1750807 (CAREER). Duligur Ibeling was supported by the NSF Graduate Research Fellowship (DGE-1656518). Thomas Icard was partially supported by the Center for the Study of Language and Information.

References

J. Aldrich. 1989. Autonomy. *Oxford Econ. Pap.* 41, 15–34. DOI: https://doi.org/10.1093/oxfordjournals.oep.a041889.

E. Bareinboim and J. Pearl. 2012. Causal inference by surrogate experiments: z-Identifiability. In N. d. F. Murphy and Kevin (Eds.), *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 113–120.

E. Bareinboim and J. Pearl. 2016. Causal inference and the data-fusion problem. *Proc. Natl. Acad. Sci.* 113, 27, 7345–7352. DOI: https://doi.org/10.1073/pnas.1510507113.

E. Bareinboim, C. Brito, and J. Pearl. 2012. Local characterizations of causal Bayesian networks. In M. Croitoru, S. Rudolph, N. Wilson, J. Howse, and O. Corby (Eds.), *Graph Structures for Knowledge Representation and Reasoning*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1–17. DOI: https://doi.org/10.1007/978-3-642-29449-5_1.

E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard. 2020. *On Pearl's Hierarchy and the Foundations of Causal Inference*. Technical Report R-60, Causal AI Lab, Columbia University.

E. W. Beth. 1956. On Padoa's method in the theory of definition. *J. Symb. Log.* 2, 1, 194–195. DOI: https://doi.org/10.2307/2268764.

R. Briggs. 2012. Interventionist counterfactuals. *Philos. Stud.* 160, 1, 139–166. DOI: https://doi.org/10.1007/s11098-012-9908-5.

N. Cartwright. 1989. *Nature's Capacities and Their Measurement*. Clarendon Press, Oxford. DOI: https://doi.org/10.1093/0198235070.001.0001.

N. Chomsky. 1959. On certain formal properties of grammars. *Inf. Control*. 2, 137–167. https://doi.org/10.1016/S0019-9958(59)90362-6.

A. P. Dawid. 2000. Causal inference without counterfactuals (with comments and rejoinder). *J. Am. Stat. Assoc.* 95, 450, 407–448. DOI: https://doi.org/10.1080/01621459.2000.10474210.

R. Fagin, J. Y. Halpern, and N. Megiddo. 1990. A logic for reasoning about probabilities. *Inf. Comput.* 87, 1/2, 78–128. DOI: https://doi.org/10.1016/0890-5401(90)90060-U.

R. A. Fisher. 1936. Design of experiments. *Br. Med. J.* 1, 3923, 554. DOI: https://doi.org/10.1136/bmj.1.3923.554-a.

D. Galles and J. Pearl. 1995. Testing identifiability of causal effects. In P. Besnard and S. Hanks (Eds.), *Uncertainty in Artificial Intelligence 11*. Morgan Kaufmann, San Francisco, 185–195.

D. Galles and J. Pearl. 1998. An axiomatic characterization of causal counterfactuals. *Found. Sci.* 3, 1, 151–182. DOI: https://doi.org/10.1023/A:1009602825894.

T. Haavelmo. 1943. The statistical implications of a system of simultaneous equations. *Econometrica* 11, 1, 1. DOI: https://doi.org/10.2307/1905714.

J. Y. Halpern. 1998. Axiomatizing causal reasoning. In G. F. Cooper and S. Moral (Eds.), *Uncertainty in Artificial Intelligence*. Cornell University, Morgan Kaufmann, San Francisco, CA, 202–210.

J. Y. Halpern. 2000. Axiomatizing causal reasoning. *J. Artif. Intell. Res.* 12, 317–337. DOI: https://doi.org/10.1613/jair.648.

J. Y. Halpern. 2013. From causal models to counterfactual structures. *Rev. Symb. Logic.* 6, 2, 305–322. DOI: https://doi.org/10.1017/S1755020312000305.

Y. Huang and M. Valtorta. 2006. Identifiability in causal Bayesian networks: A sound and complete algorithm. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI 2006)*. AAAI Press, Menlo Park, CA, 1149–1156.

D. Hume. 1739. *A Treatise of Human Nature*. Oxford University Press, Oxford.

D. Hume. 1748. *An Enquiry Concerning Human Understanding*. Open Court Press, LaSalle.

D. Ibeling and T. Icard. 2018. On the conditional logic of simulation models. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. 1868–1874. DOI: https://doi.org/10.24963/ijcai.2018/258.

D. Ibeling and T. Icard. 2019. On open-universe causal reasoning. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*.

D. Ibeling and T. Icard. 2020. Probabilistic reasoning across the causal hierarchy. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. DOI: https://doi.org/10.1609/aaai.v34i06.6577.

T. Icard. 2020. Calibrating generative models: The probabilistic Chomsky–Schützenberger hierarchy. *J. Math. Psychol.* 95. DOI: https://doi.org/10.1016/j.jmp.2019.102308.

G. W. Imbens and D. B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, Cambridge, MA. DOI: https://doi.org/10.1017/CBO9781139025751.

D. Koller and N. Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

M. Kuroki and M. Miyakawa. 1999. Identifiability criteria for causal effects of joint interventions. *J. R. Stat. Soc.* 29, 105–117. DOI: https://doi.org/10.14490/jjss1995.29.105.

S. L. Lauritzen. 1996. *Graphical Models*. Clarendon Press, Oxford.

S. Lee, J. D. Correa, and E. Bareinboim. 2019. General identifiability with arbitrary surrogate experiments. In *Proceedings of the Thirty-Fifth Conference Annual Conference on Uncertainty in Artificial Intelligence*. AUAI Press, in press, Corvallis, OR.

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

J. Locke. 1690. *An Essay Concerning Human Understanding*. London, Thomas Basset.

J. L. Mackie. 1980. *The Cement of the Universe: A Study of Causation*. Clarendon Press, Oxford. DOI: https://doi.org/10.1093/0198246420.001.0001.

J. Marschak. 1950. Statistical inference in economics. In T. Koopmans (Ed.), *Statistical Inference in Dynamic Economic Models*. Wiley, New York, 1–50.

T. Maudlin. 2019. The why of the world. *Boston Review*. https://bostonreview.net/science-nature/tim-maudlin-why-world. Accessed Febuary 10, 2020.

J. Neyman. 1923. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Stat. Sci.* 5, 4, 465–480. DOI: https://doi.org/10.1214/ss/1177012031.

J. Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.

J. Pearl. 1993. Aspects of graphical models connected with causality. In *Proceedings of the 49th Session of the International Statistical Institute*, 1 (August), 399–401.

J. Pearl. 1995. Causal diagrams for empirical research. *Biometrika* 82, 4, 669–688. DOI: https://doi.org/10.1093/biomet/82.4.669.

J. Pearl. 2000. *Causality: Models, Reasoning, and Inference*. (2nd. ed.). Cambridge University Press, NY. DOI: https://doi.org/10.10170S0266466603004109.

J. Pearl. 2001. Bayesianism and causality, or, why I am only a half-Bayesian. In *Foundations of Bayesianism, Applied Logic Series, Volume 24*. Kluwer Academic Publishers, 19–36. DOI: https://doi.org/10.1007/978-94-017-1586-7_2.

J. Pearl. 2012. The mediation formula: A guide to the assessment of causal pathways in nonlinear models. In C. Berzuini, P. Dawid, and L. Bernardinelli (Eds.), *Causality: Statistical Perspectives and Applications,* John Wiley and Sons, Ltd, Chichester, UK, 151–179. DOI: https://doi.org/10.1002/9781119945710.ch12.

J. Pearl and E. Bareinboim. 2019. A note on “ generalizability of study results.” *J. Epidemiol.* 30, 186–188. DOI: https://doi.org/10.1097/EDE.0000000000000939.

J. Pearl and D. Mackenzie. 2018. *The Book of Why*. Basic Books, New York.

J. Pearl and J. M. Robins. 1995. Probabilistic evaluation of sequential plans from causal models with hidden variables. In *Uncertainty in Artificial Intelligence 11*. Morgan Kaufmann, 444–453.

D. C. Penn and D. J. Povinelli. 2007. Causal cognition in human and nonhuman animals: A comparative, critical review. *Annu. Rev. Psychol.* 58, 97–118. DOI: https://doi.org/10.1146/annurev.psych.58.110405.085555.

J. Peters, D. Janzing, and B. Schlkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press.

G. de Pierris. 2015. *Ideas, Evidence, and Method: Hume’s Skepticism and Naturalism concerning Knowledge and Causation*. Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780198716785.001.0001.

J. M. Robins. 1986. A new approach to causal inference in mortality studies with a sustained exposure period—applications to control of the healthy workers survivor effect. *Math. Model.* 7, 1393–1512. DOI: https://doi.org/10.1016/0270-0255(86)90088-6.

P. R. Rosenbaum and D. B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1, 41–55. DOI: https://doi.org/10.1093/biomet/70.1.41.

P. K. Rubenstein, S. Weichwald, S. Bongers, J. M. Mooij, D. Janzing, M. Grosse-Wentrup, and B. Schölkopf. 2017. Causal consistency of structural equation models. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*.

D. B. Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66, 5, 688–701. DOI: https://doi.org/10.1037/h0037350.

B. Schölkopf. 2019. Causality for machine learning. *arXiv preprint arXiv:1911.10500*.

I. Shpitser and J. Pearl. 2006. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the Twenty-First AAAI Conference on Artificial Intelligence*. 2, 1219–1226.

H. A. Simon. 1953. Causal ordering and identifiability. In W. C. Hood and T. C. Koopmans (Eds.), *Studies in Econometric Method,* Wiley and Sons, Inc., New York, 49–74. DOI: https://doi.org/10.1007/978-94-010-9521-1_5.

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

L. J. Stockmeyer. 1977. The polynomial-time hierarchy. *Theor. Comput. Sci.* 3, 1–22. DOI: https://doi.org/10.1016/0304-3975(76)90061-X.

R. H. Strotz and H. O. A. Wold. 1960. Recursive versus nonrecursive systems: An attempt at synthesis. *Econometrica* 28, 417–427. DOI: https://doi.org/10.2307/1907731.

P. Suppes and M. Zanotti. 1981. When are probabilistic explanations possible? *Synthese* 48, 191–199. DOI: https://doi.org/10.1007/BF01063886.

R. S. Sutton and A. G. Barto. 2018. *Reinforcement Learning: An Introduction*. (2nd. ed.) The MIT Press.

J. Tian and J. Pearl. 2002a. A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI 2002)*. 567–573.

J. Tian and J. Pearl. 2002b. On the testable implications of causal models with hidden variables. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*. 519–527.

T. VanderWeele. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press.

J. Woodward. 2003. *Making Things Happen*. Oxford University Press, New York. DOI: https://doi.org/10.1093/0195155270.001.0001.

G. H. von Wright. 1971. *Explanation and Understanding*. Cornell University Press. DOI: https://doi.org/10.1007/978-94-010-1823-4_15.

J. Zhang. 2013. A Lewisian logic of causal counterfactuals. *Minds Mach.* 23, 77–93. DOI: https://doi.org/10.1007/s11023-011-9261-z.

J. Zhang and E. Bareinboim. 2018. Fairness in decision-making—the causal explanation formula. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. 2037–2045.