TECHNICAL REPORT
R-60
First Version: July, 2020
Last Revision: Mar, 2021

1 On Pearl's Hierarchy and the Foundations of Causal Inference

Elias Bareinboim[†] EB@CS.COLUMBIA.EDU
Juan D. Correa[†] JDCORREA@CS.COLUMBIA.EDU
[†] *Columbia University*
New York, NY 10027, USA

Duligur Ibeling[‡] DULIGUR@STANFORD.EDU
Thomas Icard[‡] ICARD@STANFORD.EDU
[‡] *Stanford University*
Stanford, CA 94305, USA

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" [Pearl and Mackenzie 2018], 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. Since 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 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 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 PCH's layers. We believe that connecting with the essential dimensions of human experience as delineated by the PCH is a critical step towards creating the next generation of AI systems that will be safe, robust, human-compatible, and aligned with the social good.

1.1 Introduction

Causal information is deemed highly valuable and desirable along many dimensions of the human endeavor, including in 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, for instance [Pearl and Mackenzie 2018, p. 1]: "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?"

At an intuitive level, the capacity for processing causal information is central to human cognition from early development, playing a critical role in higher-level cognition, allowing us to plan a course of action, to assign credit, to determine blame and responsibility, and to generalize across changing conditions. More personally, it allows us to understand ourselves, to interact with others, and to make sense of the world around us. Among the first tasks confronting an infant is to discover what kinds of objects are in the world and how those objects are causally related to one another. The past several decades of work in developmental psychology have uncovered striking ways in which children explore an unknown world in much the same manner as a scientist would [Gopnik 2012]. They ask and answer "What if?" and "Why?" questions [Buchsbaum et al. 2012], use data to formulate causal hypotheses, and even test those hypotheses by actively performing interventions on the environment [Gopnik et al. 2004]. By adulthood, our causal knowledge forms the very cement that holds our understanding of the world together [Danks 2014, Sloman and Lagnado 2015].

In a more systematic fashion, causality plays a central role on how we probe the physical world around us and ultimately understand Nature. Standard scientific methodology is built around the idea of combining observations and experiments (more on their distinction later on) and formulating hypotheses about unobserved *causal mechanisms*, submitting these hypotheses to further observation and experimentation in a continual process of refinement [Machamer et al. 2000, Salmon 1984, Woodward 2002, 2003]. In modern molecular biology, for example, scientists can explain the synthesis of proteins by first identifying the critical molecular components involved – DNA, mRNA, tRNA, rRNA, codons, amino acids – and putting them together in a series of steps that lead from initial transcription of DNA into mRNA, all the way down to the final protein folding. In a similar vein but a rather different context, economists have been able to predict and explain macro-level consumer behavior using models of individual choice behavior over a lifetime, as a function of other relevant variables like income, assets, and interest rates (see, e.g., [Deaton 1992] for a classic example).

Causal explanations like these purport to be more than mere descriptions, or summaries of the observed data. By breaking down a phenomenon into modular components, and describing how they interact to produce an emergent behavior or a final product [Simon 1953], scientists seek to uncover the underlying *data-generating processes*, or features thereof. When successful, it allows one to infer what *would* or *could* happen under various hypothetical (counter-to-fact) suppositions, going beyond the limited observations (i.e., data) afforded up to that point. For instance, biologists are able to predict the effect that bacterial or viral pathogens might have on otherwise normal protein pathways, while economists may predict what effect higher interest rates would have on consumption and economic activity. Practically speaking, mechanistic knowledge of this sort can often support cleaner and more surgical interventions, which has the potential to allow one to bring about desired states of affairs [Woodward 2003], whether social, economic, or political.

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 then 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, Mackie 1980, von Wright 1971], to cite a few prominent references, and Pearl, his collaborators, and

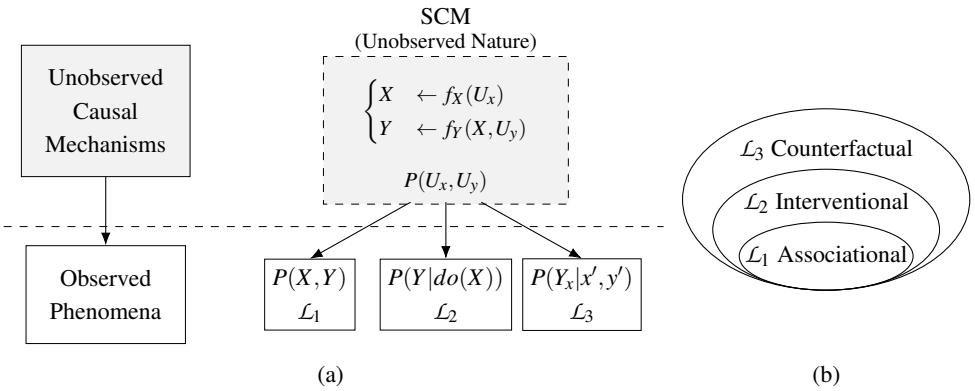


Figure 1.1: (a) Collection of causal mechanisms (or SCM) generating certain observed phenomena (qualitatively different probability distributions). (b) PCH's containment structure.

many others helped to understand and formalize this notion. In fact, following this precise mathematization, we now see a blossoming of algorithmic developments and rapid expansion towards applications.[1]

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).[2] 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, 2009]. 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.

[1] It lies outside the scope of this chapter to pursue a detailed historical account, and we refer readers to [Pearl and Mackenzie 2018] for additional context.

[2] 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, §4.16]. See [de Pierris 2015] for discussion.

This intuitive picture is illustrated in Fig. 1.1(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 1.2. 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.[3] Nevertheless, we ordinarily 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 1]. 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, viz. 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.

Roadmap of the Chapter. Against this background, we start in Sec. 1.2 by showing how the Pearl Causal Hierarchy naturally emerges from a structural causal model, 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.[4] In Sec. 1.3, we prove that the PCH

[3] At the same time, many of the natural sciences, most prominently physics, chemistry, will often purport to determine the underlying causal mechanisms quite precisely.

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

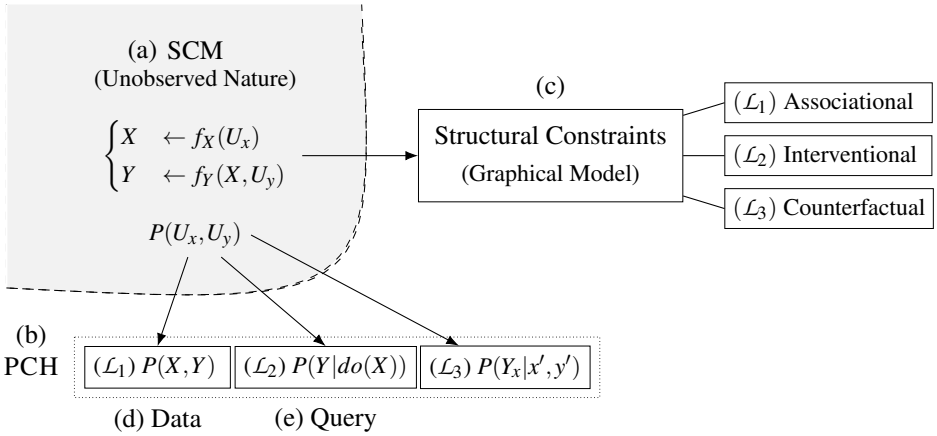


Figure 1.2: Schema depicting building blocks of canonical causal inferences – on the top, the SCM itself, i.e., the unobserved collection of mechanisms and underlying uncertainty (a); on the bottom, different probability distributions forming PCH's layers (b); on the right, structural constraints entailed by the SCM (c). Example of possible input dataset (d) and query (e).

is strict for almost-all SCMs (Thm. 1), in a technical sense of 'almost-all' (Fig. 1.1(b)).[5] It follows (Corollary 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]. This first part of the chapter does not directly address the practice of causal inference; rather, it formally establishes a general motivation for causal inference from a logical perspective.

In the second part of the chapter (Sec. 1.4), 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 (Fig. 1.2(a)) is essentially hopeless.[6] 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 (Fig. 1.2(b)) from a partial understanding of the

[5] Hierarchies abound in logic and computer science, particularly those pertaining to computational resources, 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.

[6] Of course, if we have been able to induce the structural mechanisms themselves – as may be feasible in some of the sciences, e.g., 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.

SCM (Fig. 1.2(c)). Technically speaking, if one has layer-1 type of data (Fig. 1.2(d)), e.g., collected through random sampling, and aims to infer the effect of a new intervention (layer-2 type of query, (Fig. 1.2(e)), we show that the problem is not always solvable. In words, there is not enough information about the SCM encoded in the dataset (coming from the realized world) so that one can learn how the system would react when submitted to a new intervention (a still unrealized, hypothetical world).

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 (Fig. 1.2(e)). In particular, we use these constraints to define a new family of graphical models called *Causal Bayesian Networks* (CBNs), which are comprised of a pair, a graphical model and a collection of observational and interventional distributions. We present a constructive definition of CBNs that naturally emerges from an SCM, as well as one that is purely empirical. This treatment generalizes existing characterizations [Bareinboim et al. 2012, Pearl 2000] to the semi-Markovian setting and allows for the existence of unobserved confounders. 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 given a combination of partial knowledge of the underlying structural model, in the form of a causal graph, and data at lower layers.

Finally, in Sec. 1.5, we conclude summarizing our main contributions and putting this work into the broader context of AI and data science. In particular, we outline some of the ways that progress toward the goal of developing safe, robust, explainable, and human-compatible artificial systems will be greatly amplified by further appreciation of both the inherent limitations and the exciting possibilities afforded by the study of Pearl's Hierarchy.

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 $P(x_1) = P(x_2) = \sum_{x_2} P(x_1,x_2) = \sum_{\mathbf{x}(X_1)=x_1} P(\mathbf{x}) = 1/2$.

1.2 Structural Causal Models and the Causal Hierarchy

We build on the language of *Structural Causal Models* (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.

	Layer (Symbolic)	**Typical Activity**	**Typical Question**	**Example**	**Machine Learning**
$\mathcal{L}_1$	Associational $P(y\|x)$	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 $P(y\|do(x),c)$	Doing	What if? What if I do X?	What if I take aspirin, will my headache be cured?	Reinforcement Learning
$\mathcal{L}_3$	Counterfactual $P(y_x\|x',y')$	Imagining	Why? What if I had acted differently?	Was it the aspirin that stopped my headache?	

Table 1.1: Pearl's Causal Hierarchy.

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 Pearl Causal Hierarchy, or PCH (Fig. 1.1). 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 to distinct notions within human cognition. Each of these allows one to articulate with mathematical precision qualitatively different types of question regarding the observed variables of the underlying system; for some examples, see Table 1.1.

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.

Definition 1 (Structural Causal Model (SCM))**.** A structural causal model $\mathcal{M}$ 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 $U_i \subseteq \mathbf{U}$, $Pa_i \subseteq \mathbf{V} \setminus V_i$, 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

$$v_i \leftarrow f_i(pa_i, u_i), \tag{1.1}$$

i.e., 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 structural causal model 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, 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).[7] 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 is finite[8] and that all models are *recursive* (i.e., acyclic).[9] A structural model is *Markovian* if the exogenous parent sets U_i, U_j are independent whenever $i \neq j$. In the treatment provided here, we allow for the sharing of exogenous parents and we allow for arbitrary dependences among the exogenous variables, which means that, in general, the SCM need not to be Markovian. This wider class of models is called *semi-Markovian*. For concreteness, we provide a simple SCM next.

Example 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

$$\mathcal{F} = \begin{cases} X & \leftarrow U_1 + U_2 \\ Y & \leftarrow U_1 - U_2 \end{cases}, \tag{1.2}$$

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

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

[8] In most of the literature the set $\mathbf{V}$ is assumed to be finite; however, the axiomatic characterization of SCMs can be extended to the infinitary setting in a very natural way [Ibeling and Icard 2019].

[9] 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, Halpern 1998] and [Halpern 2000].

values is observed. Here, the domains of X and Y are, respectively, $\text{Val}(X) = \{2,\dots,12\}$ and $\text{Val}(Y) = \{-5,\dots,0,\dots,5\}$. □

Pearl Hierarchy, Layer 1 – Seeing

Layer 1 of the hierarchy (Table 1.1) 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. Eq. (7.2) in [Pearl 2000]), as shown next.

Definition 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 $\mathbf{Y} \subseteq \mathbf{V}$:[10]

$$P^{\mathcal{M}}(\mathbf{y}) = \sum_{\{\mathbf{u} \,|\, \mathbf{Y}(\mathbf{u}) = \mathbf{y}\}} P(\mathbf{u}), \tag{1.3}$$

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

In words, the procedure dictated by Eq. (1.3) can be described as follows:

1. For each unit $\mathbf{U} = \mathbf{u}$, Nature evaluates $\mathcal{F}$ following a valid order (i.e., any variable in the l.h.s. is evaluated after the ones in the r.h.s.)[11], and
2. The probability mass $P(\mathbf{U} = \mathbf{u})$ is accumulated for each instantiation $\mathbf{U} = \mathbf{u}$ consistent with the event $\mathbf{Y} = \mathbf{y}$.

This evaluation is graphically depicted in Fig. 1.3(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 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 $Y = 1 - 1 = 0$. Since $P(U_1 = 1, U_2 = 1) = 1/36$ 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 $\langle Y = 0 \rangle$. Since 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, i.e., $P(Y = 0 \,|\, X = 2)$? The answer is one since 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). In fact, X and Y become deterministically related.

[10] 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.

[11] Here, we deliberately invoke the entity "Nature" as the evaluator of the SCM to emphasize the separation between the modeler/agent and the underlying dynamics of the system, which is almost invariably unknown to them.

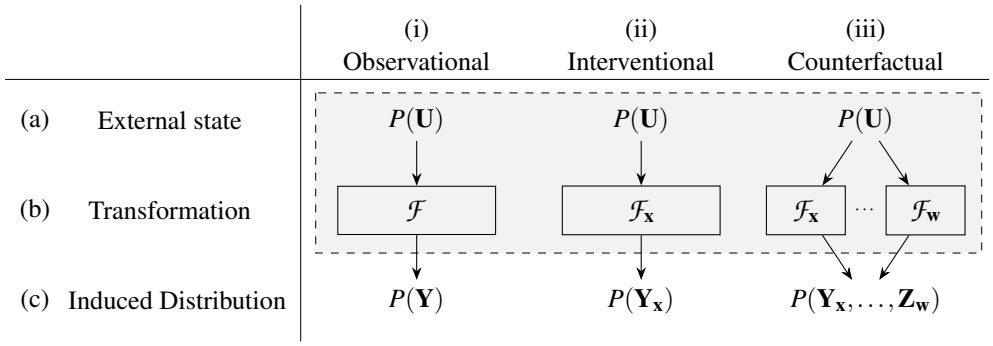


Figure 1.3: Given an SCM's initial state (i.e., population) (a), we show the different functional transformations (b) and the corresponding induced distribution (c) of each layer of the hierarchy. (i) represents the transformation (i.e., $\mathcal{F}$) from the natural state of the system ($P(\mathbf{U})$) to an observational world, (ii) to an interventional world (i.e., with modified mechanisms $\mathcal{F}_{\mathbf{x}}$), and (iii) to multiple counterfactual worlds (i.e., with multiple modified mechanisms).

Many tasks throughout data sciences can be seen as evaluating the probability of certain events occurring. For instance, expressions such as $P(\mathbf{Y} \mid \mathbf{X})$, with $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$, are naturally probabilistic reflecting the uncertainty we may have about the world. 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.

Example 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:

$$\mathcal{F} = \begin{cases} Z & \leftarrow \mathbb{1}_{\{U_r=1, U_z=1\}} \\ X & \leftarrow \mathbb{1}_{\{Z=1, U_x=1\}} + \mathbb{1}_{\{Z=0, U_x=0\}} \\ Y & \leftarrow \mathbb{1}_{\{X=1, U_r=1\}} + \mathbb{1}_{\{X=0, U_r=1, U_y=1\}} + \mathbb{1}_{\{X=0, U_r=0, U_y=0\}} \end{cases}. \tag{1.4}$$

	U_r	U_z	U_x	U_y	Z	X	Y	$P(\mathbf{u})$		U_r	U_z	U_x	U_y	Z	X	Y	$P(\mathbf{u})$
1	0	0	0	0	0	1	0	0.001125	9	1	0	0	0	0	1	1	0.000375
2	0	0	0	1	0	1	0	0.002625	10	1	0	0	1	0	1	1	0.000875
3	0	0	1	0	0	0	1	0.010125	11	1	0	1	0	0	0	0	0.003375
4	0	0	1	1	0	0	0	0.023625	12	1	0	1	1	0	0	1	0.007875
5	0	1	0	0	0	1	0	0.021375	13	1	1	0	0	1	0	0	0.007125
6	0	1	0	1	0	1	0	0.049875	14	1	1	0	1	1	0	1	0.016625
7	0	1	1	0	0	0	1	0.192375	15	1	1	1	0	1	1	1	0.064125
8	0	1	1	1	0	0	0	0.448875	16	1	1	1	1	1	1	1	0.149625

Table 1.2: Mapping of events in the space of **U** to **V** in the context of Example 2.

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 Def. 2 (Eq. 1.3) induces a mapping between $P(\mathbf{U})$ and $P(\mathbf{V})$. In this example, each entry of Table 1.2 corresponds to an event in the space of **U** and the corresponding realization of **V** according to the functions in $\mathcal{F}$.

Using this mapping and Def. 2, a query $P(Y = 1 \mid X = 1)$ can be evaluated from $\mathcal{M}$ as:

$$P(Y = 1 \mid X = 1) = \frac{P(Y = 1, X = 1)}{P(X = 1)} = \frac{\sum_{\{\mathbf{u} \mid Y(\mathbf{u})=1, X(\mathbf{u})=1\}} P(\mathbf{u})}{\sum_{\{\mathbf{u} \mid X(\mathbf{u})=1\}} P(\mathbf{u})}$$
$$= \frac{0.215}{0.29} = 0.7414, \tag{1.5}$$

which is just the ratio between the sum of the probabilities of the events in the space of **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 $P(Y = 1 \mid X = 0) = 0.3197$ 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 $P(Y = 1 \mid X = 0)$ and $P(Y = 1 \mid X = 1)$. We shall discuss this issue next. □

Pearl Hierarchy, Layer 2 – Doing

Layer 2 of the hierarchy (Table 1.1) 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 3 (Submodel – "Interventional SCM")**.** Let $\mathcal{M}$ be a causal model, **X** a set of variables in **V**, and **x** a particular realization of **X**. A submodel $\mathcal{M}_{\mathbf{x}}$ of $\mathcal{M}$ is the causal model

$$\mathcal{M}_{\mathbf{x}} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}_{\mathbf{x}}, P(\mathbf{U}) \rangle, \tag{1.6}$$

where

$$\mathcal{F}_{\mathbf{x}} = \{f_i : V_i \notin \mathbf{X}\} \cup \{\mathbf{X} \leftarrow \mathbf{x}\}. \tag{1.7}$$

■

In words, performing an external intervention (or action) is modelled 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.[12,13] The impact of the intervention on an outcome variable Y is called *potential response* (cf. Def. (7.1.4) in [Pearl 2000]):

Definition 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 $\mathbf{Y}_{\mathbf{x}}(\mathbf{u})$ 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, $\mathbf{Y}_{\mathcal{M}_{\mathbf{x}}}(\mathbf{u})$). That is, $\mathbf{Y}_{\mathbf{x}}(\mathbf{u}) = \mathbf{Y}_{\mathcal{M}_{\mathbf{x}}}(\mathbf{u})$. ■

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

Definition 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 $\mathbf{Y} \subseteq \mathbf{V}$:

$$P^{\mathcal{M}}(\mathbf{y}_{\mathbf{x}}) = \sum_{\{\mathbf{u} | \mathbf{Y}_{\mathbf{x}}(\mathbf{u}) = \mathbf{y}\}} P(\mathbf{u}). \tag{1.8}$$

■

The evaluation implied by Eq. 1.8 can be described as the following process (Fig. 1.3(ii)):

1. Replace the mechanism of each $X \in \mathbf{X}$ with the corresponding constants $\mathbf{x}$ generating $\mathcal{F}_{\mathbf{x}}$ (Eq. 1.7), which induces a submodel $\mathcal{M}_{\mathbf{x}}$ (of $\mathcal{M}$);
2. For each unit $\mathbf{U} = \mathbf{u}$, Nature evaluates $\mathcal{F}_{\mathbf{x}}$ following a valid order (where any variable in the l.h.s. is evaluated after the ones in the r.h.s.), and
3. The probability mass $P(\mathbf{U} = \mathbf{u})$ is then accumulated for each instantiation $\mathbf{U} = \mathbf{u}$ consistent with the event $\mathbf{Y}_{\mathbf{x}} = \mathbf{y}$ (i.e., $\mathbf{Y}$ in the submodel $\mathcal{M}_{\mathbf{x}}$).

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 $P(\mathbf{U})$.[14] Further, note that this procedure disconnects X from any other source of "natural" variation when it follows the original function f_x

[12] 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], and [Marschak 1950, Simon 1953]. It was then made more explicit and called "wiping out" by [Strotz and Wold 1960]; see also [Fisher 1970] and [Sobel 1990].

[13] Pearl credits his realization on the connection of this operation with graphical models to a lecture of Peter Spirtes at the International Congress of Philosophy of Science (Uppsala, Sweden, 1991), in his words [Pearl 2000, pp. 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 (...)". To avoid confusion, we note that the sense of manipulation here is not intended to be reductive, e.g., see [Pearl 2018b, 2019, Woodward 2016].

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

(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.[15] 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, since X itself is a constant, not affected by other variables. Indeed, the notion of *average causal effect* can be formally written as $E(Y_{X=1}) - E(Y_{X=0})$.[16]

The distribution $P(\mathbf{Y_x})$ defined in Eq. (1.8) is often written $P(\mathbf{Y} \mid do(\mathbf{x}))$, and we henceforth adopt this notation in the context of PCH's second layer.[17] By convention, the $do(\mathbf{x})$ maps over the entire formula in the conditional case, i.e.:

$$P(\mathbf{Y} \mid do(\mathbf{x}), \mathbf{z}) = P(\mathbf{Y_x} \mid \mathbf{z_x}) = \frac{P(\mathbf{Y_x}, \mathbf{z_x})}{P(\mathbf{z_x})} = \frac{P(\mathbf{Y}, \mathbf{z} \mid do(\mathbf{x}))}{P(\mathbf{z} \mid do(\mathbf{x}))}, \tag{1.9}$$

where the last expression is manifestly $\mathcal{L}_2$. Further, in accordance with the semantics of the intervention $do(\mathbf{x})$, it is clear that the distribution of the intervened variables must satisfy a property called *effectiveness*, which we define explicitly next:

Definition 6 (Effectiveness)**.** A joint interventional distribution $P(\mathbf{v}|do(\mathbf{x}))$ is said to satisfy *effectiveness* if for every $V_i \in \mathbf{X}$,

$$P(v_i \mid do(\mathbf{x})) = 1 \text{ if } v_i \text{ is consistent with } \mathbf{x} \text{ and 0 otherwise.} \tag{1.10}$$

■

In words, if a variable X is fixed to x by intervention, $X = x$ must be observed with probability one. This is a technical condition and reflects the probabilistic meaning of hard interventions.[18] (For a comparison against Bayesian conditioning, see [Pearl 2017].)

Example 3 (*Example 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 $\mathcal{M}_{X=2}$:

$$\mathcal{F}_{X=2} = \begin{cases} X & \leftarrow 2 \\ Y & \leftarrow U_1 - U_2, \end{cases}, \tag{1.11}$$

while $P(\mathbf{U})$ remains invariant. It is immediate to see that $Y_{X=2}(u_1, u_2)$ is the same as $Y(u_1, u_2)$; in words, misreporting the sum of the two dice will of course not change their difference. This,

[15] For a discussion of what it means for these changes to arise "from outside" the system, see, e.g., [Woodward 2003, 2016]. Of course, in many settings this simply means the intervention is performed deliberately by an *agent* outside the system, for example, in reinforcement learning [Sutton and Barto 2018].

[16] 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 $Y_{X=1} \leftarrow f_Y(1, \varepsilon_Y)$ when $do(X = 1)$.

[17] 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.

[18] For the sake of presentation, we discuss the class of atomic interventions even though there are more general within the SCM framework, including soft, conditional, stochastic [Correa and Bareinboim 2020, Pearl 2000, Ch. 4].

in turn, entails the following probabilistic invariance,

$$P(Y = 0 \mid do(X = 2)) = P(Y = 0). \tag{1.12}$$

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 (Eq. 1.12); as we say, X does not have a *causal effect* on Y. □

Example 4 (*Example 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 words, patients do not have an option of deciding their own treatment, but are compelled to take the specific drug.[19] This is represented through the new modified set of mechanisms,

$$\mathcal{F}_{X=1} = \begin{cases} Z & \leftarrow \mathbb{1}_{\{U_r=1,U_z=1\}} \\ X & \leftarrow 1 \\ Y & \leftarrow \mathbb{1}_{\{X=1,U_r=1\}} + \mathbb{1}_{\{X=0,U_r=1,U_y=1\}} + \mathbb{1}_{\{X=0,U_r=0,U_y=0\}} \end{cases}, \tag{1.13}$$

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 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

$$P(Y{=}1|do(X{=}1)) = \sum_{\{\mathbf{u}|Y_{X=1}(\mathbf{u})=1\}} P(\mathbf{u}) \tag{1.14}$$

$$= \sum_{\{u_r|Y_{X=1}(u_r)=1\}} P(u_r) = P(U_r{=}1) = 0.25. \tag{1.15}$$

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

[19] 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. This procedure is pervasive in modern society, for example, in randomized controlled trials (RCTs) when drugs are evaluated for their efficacy, or in A/B experiments when products are tested by internet companies.

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 Fig. 1.3(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) will be a central topic in causal inference, as discussed more substantively in Section 1.4.

Pearl Hierarchy, Layer 3 – Imagining counterfactual worlds

Layer 3 of the hierarchy (Table 1.1) allows operationalizing the notion of "imagination" (and the closely related activities of retrospection, introspection, 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, 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 structural causal model, as described in the sequel:

Definition 7 (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 $\mathbf{Y}_{\mathbf{x}}, \ldots, \mathbf{Z}_{\mathbf{w}}$, for any $\mathbf{Y}, \mathbf{Z}, \ldots, \mathbf{X}, \mathbf{W} \subseteq \mathbf{V}$:

$$P^{\mathcal{M}}(\mathbf{y}_{\mathbf{x}}, \ldots, \mathbf{z}_{\mathbf{w}}) = \sum_{\substack{\{\mathbf{u} \mid \mathbf{Y}_{\mathbf{x}}(\mathbf{u})=\mathbf{y}, \\ \ldots, \mathbf{Z}_{\mathbf{w}}(\mathbf{u})=\mathbf{z}\}}} P(\mathbf{u}). \tag{1.16}$$

■

Note that the l.h.s. of Eq. (1.16) contains variables with different subscripts, which, syntactically, encode different counterfactual "worlds." The evaluation implied by this equation can be described as the following process (as shown in Fig. 1.3(iii)):

1. For each set of subscripts relative to each set of variables (e.g., $\mathbf{X}$, ..., $\mathbf{W}$ for $\mathbf{Y}$, ..., $\mathbf{Z}$, respectively), replace the corresponding mechanisms with the appropriate constants and generate $\mathcal{F}_{\mathbf{x}}$, ..., $\mathcal{F}_{\mathbf{w}}$ (Eq. 1.7), creating submodels $\mathcal{M}_{\mathbf{x}}$, ..., $\mathcal{M}_{\mathbf{w}}$ (of $\mathcal{M}$);
2. For each unit $\mathbf{U} = \mathbf{u}$, Nature evaluates the modified mechanisms (e.g., $\mathcal{F}_{\mathbf{x}}$, ..., $\mathcal{F}_{\mathbf{w}}$) following a valid order (i.e., any variable in the l.h.s. is evaluated after the ones in the r.h.s.) to obtain the potential responses of the observables, and
3. The probability mass $P(\mathbf{U} = \mathbf{u})$ is then accumulated for each instantiation $U = u$ that is consistent with the events over the counterfactual variables – for instance, $\mathbf{Y}_{\mathbf{x}} = \mathbf{y}$, ..., $\mathbf{Z}_{\mathbf{w}} = \mathbf{z}$, i.e., $\mathbf{Y} = \mathbf{y}$, ..., $\mathbf{Z} = \mathbf{z}$ in the submodels $\mathcal{M}_{\mathbf{x}}$, ..., $\mathcal{M}_{\mathbf{w}}$, respectively.

Example 5 (*Example 2 continued*)**.** Since 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 $P(Y_{X=1} = 1 \mid X = 0, Y = 0)$. This is a non-trivial question since 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 $P(Y_{X=1} = 1|X = 0, Y = 0)$, since 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 since 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 and cannot be written in the languages of layers 1 and 2 since they do not allow for probability expressions involving more than one subscript (each encoding a different world). This means we need to climb up to the third layer in order to formally specify the quantity of interest. Using the procedure dictated in Eq. 1.16, we obtain

$$P(Y_{X=1}{=}1|X{=}0, Y{=}0) = \frac{P(Y_{X=1}{=}1, X{=}0, Y{=}0)}{P(X{=}0, Y{=}0)} = \frac{\sum_{\{\mathbf{u}|Y_{X=1}(\mathbf{u})=1, X(\mathbf{u})=0, Y(\mathbf{u})=0\}} P(\mathbf{u})}{\sum_{\{\mathbf{u}|X(\mathbf{u})=0, Y(\mathbf{u})=0\}} P(\mathbf{u})}$$
$$= \frac{0.0105}{0.483} = 0.0217. \tag{1.17}$$

This evaluation is shown step by step in [Bareinboim et al. 2020a, Appendix E.1], 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 Fig. 1.3(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 death. 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 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. □

The probability of necessity, $P(y'_{x'} \mid x, y)$, encodes how a disease (Y) is "attributable" to a particular exposure (X), interpreted counterfactually as "the probability that disease would not have occurred in the absence of exposure, given that disease and exposure did in fact occur." Conversely, the probability of sufficiency, $P(y_x \mid x', y')$, captures how a certain exposure might impact the healthy population, counterfactually, "whether a healthy unexposed individual would have contracted the disease had they been exposed." 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, Ch. 9], counterfactually 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, that is, the probability that Y would respond to X in both of the ways described above. 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 words, for Y the value of X is x, while for M the value of X is x'. This type of nested counterfactual allows 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 natural and general treatment of direct and indirect effects, including a precise understanding of their relationship in non-linear systems, epitomized through what is known as the *mediation formula* [Pearl 2012, VanderWeele 2015].[20] Overall, counterfactual statements are central to our ability to explain how and why certain events come about in the world. Indeed, explanations have practical ramifications for credit assignment, the determination of blame and responsibility, the analysis of biases and unfairness in decision-making, and, more broadly, the systematic understanding of the world around us.

1.3 Pearl Hierarchy – A Logical Perspective

We have thus far learned that each layer of the PCH corresponds to a different intuitive notion in human cognition: seeing, acting, and imagining. Table 1.1 presents characteristic questions associated with each of the layers. Layer 1 concerns questions like, "How likely is Y given that I have observed X?" In Layer 2 we can ask hypothetical ("conditional") questions such as, "How likely *would* Y be if one were to make it the case that X?" Layer 3 takes us even 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

[20] This result can be generalized to disentangle $X - Y$ variations of any nature, including direct, indirect, and also spurious [Zhang and Bareinboim 2018c]. This treatment led to the *explanation formula* [Zhang and Bareinboim 2018a,b], which has implications for fairness analysis and attribution in the context of non-manipulable variables.

(Defs. 2, 5, 7). 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 explicitly. We define three languages $\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3$, each based on polynomials built over basic probability terms $P(\alpha)$, the only difference among them being which terms $P(\alpha)$ we allow: as we go up in the PCH, we allow increasingly complex expressions α to appear in the probability terms. In particular, $\mathcal{L}_1$ is just a familiar probabilistic logic (see, e.g., [Fagin et al. 1990]).

Definition 8 (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, $\mathrm{P}(\mathbf{Y}_\mathbf{x}{=}\mathbf{y}, \ldots, \mathbf{Z}_\mathbf{w}{=}\mathbf{z})$, 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

$$\begin{aligned} &P(X = 1, Y = 1) = P(X = 1)P(Y = 1) \wedge P(X = 1, Y = 0) = P(X = 1)P(Y = 0) \\ &\wedge P(X = 0, Y = 1) = P(X = 0)P(Y = 1) \wedge P(X = 0, Y = 0) = P(X = 0)P(Y = 0) \end{aligned} \quad (1.18)$$

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 $3/4$.[21] As before, we could also write this as $P(Y = 1 \mid do(X = 1)) = 3/4$. Finally, $\mathcal{L}_3$ allows statements about joint probabilities over conditional terms with possibly inconsistent subscripts (also known as antecedents in logic). For instance, $P(y_x, y'_{x'}) \geq P(y \mid x) - P(y \mid x')$ is a statement expressing a lower bound on the probability of necessity and sufficiency.[22]

[21] 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 $X = 1 \,\square\!\!\rightarrow\, Y = 1$ (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 and Lewis's "system-of-spheres" interpretation, we refer readers to [Pearl 2000, § 7.4.1-7.4.3] and [Briggs 2012, Halpern 2013, 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].

[22] For details of this bound and the assumptions guaranteeing it, see [Pearl 2009, Thm. 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 $P(y_x, y'_{x'})P(x)P(x') + P(x', y)P(x) \geq P(x, y)P(x')$.

Def. 8 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 Defs. 2, 5, 7. 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 (1.18), comes out true or false.[23] At this stage, we are ready to formally define the Pearl Causal Hierarchy:

Definition 9 (Pearl Causal Hierarchy)**.** 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 Defs. 2, 5, 7 (semantics), is called the Pearl Causal Hierarchy (PCH, for short). ■

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, i.e., 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., $P(y_x)$), 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?[24] 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 $\mathcal{L}_i$-

[23] 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 Def. 9 and Thm. 1 (below) – can be cast in axiomatic terms, although these results do not depend in any direct way on questions of axiomatization.

[24] 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.)

theories coincide, i.e., 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.[25] 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, e.g., the exact value of $P(y_x \mid y'_{x'})$ – 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 $\mathcal{M}$.[26] 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 happen in exceptional circumstances, for instance, if the functional relationship is deterministic:

Example 6. Consider the SCM $\mathcal{M}^* = \langle \mathbf{U} = \{U\}, \mathbf{V} = \{X, Y\}, \mathcal{F}, P(U) \rangle$, where

$$\mathcal{F} = \begin{cases} X & \leftarrow U \\ Y & \leftarrow X \end{cases}, \tag{1.19}$$

and U is distributed as a fair coin flip. Then, there is no model $\mathcal{M}$ that agrees with $\mathcal{M}^*$ on all $\mathcal{L}_1$ and $\mathcal{L}_2$ valuations, but disagrees on $\mathcal{L}_3$. To see why, consider any model $\mathcal{M}$ with structural functions f_X, f_Y. We must have $f_Y(x,u) = x$ for any unit u, or else the $\mathcal{L}_2$-probability $P(y_x)$ differs between $\mathcal{M}, \mathcal{M}^*$. The function f_X is also easily determined by the $\mathcal{L}_1$ requirement that $P(x)$ match between $\mathcal{M}, \mathcal{M}^*$. This is enough to determine all $\mathcal{L}_3$ quantities. □

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 1.2), 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. 2020a] 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.[27] 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 $\mathcal{L}_3$.

[25] For readers familiar with causal inference, this can be seen as a generalization of the notion of identifiability (e.g., see [Pearl 2000, Def. 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, viz. Padoa's method in the theory of definability [Beth 1956].

[26] 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.

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

Does something similar happen with Layers $1, 2, 3$ 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 10 (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 $\mathcal{M} \in \Omega$. [28] ■

In the Example 6 above, we would say that Layer 3 collapses to Layer 2. 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 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.* ■

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 thus says in a general manner that there will typically be causal questions that one cannot answer with knowledge and/or data restricted to a lower layer in the hierarchy.[29] In fact, this can be seen as the formal grounding for the intuition behind the PCH as discussed in [Pearl and Mackenzie 2018, Ch. 1]:

Corollary 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

[28] 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.

[29] The CHT is not to be understood as a general impossibility result for causal inferences, quite the contrary, as will become clear in the rest of this chapter. In fact, some of the most celebrated results in the field are precisely about conditions under which these inferences are allowed from lower layer data together with minimal assumptions about the underlying SCM. For instance, 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, e.g., [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) [Peters et al. 2017, Spirtes et al. 2001, Zhang 2008].

SCMs. The theorem then states (for the complete proof and further details, we refer readers to [Bareinboim et al. 2020a, Appendix A]):

Theorem 1. *[Causal Hierarchy Theorem (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 Defs. 2, 5, 7.

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 $\mathcal{L}_1$, the second showing that $\mathcal{L}_3$ almost never collapses to $\mathcal{L}_2$. (It easily follows that neither collapse occurs, almost always.) 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 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. 2020a, 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. Examples 7 and 8 below demonstrate two quite different strategies. 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. 2020a, Lemma 3]). 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 2 (Example 9 below).

Turning now to these examples, we start with a variation of a classic construction presented by Pearl himself [Pearl 2009, §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 1).

Example 7. Let $\mathcal{M}^* = \langle \mathbf{U} = \{U_1, U_2\}, \mathbf{V} = \{X, Y\}, \mathcal{F}^*, P(U) \rangle$, where

$$\mathcal{F}^* = \begin{cases} X & \leftarrow U_1 \\ Y & \leftarrow U_2 \end{cases}. \tag{1.20}$$

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:

$$\mathcal{F} = \begin{cases} X & \leftarrow \mathbb{1}_{U_1 = U_2} \\ Y & \leftarrow U_1 + \mathbb{1}_{X=1, U_1=0, U_2=1} \end{cases}. \tag{1.21}$$

Then $P^{\mathcal{M}^*}(X, Y) = P^{\mathcal{M}}(X, Y)$, yet $P^{\mathcal{M}^*}(Y = 1 \mid do(X = 1)) = 1/2$ since X does not affect Y in $\mathcal{M}^*$, while $P^{\mathcal{M}}(Y = 1 \mid do(X = 1)) = 3/4$. 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, $P(Y \mid do(X)) = P(Y)$ – 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 $P(Y_{X=0} = 1 \mid X = 1, Y = 0)$, as discussed in Example 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

$$\mathcal{F}' = \begin{cases} X & \leftarrow U_1 \\ Y & \leftarrow XU_2 + (1 - X)(1 - U_2) \end{cases}. \tag{1.22}$$

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 – see Table 1.3. The dependence only becomes visible at Layer 3. In $\mathcal{M}^*$, we have $P^{\mathcal{M}^*}(Y_{X=0} = 1 \mid X = 1, Y = 0) = 0$, whereas in $\mathcal{M}'$ we have the exact opposite pattern, $P^{\mathcal{M}'}(Y_{X=0} = 1 \mid X = 1, Y = 0) = 1$. 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

			$\mathcal{M}^*$			$\mathcal{M}'$			
U_1	U_2	X	Y	$Y_{X=0}$	$Y_{X=1}$	Y	$Y_{X=0}$	$Y_{X=1}$	$P(\mathbf{u})$
0	0	0	0	0	0	1	1	0	1/4
0	1	0	1	1	1	0	0	1	1/4
1	0	1	0	0	0	0	1	0	1/4
1	1	1	1	1	1	1	0	1	1/4

Table 1.3: Counterfactual evaluation of $\mathcal{M}^*$ and $\mathcal{M}'$ in Example 7.

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 7 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.

Example 8 (*Example 1 continued*)**.** Let SCM $\mathcal{M}^* = \mathcal{M}^1$ from Example 1, and consider another model $\mathcal{M}$ such that $Y \leftarrow (U_1 - U_2)$ as before, but now $X \leftarrow (2U_1 - Y)$. It is then easy to see that $P^{\mathcal{M}^*}(X,Y) = P^{\mathcal{M}}(X,Y)$, that is, $\mathcal{M}^* \sim_1 \mathcal{M}$. However, at Layer 2 we have $P^{\mathcal{M}^*}(X = 10 \mid do(Y = 0)) = P^{\mathcal{M}^*}(X = 10) = \frac{1}{12} \neq \frac{1}{6} = P^{\mathcal{M}}(X = 10 \mid do(Y = 0))$. There is no causal relationship between X and Y in $\mathcal{M}^*$, while Y exerts a clear influence on X in $\mathcal{M}$, one that is simply not visible at Layer 1.

In a similar vein, we can find another SCM $\mathcal{M}'$ with $\mathcal{M}^* \sim_2 \mathcal{M}'$ but $\mathcal{M}^* \not\sim_3 \mathcal{M}'$. In $\mathcal{M}'$, let us add new exogenous variables $U'_{-5}, \ldots, U'_0, \ldots, U'_5$, with each U_k distributed as $P^{\mathcal{M}^*}(X \mid (U_1 - U_2) \neq k)$, and let:

$$\mathcal{F}' = \begin{cases} X & \leftarrow \mathbb{1}_{Y=(U_1-U_2)}(U_1+U_2) + \sum_{-5\leq k\leq 5} \mathbb{1}_{Y=k\neq(U_1-U_2)} U_k \\ Y & \leftarrow U_1 - U_2 \end{cases}. \tag{1.23}$$

Then, $\mathcal{M}^* \sim_1 \mathcal{M}'$ – in fact, $\mathcal{M}^* \sim_2 \mathcal{M}'$, since $P^{\mathcal{M}'}(X_{Y=k}) = P^{\mathcal{M}^*}(X, (U_1 - U_2) = k) + P^{\mathcal{M}^*}(X, (U_1 - U_2) \neq k) = P^{\mathcal{M}^*}(X) = P^{\mathcal{M}^*}(X_{Y=k})$. However, e.g., $P^{\mathcal{M}^*}(X_{Y=0} = 7 \mid Y = 5) = P^{\mathcal{M}^*}(X = 7 \mid U_1 = 6, U_2 = 1) = 1 \neq \frac{1}{5} = P^{\mathcal{M}^*}(X = 7 \mid U_1 \neq U_2) = P^{\mathcal{M}'}(X_{Y=0} = 7 \mid Y = 5)$. □

Finally, for this last illustration we see two quite different patterns. 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. 2020a, Lemma 3].

Example 9 (*Example 2 continued*)**.** For the SCM $\mathcal{M}^* = \mathcal{M}^2$ of Example 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 4 we saw that $P^{\mathcal{M}^*}(Y \mid X) \neq P^{\mathcal{M}^*}(Y \mid do(X))$, here we have $P^{\mathcal{M}}(Y \mid X) = P^{\mathcal{M}}(Y \mid do(X))$. 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 $\mathrm{Val}(U) = \{0,1\}^4 \cup \{u_1^*, u_2^*\}$. 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 $\mathcal{M}^*$ — $P(U_r = u_r) \cdot P(U_z = u_z) \cdot P(U_x = u_x) \cdot P(U_y = u_y)$ — as calculated explicitly in Table 1.2, 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 write, e.g., $U_r = u$ 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 causal mechanisms can now be written:

$$\mathcal{F}' = \begin{cases} Z \leftarrow \mathbb{1}_{\{U_r=1,U_z=1\}} + \mathbb{1}_{U \in \{u_1^*, u_2^*\}} \\ X \leftarrow \mathbb{1}_{\{Z=1,U_x=1\}} + \mathbb{1}_{\{Z=0,U_x=0\}} + \mathbb{1}_{U=u_2^*} \\ Y \leftarrow \mathbb{1}_{\{X=1,U_r=1\}} + \mathbb{1}_{\{X=0,U_r=1,U_y=1\}} + \mathbb{1}_{\{X=0,U_r=0,U_y=0\}} + \mathbb{1}_{\{X=1,U \in \{u_1^*, u_2^*\}\}} \end{cases} . \quad (1.24)$$

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 $Z = X = Y = 1$ (recall Table 1.2), 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, $P^{\mathcal{M}'}(Y_{Z=1} = 1, Y_{Z=0} = 1) = P^{\mathcal{M}^*}(Y_{Z=1} = 1, Y_{Z=0} = 1) + \varepsilon$. □

Whereas Example 6 shows that collapse of the layers is possible if $\mathcal{M}^*$ is exceptional, the CHT shows that this is the exception indeed. Typical cases are like Examples 7, 8, 9, each showing a different way of perturbing an SCM to obtain a second SCM revealing non-collapse. As these examples demonstrate, 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 – since 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*.

1.4 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, we usually cannot determine the structural model at this level of precision in most practical settings, and the CHT severely curtails the ability to "climb up" the PCH via lower-level data. Still, learning about cause and effect relationships is arguably one of the main goals found throughout the empirical sciences. How might we progress?

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. In particular, our goal here is to understand the conditions under which valid causal claims can be made (following the semantics given by the SCM) when the data scientist does not have direct access to the underlying causal mechanisms (the SCM itself), but only some features and data thereof. For instance, one typical 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 Fig. 1.4(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 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 Fig. 1.4(d) as *structural constraints*.

1.4.1 Causal Inference via $\mathcal{L}_1$-constraints

Following this discussion, a first attempt to solve our 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

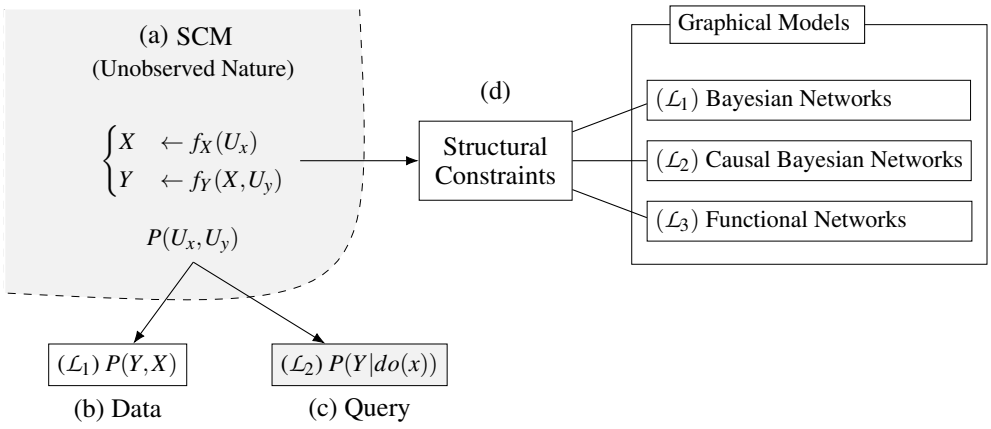


Figure 1.4: Example of Prototypical Causal Inference – on top the SCM itself, representing the unobserved collection of mechanisms and corresponding uncertainty (a); on the bottom, the different probability distributions entailed by the model (b,c); on the right side, the graphical model representing the specific constraints of the SCM (d).

used to encode this type of knowledge.[30] The example below shows that $\mathcal{L}_1$ constraints (and BNs) alone are insufficient to support causal reasoning in general.

Example 10. 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,

$$\mathcal{F}_1 = \begin{cases} X & \leftarrow U_x \\ Z & \leftarrow X \oplus U_z\,, \\ Y & \leftarrow Z \oplus U_y \end{cases} \qquad \mathcal{F}_2 = \begin{cases} X & \leftarrow Z \oplus U_x \\ Z & \leftarrow Y \oplus U_z\,, \\ Y & \leftarrow U_y \end{cases} \tag{1.25}$$

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 is immediate to see (via Def. 2 and Eq. (1.3)) that both models generate the same observational distribution,

$$P^{1,2}(X = 0,Z = 0,Y = 0) = P^{1,2}(X = 1,Z = 1,Y = 1) = (1-a)(1-b)/2,$$
$$P^{1,2}(X = 0,Z = 0,Y = 1) = P^{1,2}(X = 1,Z = 1,Y = 0) = (1-a)b/2,$$

[30] This formalism is especially effective for encoding conditional independences in a parsimonious way, capable of spanning the (exponentially many) constraints implied by it. This encoding leads to efficiency gains in the context of probabilistic reasoning in terms of both computational and sample complexity. Due to space constraints, we refer readers to [Bareinboim et al. 2020a, Appendix C], where a more detailed account of the semantics and construction of $\mathcal{L}_1$-models is provided. Understanding the relationship between a fully specified SCM and its corresponding BN should be helpful for comprehending the nature of the constraints and characteristic models of the other layers.

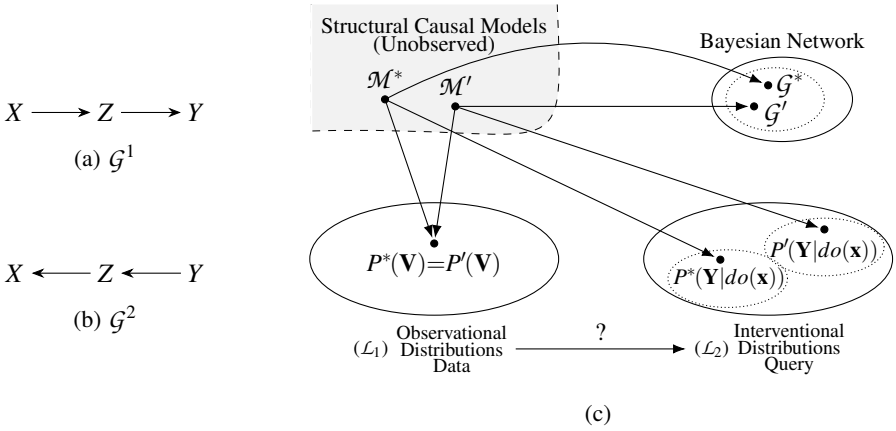


Figure 1.5: Two causal diagrams encoding knowledge about the causal mechanisms governing three observable variables X, Z and Y. In (a) X is an argument to f_Z, and Z an argument to f_Y. In (b) the opposite is true. In (c), schema representing the impossibility of identifying causal queries from $\mathcal{L}_1$ data, constraints, and graphical models.

$$P^{1,2}(X=0,Z=1,Y=1) = P^{1,2}(X=1,Z=0,Y=0) = a(1-b)/2,$$
$$P^{1,2}(X=0,Z=1,Y=0) = P^{1,2}(X=1,Z=0,Y=1) = ab/2. \tag{1.26}$$

We further compute the effect of the intervention $do(x)$ (via Def. 5 and Eq. 1.8),

$$P^1(Y=1 \mid do(X=1)) = ab + (1-a)(1-b),$$
$$P^2(Y=1 \mid do(X=1)) = 1/2, \tag{1.27}$$

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 Figs. 1.5(a) and (b).[31] 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, $X \perp\!\!\!\perp Y \mid Z$) and nothing more.[32] 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 Eqs. (1.27).

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 $P(\mathbf{V})$. And yet, each such model could entail a different causal effect. This suggests that, in general,

[31] This construction follows from the order in which the functions are determined in the SCM, systematized in Def. 24([Bareinboim et al. 2020a, 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. 2020a, Thm. 8, Appendix C]).

[32] We refer readers to [Bareinboim et al. 2020a, Appendix C, Definition 23], 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.

causal inference cannot be carried out with mere $\mathcal{L}_1$ objects – the observational distribution, its constraints, and corresponding models (Bayesian networks). This result can be seen as a graphical instantiation of Corollary 1 and is schematically summarized in Fig. 1.5(c). □

It is sometimes believed in the literature that Bayesian networks might provide a basis for causal reasoning, but as this example suggests, such an assumption is unfounded even for the simplest models. Pearl himself acknowledged this impossibility, which was one of the motivating factors for his journey towards the principles and additional constraints needed to support causal inference; in his own words [Pearl 2018a, pp. 7]: "I had been under the impression that probabilities could express every aspect of human knowledge, even causality, but they can't. After I realized this, my research shifted completely to a new direction, as I tried to understand how to represent causal knowledge and draw causal conclusions." He further hints at the importance of causal Bayesian networks, the very topic of the next section: "In retrospect, fighting for the acceptance of Bayesian networks was a picnic – no, a luxury cruise! – compared with the fight I had to go through for causal Bayesian networks."

1.4.2 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 (Fig. 1.4(d)) imprinted by the underlying SCM could license causal inferences? To answer this question, it is instructive to compare more closely the effect of an intervention $do(X = 1)$ in the two SCMs from Example 10. First, note that the function f_Y does not depend on X in the submodel $\mathcal{M}^2_{X=1}$ (constructed following Def. 3), so, probabilistically, Y will not depend on X. This implies the following relationship between distributions,

$$P^2(Y = 1 \mid do(X = 1)) = P(Y = 1), \tag{1.28}$$

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

$$P^1(Y = 1 \mid do(X = 1)) = P(Y = 1 \mid X = 1). \tag{1.29}$$

The exact computation of Eqs. (1.28) and (1.29) follows immediately from Defs. 2 and 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 (but not mere conditional independences) and can further be pieced together, and given a graphical interpretation. Consider again the Eq. (1.28) 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 11 (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 $(V_j \rightarrow V_i)$ for every $V_i \in \mathbf{V}$ if V_j appears as an argument of $f_i \in \mathcal{F}$. ■

The procedure encapsulated in Def. 11 is central to the elicitation of the knowledge necessary to perform causal inference (Fig. 1.4(d)). Intuitively, $\mathcal{G}$ has an arrow from A to B ($A \rightarrow B$) 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, pp. 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.[33,34]

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 12 (Causal Bayesian Network (CBN - Markovian))**.** Let $\mathbf{P}_*$ be the collection of all interventional distributions $P(\mathbf{V} \mid do(\mathbf{x}))$, $\mathbf{X} \subseteq \mathbf{V}$, $\mathbf{x} \in \text{Val}(\mathbf{X})$, 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 *Causal Bayesian Network* 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 $\mathcal{G}$.

[33] 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.

[34] Here is where we depart ways with the tradition in, e.g., the physical sciences, of focusing on the precise functional relationships among variables, and move to a more relaxed approach where only the qualitative relationships between them are invoked; obviously, the former is more powerful than the latter, even though not always realizable.

(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 $\mathcal{G}$:

$$P(v_i \mid do(pa_i), do(\mathbf{x})) = P(v_i \mid do(pa_i)). \tag{1.30}$$

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

$$P(v_i \mid do(\mathbf{x}), do(pa_i)) = P(v_i \mid do(\mathbf{x}), pa_i). \tag{1.31}$$

■

The first condition requires the graph to be *Markov relative*[35] to every interventional distribution $P(\mathbf{V} \mid do(\mathbf{X} = \mathbf{x}))$ that holds if every variable is independent of its non-descendants given its parents.[36] The second condition, missing-link, encapsulates the type of constraint exemplified by Eq. (1.28): after fixing the parents of a variable by intervention, the corresponding function should be insensitive to any other intervention elsewhere in the system. In 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 Eq. (1.29): 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.[37]

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 (Def. 12) 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 ii conditions (ii) and (iii).

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 Bayesian networks ($\mathcal{L}_1$), each missing arrow corresponds to a conditional independence, a probabilistic type of invariance.[38] 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 Def. 12 through cond. (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

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

[36] In some accounts of causation, this condition is known as the *causal Markov condition* (CMC), and it usually phrased in terms of "causal" parents. We invite the reader to check that conditions (ii), (iii) are in no way implied by (i). One could in fact see Def. 12 as offering a precise characterization of what CMC formally means.

[37] Conds. (ii) and (iii) are equivalent to a condition called *modularity* in [Pearl 2000, pp. 24], i.e., $P(v_i|do(\mathbf{x}), pa_i) = P(v_i|pa_i)$. These definitions were shown equivalent to the truncated product [Bareinboim et al. 2012].

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

SCM – i.e., the collection of observational and interventional distributions (Def. 5) – indeed matches the content of the CBN (Def. 11), as defined above.

Theorem 2. *[$\mathcal{L}_2$-Connection – SCM-CBN (Markovian)] The causal diagram $\mathcal{G}$ induced by the SCM $\mathcal{M}$ (following the constructive procedure in Def. 11) is a CBN for $\mathbf{P}_*^{\mathcal{M}}$ – the collection of observational and experimental distributions induced by $\mathcal{M}$.* ■

As this result demonstrates, CBNs serve as proxies for SCMs in terms of the observed $\mathcal{L}_2$ distributions. [39] 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.[40]

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, i.e., any interventional distribution is computable from $\mathcal{L}_1$-data.

Theorem 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,*

$$P(\mathbf{v} \mid do(\mathbf{x})) = \prod_{\{i|V_i \notin \mathbf{X}\}} P(v_i \mid pa_i)\Big|_{\mathbf{X}=\mathbf{x}}. \tag{1.32}$$

■

In other words, the interventional distribution in the l.h.s. of Eq. (1.32) can be expressed as the product given in the r.h.s. 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, Eq. 1.37]).[41] Obviously, any marginal distribution of interest can be obtained by summing out the irrelevant factors, including the causal effect of X on Y.

Corollary 2 (Back-door Criterion (Markovian))**.** *In Markovian models (i.e., models without unobserved confounding), for any treatment $\mathbf{X}$ and outcome $\mathbf{Y}$, the interventional distribution*

[39] It can also be shown that there is an SCM inducing $\mathbf{P}_*$ and $\mathcal{G}$ for any CBN $(\mathcal{G}, \mathbf{P}_*)$ of a Markovian model.

[40] In fact, one could take an axiomatic view of the CBNs and consider alternative ways for satisfying their conditions, detached from the structural semantics. It is conceivable that other constructions that do not use the "listening" (i.e., functional) metaphor as a building block, the signature of structural models, may also entail the same conditions of a CBN, therefore, supporting valid cross-layer (causal) inferences. Due to space constraints, we present Algorithm 10 in [Bareinboim et al. 2020a], which constitutes one such alternative supporting a purely experimental construction of a CBN. This procedure can be seen as the probabilistic-empirical counterpart of the SCM-functional Def. 11; see also Thm. 10 in the corresponding Appendix. The same type of experimental construction can be explored in the context of structure learning [Kocaoglu et al. 2017]; see also [Ghassami et al. 2018, Kocaoglu et al. 2019].

[41] The truncated formula is also known as the "manipulation theorem" [Spirtes et al. 1993] or G-computation formula [Robins 1986, pp. 1423]. For further details, we refer readers to [Pearl 2000, § 3.6.4].

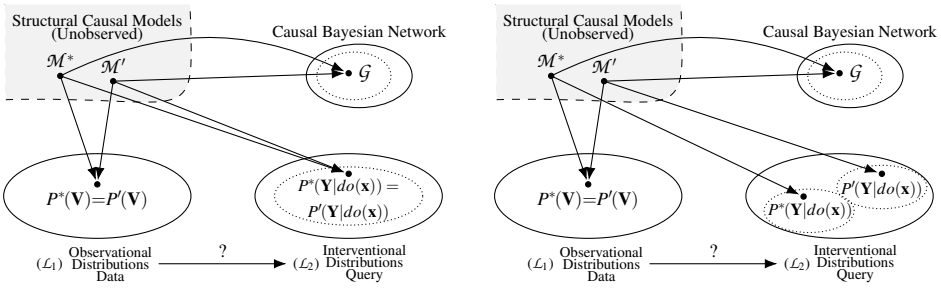


Figure 1.6: Two identifiability scenarios. In some settings, the causal graph is sufficient for solving this task (a), while in others the problem is unsolvable (b).

$P(\mathbf{Y} \mid do(\mathbf{x}))$ is always identifiable and given by the expression

$$P(\mathbf{Y} \mid do(\mathbf{x})) = \sum_{\mathbf{z}} P(\mathbf{Y} \mid \mathbf{x}, \mathbf{z}) P(\mathbf{z}), \tag{1.33}$$

where $\mathbf{Z}$ is the set of all variables not affected by the action $\mathbf{X}$ (non-descendants of $\mathbf{X}$). ■

The importance of this result stems from the fact that the estimand in Eq. (1.33), called adjustment, is generally used in many fields – viz., averaging the conditional probability of $\mathbf{Y}$ given $\mathbf{X}$ by the margins of the covariates. This corollary specifies a condition on its causal validity; in words, if the set of covariates $\mathbf{Z}$ is constituted by all pre-treatment variables and the model is Markovian (all relevant sources of variations are measured), adjusting for these variables will lead to the causal effect.[42]

1.4.3 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 summarized in Fig. 1.6(a), which shows 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 (Thm. 3). This is a quite remarkable result since 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.

[42] A condition known as *conditional ignorability* follows, i.e., $\mathbf{Y}_{\mathbf{x}} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}$ [Imbens and Rubin 2015, Rosenbaum and Rubin 1983], whenever these conditions of the corollary are satisfied; see discussion in [Pearl 2000, § 11.3.2].

Figure 1.7: 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.

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 11 (Example 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 the graphical structure shown in Fig. 1.7(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 Fig. 1.7(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 (Eq. 1.12). Obviously, this would be violated were the world to mirror this graphical structure. To witness, consider the alternative SCM $\mathcal{M}'$, similar to the construction discussed in Example 8, 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 not know, neither want to assume, their particular form, dimensionality, or distribution. This new type of arrow will allow for the probabilistic dependence between them, $(X \not\perp\!\!\!\perp Y)$, while being neutral with respect to their interventional invariance. That is, it would accept constraints such as $P(Y|do(X)) = P(Y)$ and $P(X|do(Y)) = P(X)$. See Fig. 1.7(d) for a graphical example.[43]

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 Fig. 1.7(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 this model. In 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 Eq. (1.29) and, more generally, in cond. (iii) of the definition of Markovian CBNs (Def. 12).

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 12. Consider two SCMs $\mathcal{M}^*$ and $\mathcal{M}'$ such that $\mathbf{V} = \{X, Y\}$, $\mathbf{U} = \{U_{y_0}, U_{y_1}, U_{xy}\}$, the structural mechanisms are:

$$\mathcal{F}^* = \begin{cases} X & \leftarrow U_{xy} \\ Y & \leftarrow (U_{y_1} \wedge X) \vee (U_{y_0} \wedge \neg X) \vee (U_{xy} \oplus X) \end{cases}, \tag{1.34}$$

$$\mathcal{F}' = \begin{cases} X & \leftarrow U_{xy} \\ Y & \leftarrow (U_{y_1} \wedge X) \vee (U_{y_0} \wedge \neg X) \vee ((U_{xy} \oplus X) \wedge U_{y_1}) \end{cases}. \tag{1.35}$$

The exogenous distributions of both models, $P^*(\mathbf{U})$ and $P'(\mathbf{U})$, are the same and given by $P(U_{xy} = 1) = 1/2$, $P(U_{y_0} = 1) = 4/5$, $P(U_{y_1} = 1) = 1/4$, and they both follow the diagram shown in Fig. 1.7(e). It is easy to verify that both models induce the same $P(\mathbf{V})$, while $P^*(Y = 1 \mid do(X = 1)) = 5/8 \neq 1/4 = P'(Y = 1 \mid do(X = 1))$. □

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 correspond-

[43] This constitutes a subtle distinction between $\mathcal{L}_1$ and $\mathcal{L}_2$ views of the underlying SCM. If we consider an SCM such that X and Y are not causally related but share an unobserved common cause, they will appear in $\mathcal{L}_1$ as Fig. 1.7(a) (or with the arrow reversed), since no prediction about the effect of interventions could be made, while in $\mathcal{L}_2$ this unobserved confounder needs to be encoded through a bidirected arrow, such as in Fig. 1.7(d). The existence of minimal i-maps and the obedience of the corresponding blankets is somewhat more subtle, however, which is discussed in more detail in [Bareinboim et al. 2020a, Example 16 in Appendix C].

ing $\mathcal{L}_1$-distribution, $P(X,Y)$. There exists at least one other SCM $\mathcal{M}'$ (Eq. 1.35) 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. This scenario is summarized in Fig. 1.6(b). In words, one cannot commit and make a claim about the target effect since 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 Def. 11, to formally account for the existence of unobserved confounders.

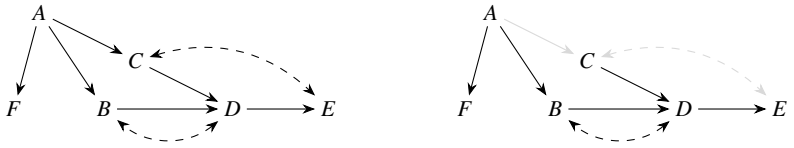
Definition 13 (Causal Diagram (Semi-Markovian Models))**.** Consider an SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$. Then, $\mathcal{G}$ is said to be a *causal diagram* (of $\mathcal{M}$) if constructed as follows:
(1) add a vertex for every endogenous variable in the set $\mathbf{V}$,
(2) add an edge $(V_j \rightarrow V_i)$ 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 $(V_j \dashleftarrow\dashrightarrow V_i)$ 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.

1.4.3.1 Revisiting Locality in Semi-Markovian Models

Graphical models provide a transparent and systematic way of encoding structural constraints about the underlying SCM (Fig. 1.4(d)). In practice, these constraints follows 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 Def. 12). 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. In Fig. 1.5(a), for example, we can see that $Y \perp\!\!\!\perp NDesc_y \mid Pa_y$, where the non-descendants of Y is $\{X\}$ and the parent set is $\{Z\}$. This implies, for example, that if we know the value of Z, the value of X is (probabilistically) irrelevant for computing the likelihood of Y. Further, once we intervene on Z, the variable Y is invariant to interventions on X, i.e., $P(y \mid do(x,z)) = P(y \mid do(z)), \forall x, y, z$.

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 Fig. 1.8(a), and note that, where $Pa_d = \{B,C\}$ and the remaining $NDesc_d = \{A,F\}$, $D \perp\!\!\!\perp$



(a) graph with four C-components. (b) graph under intervention $do(c)$.

Figure 1.8: Causal diagram with bidirected arrows and its mutilated counterpart under $do(c)$.

$NDesc_d \mid Pa_d$ does not hold since D and A are connected through the open path $D \dashleftarrow\dashrightarrow B \leftarrow A$. 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 14 (Confounded Component)**.** Let $\{\mathbf{C}_1, \mathbf{C}_2, \ldots \mathbf{C}_k\}$ be a partition over the set of variables $\mathbf{V}$, where $\mathbf{C}_i$ is said to be a confounded component (for short, C-component) of $\mathcal{G}$ if for every $V_i, V_j \in \mathbf{C}_i$ there exists a path made entirely of bidirected edges between V_i and V_j 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 Fig. 1.8(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 Figs. 1.7(a-c) contains two C-components, $\mathbf{C}_1 = \{X\}$ and $\mathbf{C}_2 = \{Y\}$, while each in Figs. 1.7(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 since 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 Fig. 1.8(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 since 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, i.e., 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 $\mathcal{G}$,[44] 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 Fig. 1.8(a), $Pa_e^+ = \{D,C,A\}$ and $Pa_d^+ = \{B,C,A\}$.

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 15 (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

$$P(\mathbf{v}) = \prod_{V_i\in\mathbf{V}} P(v_i \mid pa_i^+), \tag{1.36}$$

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 $P(\cdot \mid do(\mathbf{x}))$ under an intervention $do(\mathbf{X}=\mathbf{x})$. The corresponding graph is $\mathcal{G}_{\overline{\mathbf{X}}}$, where the incoming arrows towards $\mathbf{X}$ are cut, and the semi-Markovian factorization is

$$P_{\mathbf{x}}(\mathbf{v}) = \prod_{V_i\in\mathbf{V}} P_{\mathbf{x}}(v_i \mid pa_i^{\mathbf{x}+}), \tag{1.37}$$

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

Example 13 (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 Fig. 1.8(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, $(C \perp\!\!\!\perp B \mid A)$, $(E \perp\!\!\!\perp B \mid D,C,A)$ and $(F \perp\!\!\!\perp E,D,C,B \mid A)$. The same applies to interventional distributions. First, let $P_c(A,B,C,D,E,F)$ be semi-Markov relative to $\mathcal{G}_{\overline{C}}$ (Fig. 1.8(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 $(E \perp\!\!\!\perp A \mid D)$. This last constraint holds true since $(C \dashleftarrow\!\!\dashrightarrow E)$ is absent in $\mathcal{G}_{\overline{C}}$. The extended parents in both distributions are $Pa_e^+ = \{A,C,D\}$ and $Pa_e^{C+} = \{D\}$, respectively. □

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

1.4.3.2 CBNs with Latent Variables – Putting all the pieces together

The constructive procedure described in Def. 13 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.[45] 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 *Causal Bayesian Network* (CBN) if it satisfies the definition below. This substitutes for Def. 12 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 Bayesian network and the additional missing-link and do-see constraints are encoded in the Markov-CBNs (Def. 12).

Definition 16 (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 Causal Bayesian Network 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 $V_i \in \mathbf{V} \setminus \mathbf{X}$, $\mathbf{W} \subseteq \mathbf{V} \setminus (Pa_i^{\mathbf{x}+} \cup \mathbf{X} \cup \{V_i\})$:

$$P(v_i \mid do(\mathbf{x}), pa_i^{\mathbf{x}+}, do(\mathbf{w})) = P(v_i \mid do(\mathbf{x}), pa_i^{\mathbf{x}+}), \tag{1.38}$$

(iii) [Missing bidirected-link] For every $V_i \in \mathbf{V} \setminus \mathbf{X}$, 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

$$P(v_i \mid do(\mathbf{x}), pa_i^c, do(pa_i^u)) = P(v_i \mid do(\mathbf{x}), pa_i^c, pa_i^u). \tag{1.39}$$

■

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 13. 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 words, $\mathbf{W}$ has no direct effect on V_i. For instance, note that for $V_i = D$ in Fig. 1.8(a), $P(d \mid do(f,e), b, c, a) = P(d \mid b, c, a)$ as well as $P(d \mid do(b,c), do(a,f,e)) = P(d \mid do(b,c))$. Further, the missing bidirected-link condition relaxes the stringent parents do/see condition in Markovian CBNs (Def. 12(iii)). Note that the do/see condition does not hold due

[45] 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.

	CBN		BN
	Markovian	Semi-Markovian	
$P(\mathbf{V}) \to P(\mathbf{Y} \mid do(\mathbf{x})) \in \mathbf{P}_*$	Always	Sometimes	Never

Figure 1.9: Summary of when we can identify a causal query from the observational $\mathcal{L}_1$ distribution and different types of graphical models compatible with the underlying phenomena.

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 Fig. 1.8(a).[46] 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 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 Def. 13) 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. 2020a, Appendix D, Algorithm 2] a procedure called CONSTRUCTCBN that constitutes such an alternative. It can be seen as the experimental-stochastic counterpart of the SCM-functional Def. 13 (see also Thm. 10). We show in the next section that CBNs can act as a basis for causal inference regardless of their underlying generating model.

1.4.3.3 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 12. This dichotomy is summarized in Fig. 1.9. We introduce next the notion of identifiability [Pearl 2000, pp.77] to more visibly capture each of these instances.

Definition 17 (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

[46] 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 $P(e|do(d))$ does not depend on U_d, while $P(e|d)$ does due to the path $U_d \to D \leftarrow C \dashleftarrow\dashrightarrow E$.

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 $P^{(1)}(\mathbf{Y} \mid do(\mathbf{x}),\mathbf{z})=P^{(2)}(\mathbf{Y} \mid do(\mathbf{x}),\mathbf{z})$. ■

This formalizes the very natural type of cross-layer inference we have discussed in Fig. 1.4, 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.[47,48]

Theorem 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:*

$$\textbf{\textit{Rule 1}} \quad P(\mathbf{y} \mid do(\mathbf{x}),\mathbf{z},\mathbf{w}) = P(\mathbf{y} \mid do(\mathbf{x}),\mathbf{w}) \qquad \text{if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{X},\mathbf{W}) \text{ in } \mathcal{G}_{\overline{\mathbf{X}}}. \quad (1.40)$$

$$\textbf{\textit{Rule 2}} \quad P(\mathbf{y} \mid do(\mathbf{x}),do(\mathbf{z}),\mathbf{w}) = P(\mathbf{y} \mid do(\mathbf{x}),\mathbf{z},\mathbf{w}) \qquad \text{if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{X},\mathbf{W}) \text{ in } \mathcal{G}_{\overline{\mathbf{X}}\underline{\mathbf{Z}}}. \quad (1.41)$$

$$\textbf{\textit{Rule 3}} \quad P(\mathbf{y} \mid do(\mathbf{x}),do(\mathbf{z}),\mathbf{w}) = P(\mathbf{y} \mid do(\mathbf{x}),\mathbf{w}) \qquad \text{if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{X},\mathbf{W}) \text{ in } \mathcal{G}_{\overline{\mathbf{X}\mathbf{Z}(\mathbf{W})}}, \quad (1.42)$$

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[49] 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 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, Ch. 3].[50]

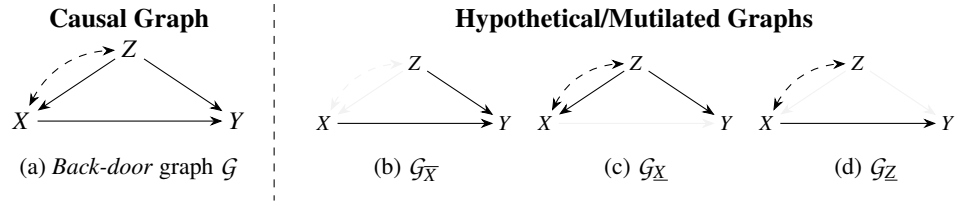
We have previously shown that in some very simple settings causal inference is unattainable with only $\mathcal{L}_1$-data, and that causal knowledge conveniently encoded in the form of a

[47] 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, possibly far away in the causal diagram.

[48] 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 [Lauritzen 1996, Pearl 1988]. The graphical model could be seen as a basis, i.e., 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.

[49] This criterion is trivially extendable to graphs with bidirected edges by interpreting them as common causes of the variables, as if connected by two corresponding directed edges.

[50] Interestingly, the do-calculus theorem (Thm. 5) as stated here was derived entirely within the domain of CBNs and layer 2 constraints, which contrasts with the traditional proposition ([Pearl 1995, Thm. 3]) based on layer 3 facts. While this is an obviously valid route, we believe the statement in its current form may allow some researchers to use this result even when taking an alternative, non-structural route; see also footnote 40.



(a) *Back-door* graph $\mathcal{G}$ (b) $\mathcal{G}_{\overline{X}}$ (c) $\mathcal{G}_{\underline{X}}$ (d) $\mathcal{G}_{\underline{Z}}$

Figure 1.10: (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 the graphs (b), (c) and (d).

causal diagram is required. Next, we show how the knowledge from the diagram together with the inference rules of do-calculus allow for the identification of the query $P(y \mid do(x))$ in the context of the model represented in Fig. 1.10(a). First, we start with the target query and then apply do-calculus:

$$P(y \mid do(x)) = \sum_z P(y \mid do(x), z) P(z \mid do(x)) \qquad \text{Summing over } Z \tag{1.43}$$

$$= \sum_z P(y \mid do(x), z) P(z) \qquad \text{Rule 3: } (Z \perp\!\!\!\perp X)_{\mathcal{G}_{\overline{X}}} \tag{1.44}$$

$$= \sum_z P(y \mid x, z) P(z) \qquad \text{Rule 2: } (Y \perp\!\!\!\perp X \mid Z)_{\mathcal{G}_{\underline{X}}}. \tag{1.45}$$

Each step above is accompanied by the corresponding probability axiom or rule, supported by the licensing graphs $\mathcal{G}_{\overline{X}}$ and $\mathcal{G}_{\underline{X}}$ (Figs. 1.10(b) and (c), respectively). As desired, the right-hand side of Eq. (1.45) 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 Fig. 1.10(a), the causal effect of X on Y will always be equal to Eq. (1.45). This can be seen as an instance of a generalized version of the back-door criterion provided in the Markovian case (Corollary 2). Next, we state the criterion that validates such adjustment more generally.

Corollary 3 (Back-door Criterion)**.** *Let $\mathcal{G}$ be a causal diagram and $\mathbf{X}$ and $\mathbf{Y}$ be the sets of treatment and outcomes variables, respectively. A set of variables $\mathbf{Z}$ is said to satisfy the back-door criterion relative to the pair $(\mathbf{X}, \mathbf{Y})$ in $\mathcal{G}$ if:*

(i) No node in $\mathbf{Z}$ is a descendant of $\mathbf{X}$, and

(ii) $\mathbf{Z}$ blocks every path between $\mathbf{X}$ and $\mathbf{Y}$ that contains an arrow into $\mathbf{X}$.

If such $\mathbf{Z}$ exists, the causal effect of $\mathbf{X}$ on $\mathbf{Y}$ is identifiable and given by the expression:

$$P(\mathbf{Y} \mid do(\mathbf{x})) = \sum_{\mathbf{z}} P(\mathbf{Y} \mid \mathbf{x}, \mathbf{z}) P(\mathbf{z}). \tag{1.46}$$

■

Many observations follow from this result. First, it is not hard to see that the derivation provided above for the causal diagram in Fig. 1.10(a) (i.e., Eqs. (1.43)-(1.45)) constitutes an

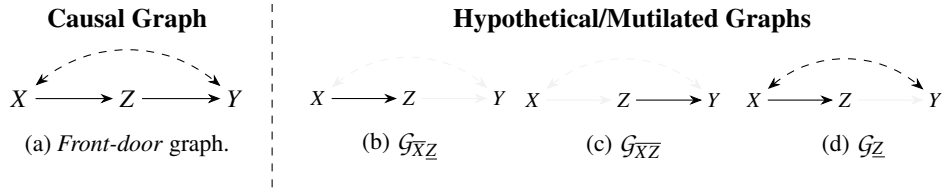


Figure 1.11: Front-door graph (a) and derived graphs used to identify $P(y \mid do(x))$.

outline of the proof for Corollary 3. Note that the application of the third rule of do-calculus that led to Eq. (1.44) is satisfied whenever the set $\mathbf{Z}$ does not include descendants of $\mathbf{X}$, as required by cond. (i) of the criterion. Further, the application of the second rule of do-calculus that culminated in the adjustment given by Eq. (1.45) represents exactly cond. (ii) of the criterion, namely, that the back-door paths (with arrows into $\mathbf{X}$) are blocked. Interestingly, the most practical way of checking the back-door admissibility of a candidate set $\mathbf{Z}$ is by removing the outgoing arrows from $\mathbf{X}$ and confirming whether $\mathbf{Z}$ separates $\mathbf{X}$ and $\mathbf{Y}$ in this graph (Fig. 1.10(c)). This is precisely the graph associated with the second rule of do-calculus, $(Y \perp\!\!\!\perp X \mid Z)_{\mathcal{G}_{\underline{X}}}$, and gives the name of the criterion since it isolates specifically these paths.

Second, the importance of the back-door criterion stems from the fact that adjustment is arguably the most common technique used to identify causal effects found throughout the sciences. While the expression given in Eq. (1.46) has been used since much earlier than the discovery of the criterion itself [Pearl 1993], this condition 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 through adjustment based on a model of the world.[51]

Third, for the effect of Z on X, $P(X \mid do(z))$, in the same graph (Fig. 1.10(a)), there exists no set $\mathbf{Z}$ that satisfies the conditions of Corollary 3. Note that in the graph where the arrows outgoing from Z are cut, Fig. 1.10(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 the proof provided in Example 12. Broadly speaking, 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. This dichotomy is summarized in Figs. 1.6(a) and (b).

To further ground our understanding of do-calculus, consider the causal diagram in Fig. 1.11(a), and assume that our goal is to identify the interventional distribution $P(Y \mid do(x))$ from the observational one, $P(Z,X,Y)$. We start by noting that there exists no back-door admissible set with respect to (X,Y) since X and Y are connected through a bidirected arrow in $\mathcal{G}_{\underline{X}}$. One may be tempted to surmise that this effect is not identifiable as discussed in the

[51] The back-door criterion provides a formal and transparent condition to judge the validity of a condition called conditional ignorability; see further details in [Pearl 2000, § 11.3.2].

previous example. To witness that this is not the case (and that the back-door is not necessary for identification), we start by writing the target distribution and expanding it through do-calculus:

$$P(y \mid do(x)) = \sum_z P(y \mid do(x), z) P(z \mid do(x)) \qquad \text{Summing over } Z \tag{1.47}$$

$$= \sum_z P(y \mid do(x, z)) P(z \mid do(x)) \qquad \text{Rule 2: } (Y \perp\!\!\!\perp Z \mid X)_{\mathcal{G}_{\overline{X}\underline{Z}}} \tag{1.48}$$

$$= \sum_z P(y \mid do(z)) P(z \mid do(x)) \qquad \text{Rule 3: } (Y \perp\!\!\!\perp X \mid Z)_{\mathcal{G}_{\overline{ZX}}}. \tag{1.49}$$

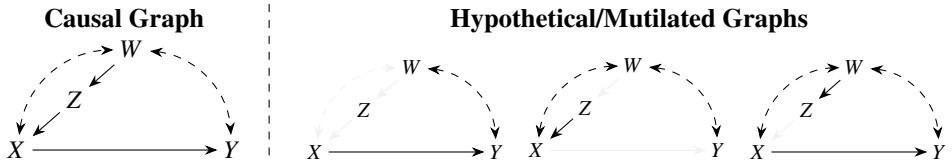
The rules used in each step and licensed by the corresponding graphs are shown in Figs. 1.11(b)-(c). At this point, Corollary 3 can be used to solve for both factors in Eq. (1.49). In the first factor, $P(y \mid do(z))$, $\{X\}$ itself is back-door admissible with respect to (Z, Y), since $(Y \perp\!\!\!\perp Z \mid X)_{\mathcal{G}_{\underline{Z}}}$ (see Fig. 1.11(d)). For the second factor, $P(z \mid do(x))$, the empty set is back-door admissible with respect to (X, Z), since $(Z \perp\!\!\!\perp X)_{\mathcal{G}_{\underline{X}}}$. Putting these two results together and replacing it back into Eq. (1.49) leads to:

$$P(y \mid do(x)) = \sum_z P(z \mid x) \sum_{x'} P(y \mid z, x') P(x'). \tag{1.50}$$

The right-hand side of Eq. (1.50) 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 its idiosyncrasies. This particular expression is known as the front-door adjustment since it leverages the outgoing arrows from X, in contrast to the back-door which considered the incoming arrows towards X. The conditions under which the front-door adjustment (Eq. (1.50)) can be applied are more general than the graph in Fig. 1.11(a), as shown in [Pearl 2000, § 3.3.2], and generalized in [Hünermund and Bareinboim 2019, Thm. 3.5]. The front-door setting can be seen as the composition of two consecutive applications of the back-door adjustment, one with the empty set and another with the set $\{X\}$.[52]

Finally, there are more involved scenarios that generated surprise in the literature since they go beyond some of the intuitions discussed in the examples above; for instance, see diagram in Fig. 1.12(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 since the empty set is not admissible, 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. It is also clear that no front-door strategy can be used due to the direct arrow from X to Y, $X \to Y$. After all, one may be tempted to believe that the effect

[52] The conditions under which estimands can be re-expressed in terms of a composition of back-door adjustments has been pursued recently in [Jung et al. 2020], where the front-door constitutes just one special case. This composition allows one to leverage statistically efficient and robust estimation methods such as the propensity score and inverse probability weighting (IPW), going beyond the classic back-door setting.



(a) *Napkin* graph. (b) $\mathcal{G}_{\overline{XZ}}$ (c) $\mathcal{G}_{\overline{Z}\underline{X}}$ (d) $\mathcal{G}_{\underline{Z}}$

Figure 1.12: Napkin graph (a) and derived graphs used to identify $P(y \mid do(x))$.

of X on Y is not identifiable in this case.[53] Contrary to this intuition, consider the following derivation in do-calculus:

$$P(y \mid do(x)) = P(y \mid do(x), do(z)) \qquad \text{Rule 3: } (Y \perp\!\!\!\perp Z \mid X)_{\mathcal{G}_{\overline{XZ}}} \tag{1.51}$$

$$= P(y \mid do(z), x) \qquad \text{Rule 2: } (Y \perp\!\!\!\perp X)_{\mathcal{G}_{\overline{Z}\underline{X}}} \tag{1.52}$$

$$= \frac{P(y, x \mid do(z))}{P(x \mid do(z))} \qquad \text{Def. of cond. probability.} \tag{1.53}$$

The rules used in each step and the licensing graphs are shown in Figs. 1.12(b)-1.12(c). At this point, Corollary 3 can be applied to solve for both factors in Eq. (1.53). To witness, note that in the numerator, $P(y, x \mid do(z))$, $\{W\}$ is back-door admissible with respect to $(Z, \{Y, X\})$, since $(Y, X \perp\!\!\!\perp Z \mid W)_{\mathcal{G}_{\underline{Z}}}$, as shown in Fig. 1.12(d). The denominator follows by marginalizing Y out. Putting these two results together and replacing it back into Eq. (1.53) lead to:

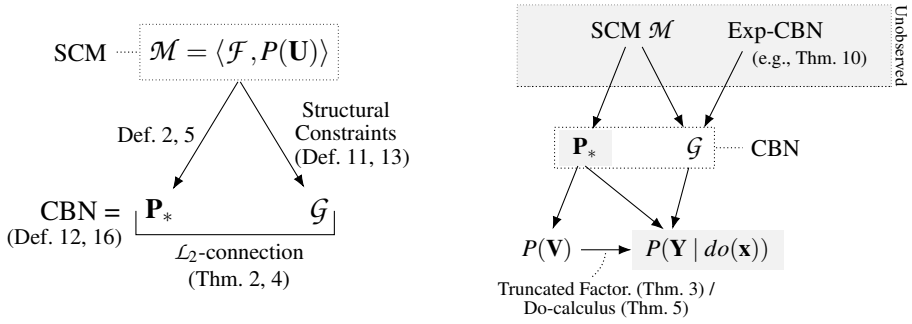
$$P(y \mid do(x)) = \frac{\sum_w P(y, x \mid z, w) P(w)}{\sum_w P(x \mid z, w) P(w)}. \tag{1.54}$$

The r.h.s. of Eq. (1.54) 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 Eq. (1.54) is a ratio following from the application of the back-door criterion twice.

1.4.4 Summary – Background Context and Recent Developments

Broadly speaking, once one acknowledges the existence of the causal mechanisms (SCM $\mathcal{M}$) underlying the phenomenon under investigation, all conceivable quantities from the PCH defined over the endogenous variables obtain precise numerical values. The collection of such quantities is called $\mathbf{P}_*$, which includes the observational (Def. 2) and interventional (Def. 5)

[53] As alluded to by Pearl himself: "There was no way to block the back-door paths and no front-door condition. I tried all my favorite shortcuts and my otherwise trustworthy intuitive arguments, both pro and con, and I couldn't see how to do it. I could not find a way out of the maze. But as soon as Bareinboim whispered to me, "Try the do-calculus," the answer came shining through like a baby's smile. Every step was clear and meaningful. This is now the simplest model known to us in which the causal effect needs to be estimated by a method that goes beyond the front- and back-door adjustments." [Pearl and Mackenzie 2018, pp. 240]. The structure is known as the new *Napkin graph*.



(a) Graphical Semantics (b) Causal Inference

Figure 1.13: (a) Summary of the results connecting the SCM $\mathcal{M}$ and its corresponding coarsening $\langle \mathbf{P}_*, \mathcal{G} \rangle$, where $\mathbf{P}_*$ is the collection of distributions and $\mathcal{G}$ is the causal diagram. (b) Summary of the causal inference task that entail another level of coarsening, since $\mathbf{P}_*$ is not fully observable. The goal here is to bridge distributions within $\mathbf{P}_*$ by leveraging $\mathcal{G}$. The marked areas (shaded in gray) represent the unobserved components of the analysis.

distributions; see the left side of Fig. 1.13(a). Since any query is computable from the fully specified SCM, why is there a need to pursue the graphical approach developed in this section?

As discussed earlier in the chapter, in many practical situations it is implausible for scientists to have a full specification of the SCM (i.e., $\langle \mathcal{F}, P(\mathbf{U}) \rangle$). Instead, only data generated by the underlying model is at disposal. Unfortunately, the lack of detailed knowledge of the causal mechanism cannot be surmounted with data alone; except, perhaps, in specific settings where data and query match one another precisely. As stated in Corollary 1, if the query belongs to a layer above the one where data come from, causal inference is provably impossible. Therefore, one is compelled to investigate more relaxed inferential tasks that may rely on some sufficient, but attainable knowledge of the underlying causal mechanisms.

For this purpose, we introduced a procedure for constructing a graphical model $\mathcal{G}$ encoding the structural constraints of $\mathcal{M}$, bestowed with causal interpretation (Defs. 11 and 13, for Markovian and semi-Markovian models, respectively). In fact, $\mathcal{G}$ can be seen as a non-parametric coarsening of the SCM $\mathcal{M}$, where the signatures (i.e., arguments) of the functions are preserved while their specific forms are dismissed (see the right side of Fig. 1.13(a)).

To give the graphical model a precise interpretation, we introduced an object called Causal Bayesian Network (CBN) to marry the structural constraints encoded in $\mathcal{G}$ and the collection of distributions in $\mathbf{P}_*$ (Defs. 12 and 16). Furthermore, and irrespective of the SCM, the CBN construct provides means of deciding whether a pair $(\mathcal{G}, \mathbf{P}_*)$ is compatible in some causal fashion. Whenever this is the case, a CBN can serve as a surrogate for the empirical content

of the underlying SCM with respect to knowledge in layers $\mathcal{L}_1$ and $\mathcal{L}_2$ (Thms. 2 and 4, for Markovian and semi-Markovian models, respectively).

Obviously, operating at this level of coarsening does not come without a price, as challenges will appear when attempting to bridge the layers of the hierarchy. For instance, inferring an interventional distribution $P(\mathbf{Y} \mid do(\mathbf{x}))$ from the observational distribution $P(\mathbf{V})$ and the causal diagram $\mathcal{G}$, will not always be possible. Depending on the existence of exogenous variables and how they influence the observed system, cross-layer inferences will entail very different answers. In Markovian CBNs, the truncated factorization product (Thm. 3) maps any interventional distribution ($\mathcal{L}_2$) to the observational ($\mathcal{L}_1$) one, effectively bridging Layers 1 and 2 (Fig.1.6(a)). For semi-Markovian CBNs, we proved that the do-calculus holds (Thm. 5), again, irrespective of the SCM. We showed that in many, but not all, cases the target effect is identifiable (Fig.1.6(b)). The corresponding impossibility results evidence the semantic gap between the SCM $\mathcal{M}$ (that generates the entire $\mathbf{P}_*$) and its diagram $\mathcal{G}$ (that identifies only a portion of it; see Fig. 1.13(b) for a graphical summary.

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, Kuroki and Miyakawa 1999, Pearl and Robins 1995, Spirtes et al. 1993]. 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 [Bareinboim and Pearl 2012, Huang and Valtorta 2006, Lee et al. 2019, Shpitser and Pearl 2006, Tian and Pearl 2002a]. In 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 $\mathbf{Z}$, $P(\mathbf{y}|do(\mathbf{x}), \mathbf{z})$, 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.

Recent Developments

The clarity gained from theoretical understanding of the do-calculus machinery has inspired numerous developments in the field. We summarize next a few important threads in the literature that seem especially significant for the data sciences, given their generality.

Stochastic, conditional, and non-atomic interventions. There are complex interventions that cannot be described in terms of atomic ones, as delineated by the do-operator, including conditional and stochastic [Pearl 2000, Ch. 4]. It may be challenging in many settings to asses the effect of new soft-interventions from non-experimental data, including when designing adaptive strategies to treat patients in medicine, developing new incentives structures in an economy, or inferring a new dynamic policy in reinforcement learning. In response to that challenge, a novel treatment building on the do-calculus has recently been introduced [Correa

and Bareinboim 2019], and developed in great generality under the rubric of σ-calculus (shortcut for *soft-do-calculus*) [Correa and Bareinboim 2020].

Data-driven identification. On another line of investigation, one can use the structural constraints (Fig. 1.4(d)) learned from $\mathcal{L}_1$-data to induce an equivalence class of causal diagrams (i.e., all the diagrams that respect the constraints imposed by the observed data) [Peters et al. 2017, Spirtes et al. 2001, Zhang 2008]. The problem of identification in these structures is considerably more challenging due to the inherent uncertainty about the causal structure itself. An initial solution has been proposed in [Jaber et al. 2018, Perkovic et al. 2017], and further developed with greater generality in [Jaber et al. 2019a,b,c].

Estimation beyond the back-door adjustment. Most of the results found in the literature are about identification but, ultimately, one needs to estimate the effect from a finite number of samples. One line of investigation led to efficient and robust estimation techniques [Bang and Robins 2005, Horvitz and Thompson 1952, Imbens and Rubin 2015, Rosenbaum and Rubin 1983, Van Der Laan and Rubin 2006], which was specific to adjustment-like estimands. A recent treatment leveraging insights following form the do-calculus generalized some of these technique based on weighting operators for the non-backdoor cases [Jung et al. 2020].

Data-Fusion – Integrating Heterogeneous Data Collections. The do-calculus was developed for solving the problem of confounding bias, which had been the primary concern in the modern literature of causation at the time [Pearl 1995], and going back at least to [Rubin 1974]. In practice, however, there exist other issues that plague most data collection, including selection bias, external validity, transportability, and, broadly speaking, due to issues of generalizability. Building on the insights learned by handling the problem of confounding, a general framework has been introduced and developed to solve the *data-fusion* problem, i.e., cohesively combining multiple datasets coming from various, heterogeneous sources, and plagued with different types of biases [Bareinboim and Pearl 2016]. For some recent developments on the fusion framework, refer to [Correa et al. 2019a,b] and [Lee et al. 2020], and for some applications in the context of econometrics, see [Hünermund and Bareinboim 2019].

1.5 Reflections and Conclusions

We investigated a mathematical structure called the Pearl Causal Hierarchy (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, a structural causal model (Def. 1) – the hierarchy relative to the modeled phenomena emerges (Def. 9).

The main contribution of this chapter is a detailed analysis of the PCH through three lenses: one semantical (Sec. 1.2), another logical-probabilistic (Sec. 1.3), and another inferential-graphical (Sec. 1.4). 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 believe these distinct interpretive angles provide a powerful and appealing tool for studying causation across different research communities, with far reaching implications for scientific practice in a wide range of data-driven fields. This work also promises to shape the development of the next generation of AI systems.

Semantically speaking (Sec. 1.2), we start from the building block of any causal analysis, a fully specified SCM $\mathcal{M}$, which represents the dynamics of a specific system or phenomenon under investigation. Following Pearl's own presentation, we then develop basic intuitions for how the PCH naturally emerges from an SCM, conveying how an agent could (1) perceive a certain part of the world unfolding in time and predict its next stage, (2) deliberately intervene in this world and evaluate the effects of such an action, and (3) imagine alternative ways the world could be and speculate what would have happened had something different occurred. Provided the SCM $\mathcal{M}$ is fully known, all these and other conceivable quantities of the three layers are well-defined and immediately computable from the model (through Defs. 2, 5, 7).

From this starting point, we move on to the original contributions of the chapter. First, in Sec. 1.3, we define (symbolic) logical languages that were implicit in our discussions in the previous discussion. These languages encode the types of statements and questions associated with the three layers of the hierarchy. We then formally state Thm. 1, the Causal Hierarchy Theorem (CHT), which tells us that the ability to answer questions at one layer, by itself, *almost never* guarantees the ability to answer all the relevant questions at higher layers. Depending on the specific dynamics of the underlying SCM, it is conceivable that the effects of actions (Layer 2) can be fully determined merely by passively observing the world unfolding in time (Layer 1); or, all the facts about which one could imagine or speculate (Layer 3) are fully determined by the experimental evidence (Layer 2). The CHT, on the other hand, says that this is generically impossible (i.e., it is measure zero): each layer makes claims about the world not entailed by lower layers. The theorem can be seen as formal grounding for the mantra "no causes in, no causes out" [Cartwright 1989], commonly taught in introductory causal inference courses.

As a practical matter, one might wonder under what conditions a quantity at one layer of the PCH could be learned from data collected at another. For instance, how could the effect of a new intervention of X on an outcome Y, $P(\mathbf{y} \mid do(\mathbf{x}))$, a layer-2 quantity, be predicted from data collected observationally, $P(\mathbf{x}, \mathbf{y})$, i.e., from Layer 1?

In Sec. 1.4, we explore a solution to this inferential task through the lens of graphical models. The departing point of most of the causal inference literature is the acknowledgment

that the SCM $\mathcal{M}$ as well as the corresponding PCH are there even though $\mathcal{M}$ may be hard to obtain in practice.[54] Even when $\mathcal{M}$ is not fully known, one may still have some understanding of its structure at a coarse level. Accordingly, we study the structural constraints – different types of invariances – implied by the SCM at each layer of the PCH. For instance, a projection of the SCM downwards to the first layer, the world of associations, displays conditional independences. There exists a rich literature on probabilistic reasoning and how to parsimoniously organize knowledge in this fashion, including the use of Bayesian networks [Koller and Friedman 2009, Pearl 1988]. We observe that layer-1 constraints and models are insufficient to answer questions of interventional nature (see Example 10).

Going up the hierarchy, we study layer 2 type constraints, which leads to a family of graphical models called Causal Bayesian Networks (CBNs). There are two types of CBNs depending on whether there are unobserved confounders (i.e., latent variables that affect more than one observable) in the system. We study Markovian CBNs [Bareinboim et al. 2012, Pearl 2000] and show that it is always possible to answer interventional questions from observational data via the truncated factorization product (Thm. 3). Given that unobserved confounders exist in most practical settings, we introduce a general version of CBNs for semi-Markovian systems (Def. 16). We show that in the presence of unobserved confounders the empirical gap between the CBN and the true SCM widens, and hence the knowledge encoded in the CBN is not always sufficient to support cross-layer inferences (Example 12). We subsequently discuss an inferential system that allows one to perform inferences across layers of the PCH, known as *do-calculus* [Pearl 1995]. We prove that do-calculus can be used to facilitate cross-layer inferences from a CBN perspective (Thm. 5), agnostic to the underlying generating model. In other words, one could take an axiomatic view and satisfy the constraints that CBNs require through means other than a structural model (for example, from a set of experimental distributions [Kocaoglu et al. 2017]), which will still lead to causally-valid inferences. By the end of the section, we study instances of identifiable and non-identifiable effects, and summarize new challenges and relaxations of the identification problem leading to more robust and general cross-layer inferential settings.

The work presented so far on the foundations of causal inference has been somewhat technical, and we believe it is worth closing the chapter with some reflections on how these ideas and results relate to some practical concerns arising in modern AI and machine learning. As a discipline largely concerned with the replacement of explicit "hand-coded" rules and algorithms with rules and behaviors learned directly from "patterns in data" [Bishop 2006], ideally with minimal human supervision, machine learning faces formidable inferential challenges that can be understood through the lenses of the PCH. In fact, a central question for machine learning concerns what kinds of inferences one can possibly hope to make from

[54] Given that this is a very dense and technical section, a detailed summary of the results is provided in Sec. 1.4.4.

what kinds of data (which may include, e.g., data produced via interactions in the world). The PCH addresses this question in a direct and transparent way.

As a suggestive final example to illustrate the subtleties as well as the significance of the PCH in practically important tasks, consider a problem that has been increasingly prominent in discussions around the role of AI in contemporary society, namely whether we ought to prefer an AI system or a human medical doctor in a decision-making setting.[55]

Example 14 (MD versus AI physicians)**.** Imagine a case of a patient who needs to decide between two styles of treatment: an AI physician, who is a black-box and cannot explain itself, or a human physician, who is a medical doctor (MD) and can, for instance, offer clear explanations for their decisions. Let X represent the treatment received, $X = 0$ meaning the MD, $X = 1$ meaning treatment by the AI system – and let $Y = 0$ mean that the patient has died, while $Y = 1$ means that the patient is alive after one year of treatment. How should one choose between these two completely different styles of treatment?

One way of proceeding might be to collect data on previous patients in the population who received these two treatments, computing statistics, for example, about the recovery rate for each treatment. This could be implemented with standard methods (including some of the most modern neural methods). The corresponding statistics can be computed after careful and extensive data collection and analysis, and in this case they come back as follows:

$$\begin{aligned} &P(Y = 1 \mid X = 1) = 0.9, \text{ and} \\ &P(Y = 1 \mid X = 0) = 0.8. \end{aligned} \tag{1.55}$$

Naturally, one may be tempted to conclude based on these results that the AI physician is to be preferred to the MD. If we are concerned solely with maximizing the chances of survival, not availability of explanation (nor, e.g., the human-human contact that might be lacking in an artificial system), then why would we opt for a physician with a lower success rate?

Digging a bit deeper in the Pearl Hierarchy, suppose that, unbeknownst to both the MD and the AI physician (not to mention the patient), the following SCM $\mathcal{M}^*$ encodes the true causal mechanisms of how patients decide their treatment and respond to it – formally, $\mathcal{M}^* = \langle \mathbf{V} = \{X,Y\}, U, \mathcal{F} = \{f_x, f_y\}, P(U)\rangle$, where:

$$f_X(U) = \begin{cases} 0, & U = 0,1,2,3 \\ 1, & U = 4,5,6,7 \end{cases}, \qquad f_Y(X,U) = \begin{cases} 0, & U = 0,4 \\ X, & U = 1,5 \\ 1-X, & U = 2,6 \\ 1, & U = 3,7 \end{cases}, \tag{1.56}$$

[55] To cite just one recent discussion appearing in the public sphere, see the spirited discussion following this provocative tweet by Geoffrey Hinton (2/20/2020, 3:37pm): "Suppose you have cancer and you have to choose between a black box AI surgeon that cannot explain how it works but has a 90% cure rate and a human surgeon with an 80% cure rate. Do you want the AI surgeon to be illegal?"

and the distribution over the exogenous variable is $P(U = u) = (1/40,\ 1/40,\ 1/80,\ 3/16,\ 1/160,\ 1/160,\ 11/160,\ 107/160)$, where $u = 0, 1, ..., 7$. The endogenous variables X and Y are both binary and the exogenous variable U is 8-dimensional, so as to represent the different types of preferences patients could have over treatments, and how they respond to them.

As expected, the underlying SCM $\mathcal{M}^*$ produces exactly the same conditional probability statements as the statistics supported by the data and shown in Eq. 1.55. Suppose that, impressed by these numbers, a new regulation is put in place, whereby AI physicians are deployed across the country, implying that patients are uniformly submitted to treatment $do(X = 1)$. What is the effect of this new policy in the general population? Again, following the discussion in Section 1.2 (and Def. 5), $\mathcal{M}^*$ gives meaning to this statement as:

$$P(Y = 1 \mid do(X = 1)) \approx 0.89. \tag{1.57}$$

Likewise, we can compute what would happen in the hypothetical scenario in which no AI system replaces any human MDs, which would lead to the following average treatment effect:

$$P(Y = 1 \mid do(X = 0)) \approx 0.94. \tag{1.58}$$

As Eqs. 1.57 and 1.58 reveal, the result of this policy is quite disastrous – given the true causal mechanism $\mathcal{M}^*$, it is clear that significantly more lives would be saved by relying on human doctors; on average AI is hurting more people than helping.

Tension between Layers 1 and 2. The trend reversal revealed in Eq. 1.55 (Layer 1) and Eqs. 1.57-1.58 (Layer 2) is no surprise from the standpoint of the PCH (for example, see [Pearl 2000, Ch. 6]). The purpose of the example is to clarify how the whole analysis is contingent on the type of data collected with respect to PCH's layers. For instance, if the samples are drawn from a layer 1 distribution, the results are likewise constrained to layer 1 claims, viz. predictions. To be sure, the best learning algorithm (e.g., a deep neural network) will be able to approximate with great precision the probability of recovery for those individuals deciding by themselves whether they opt for the AI or the human-type of treatment. Still, no claim can be made about the effect of the intervention and what would be the medical recommendation; only a prediction can be offered.

Similarly, if the data is collected from experiments with respect to a layer 2 distribution, the results are again constrained to the corresponding layer 2 claims (viz. interventions). After all, we learned from the Causal Hierarchy Theorem (Thm. 1) that claims from a certain layer need to be supported by knowledge at the same layer or above, regardless of the amount of data or computational resources available.

Climbing further up the PCH, we could imagine another group of more enlightened policy-makers who have internalized the lesson that correlation is not causation. Since they do not have access to the real collection of causal mechanisms underlying the system (as shown in Eqs. 1.56), they decide to conduct a careful randomized clinical trial (RCT) leveraging modern

reinforcement learning techniques. Lo and behold, the statistics come as describe in Eqs. 1.57 and 1.58. Naturally, they may be tempted to conclude based on these results that the human physician is to be preferred to the AI, flipping the decision of the first group of policy-makers. Again, if we are concerned with maximizing the chances of survival, then why would we opt for an AI with a lower success rate?

Further, digging even deeper in the Pearl Hierarchy, it makes sense to ask about the probability of survival under a certain treatment, given that the patient is inclined to opt for one or the other treatment. For instance, the patient may be tempted to go with the AI system ($X = 1$), given their age group, gender, socioeconomic status, personality, and other factors, and may wonder whether they would survive ($Y = 1$) had they been treated by the human ($X = 0$). This query has meaning in counterfactual language (Layer 3), and can be formally written as $P(Y_{X=1} = 1 \mid X = 0)$. In particular, they can be evaluated by $\mathcal{M}^*$ (Eq. (1.56)), as discussed in Section 1.2 (and following Def. 7),

$$\begin{aligned} &P(Y_{X=1} = 1 \mid X = 0) = 0.85 \qquad &&P(Y_{X=1} = 1 \mid X = 1) = 0.90 \\ &P(Y_{X=0} = 1 \mid X = 0) = 0.80 \qquad &&P(Y_{X=0} = 1 \mid X = 1) \approx 0.98. \end{aligned} \tag{1.59}$$

In words, among those patients inclined to opt for the MD ($X = 0$ after the conditioning bar), treatment by the AI physician ($X = 1$) would actually be better (since $P(Y_{X=1} = 1 \mid X = 0) > P(Y_{X=0} = 1 \mid X = 0)$), while the opposite pattern can be seen for those inclined to prefer the AI physician (since $P(Y_{X=1} = 1 \mid X = 1) < P(Y_{X=0} = 1 \mid X = 1)$). Knowing these additional, quintessentially layer 3 facts, the best policy would actually be to give patients whatever treatment they would be inclined *not* to choose. (We leave as a challenge to the readers to interpret how this population is structured, in terms of the generating $\mathcal{M}^*$ (Eq. (1.56), and how this type of pattern could arise in the real world. One hint would be to dissect the U-space and understand the endogenous and exogenous components of the SCM.)

Tension between Layers 2 and 3. The reversal shown of the statistics obtained in Eqs. (1.57)-(1.58) (Layer 2) and Eqs. (1.59) (Layer 3) is again not surprising from the standpoint of the PCH (for example, see [Bareinboim et al. 2015, Forney et al. 2017]). Still, it does make clear that the output of the analysis is contingent on the type of data collected, even when it comes from direct experimentation (Layer 2). To be sure, the best learning algorithm will be able to approximate with great precision the treatment effect when we randomize across units in the population, but this will miss an opportunity to leverage information revealed by the human's decision, imperfect as this may be. This turns out to be a great opportunity for collaboration – human decision-makers may perform poorly (Eq. 1.55); purely automated decision-making procedures may also perform poorly (Eqs. 1.57-1.58). From the perspective of Layer 3, as evident from Eqs. (1.59), optimal decision-making emerges from a combination of humans intuitions, even when wrong, with machine capabilities. The lesson here is, of course, more

general, as the example does not depend on what exactly X and Y are. For a detailed discussion on the implications of causality to decision-making, see [Bareinboim et al. 2020b]. □

To summarize, whenever a domain of interest can be construed as a collection of causal mechanisms (i.e., as an SCM), the PCH emerges at once. To the extent that the SCM is already known, all the relevant quantities at all layers can in principle be computed. For almost any SCM, however, higher layers in the hierarchy remain independent of those lower in the hierarchy; it is a generic property of SCMs that information at lower layers underdetermines information at higher layers. Because in practice the SCM is rarely known, we thus face the formidable challenge of cross-layer inference, a challenge that has been illustrated in Example 14 and the other examples in this chapter. Any sensible approach to this problem must involve a subtle interplay between (1) prior knowledge we may have in the form of a partial specification of the underlying SCM, (2) the type of data that we can collect, and (3) the inferential target of the analysis (Fig. 1.4). Pearl's original presentation of the PCH and the results presented here that build on this conceptual and formal foundation allow us to understand from first principles when a question about some aspect of the world (whether associational, interventional, or counterfactual) can – at least in principle – be answered from a combination of prior knowledge about the domain and the available data.

We note that the Pearl Causal Hierarchy has a certain *inevitability* to it. Many of the most fundamental problems that we need to solve – "What will happen if I do this?," "Why did that happen?," etc. – compel us upward in the hierarchy. Even when we lack data or experience at the relevant layer, there are still powerful methods we can use to answer such questions. If our goal is to build the next generation of AI systems that are human-compatible, capable of explaining themselves, aligned with the social good, safe and robust to changing conditions, we submit that connecting with the fundamental dimensions of the human experience delineated by the PCH is a critical step towards this goal.

Acknowledgements

This work has benefited immensely from conversations with David Blei, Carlos Cinelli, Philip Dawid, Sanghack Lee, Judea Pearl, Jin Tian, Yuhao Wang, and Junzhe Zhang. We are also grateful to the Editors, Professors Rina Dechter, Hector Geffner and Joe Halpern, for the opportunity to contribute to this special volume. Elias Bareinboim and Juan D. Correa were partially supported by grants from NSF IIS-1704352 and IIS-1750807 (CAREER). Duligur Ibeling was supported by the NSF Graduate Research Fellowship Program under Grant No. DGE-1656518. Thomas Icard was partially supported by the Center for the Study of Language and Information.

Bibliography

J. Aldrich. 1989. Autonomy. *Oxford Economic Papers*, 41: 15–34.

H. Bang and J. M. Robins. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4): 962–973.

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*, pp. 113–120. AUAI Press.

E. Bareinboim and J. Pearl. 2016. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27): 7345–7352.

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*, pp. 1–17. Springer Berlin Heidelberg, Berlin, Heidelberg.

E. Bareinboim, A. Forney, and J. Pearl. 2015. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems*, pp. 1342–1350.

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

E. Bareinboim, S. Lee, and J. Zhang. 2020b. An introduction to causal reinforcement learning: New challenges and opportunities. Technical Report R-65, Causal AI Lab, Columbia University.

E. W. Beth. 1956. On Padoa's method in the theory of definition. *Journal of Symbolic Logic*, 2(1): 194–195.

C. M. Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

R. Briggs. 2012. Interventionist counterfactuals. *Philosophical Studies*, 160(1): 139–166.

D. Buchsbaum, S. Bridgers, D. S. Weisberg, and A. Gopnik. 2012. The power of possibility: causal learning, counterfactual reasoning, and pretend play. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1599): 2202–2212.

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

N. Chomsky. 1959. On certain formal properties of grammars. *Information & Control*, 2: 137–167.

J. D. Correa and E. Bareinboim. 2019. From Statistical Transportability to Estimating the Effect of Stochastic Interventions. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 1661–1667. IJCAI Organization.

J. D. Correa and E. Bareinboim. 2020. A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. AAAI Press.

J. D. Correa, J. Tian, and E. Bareinboim. 2019a. Adjustment Criteria for Generalizing Experimental Findings. In *Proceedings of the 36th International Conference on Machine Learning*.

J. D. Correa, J. Tian, and E. Bareinboim. 2019b. Identification of Causal Effects in the Presence of Selection Bias. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*. AAAI Press.

D. Danks. 2014. *Unifying the Mind: Cognitive Representations as Graphical Models*. MIT Press.

A. P. Dawid. 2000. Causal Inference Without Counterfactuals (with Comments and Rejoinder). *Journal of the American Statistical Association*, 95(450): 407–448.

A. Deaton. 1992. *Understanding Consumption*. Oxford University Press.

R. Fagin, J. Y. Halpern, and N. Megiddo. 1990. A logic for reasoning about probabilities. *Information and Computation*, 87(1/2): 78–128.

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

R. A. Fisher. 1936. Design of Experiments. *British Medical Journal*, 1(3923): 554.

A. Forney, J. Pearl, and E. Bareinboim. 2017. Counterfactual Data-Fusion for Online Reinforcement Learners. In *Proceedings of the 34th International Conference on Machine Learning*.

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

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

A. Ghassami, S. Salehkaleybar, N. Kiyavash, and E. Bareinboim. 2018. Budgeted experiment design for causal structure learning. In J. Dy and A. Krause, eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 1724–1733. PMLR, Stockholm, Sweden.

A. Gopnik. 2012. Scientific thinking in young children: Theoretical advances, empirical research, and policy implications. *Science*, 337(6102): 1623–1627.

A. Gopnik, C. N. Glymour, D. M. Sobel, L. E. Schulz, T. Kushnir, and D. Danks. 2004. A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111(1): 3–32.

T. Haavelmo. 1 1943. The Statistical Implications of a System of Simultaneous Equations. *Econometrica*, 11(1): 1.

J. Y. Halpern. 1998. Axiomatizing Causal Reasoning. In G.F. Cooper and S. Moral, eds., *Uncertainty in Artificial Intelligence*, pp. 202–210. Cornell University, Morgan Kaufmann, San Francisco, CA.

J. Y. Halpern. 2000. Axiomatizing causal reasoning. *J. Artif. Intell. Res.*, 12: 317–337.

J. Y. Halpern. 2013. From causal models to counterfactual structures. *Review of Symbolic Logic*, 6(2): 305–322.

D. G. Horvitz and D. J. Thompson. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260): 663–685.

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)*, pp. 1149–1156. AAAI Press, Menlo Park, CA.

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

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

P. Hünermund and E. Bareinboim. Dec 2019. Causal inference and data-fusion in econometrics. Technical Report R-51 , <https://causalai.net/r51.pdf>, Causal Artificial Intelligence Lab, Columbia University.

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*, pp. 1868–1874.

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*.

T. Icard. 2020. Calibrating generative models: The probabilistic Chomsky-Schützenberger hierarchy. *Journal of Mathematical Psychology*, 95.

G. W. Imbens and D. B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, USA.

A. Jaber, J. Zhang, and E. Bareinboim. Aug 2018. Causal identification under Markov equivalence. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, pp. 978–987. AUAI Press.

A. Jaber, J. Zhang, and E. Bareinboim. 2019a. On causal identification under Markov equivalence. In S. Kraus, ed., *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 6181–6185. IJCAI Organization.

A. Jaber, J. Zhang, and E. Bareinboim. 2019b. Identification of conditional causal effects under Markov equivalence. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, eds., *Advances in Neural Information Processing Systems 32*, pp. 11512–11520. Curran Associates, Inc.

A. Jaber, J. Zhang, and E. Bareinboim. 2019c. Causal identification under Markov equivalence: Completeness results. In K. Chaudhuri and R. Salakhutdinov, eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2981–2989. PMLR, Long Beach, CA.

Y. Jung, J. Tian, and E. Bareinboim. 2020. Estimating causal effects using weighting-based estimators. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. AAAI Press.

M. Kocaoglu, K. Shanmugam, and E. Bareinboim. 2017. Experimental design for learning causal graphs with latent variables. In *Advances in Neural Information Processing Systems 30*, pp. 7018–7028. Curran Associates, Inc.

M. Kocaoglu, A. Jaber, K. Shanmugam, and E. Bareinboim. 2019. Characterization and learning of causal graphs with latent variables from soft interventions. In *Advances in Neural Information Processing Systems 32*, pp. 14346–14356. Curran Associates, Inc., Vancouver, Canada.

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. *Journal of the Royal Statistical Society*, 29: 105–117.

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.

S. Lee, J. D. Correa, and E. Bareinboim. 2020. Generalized transportability: Synthesis of experiments from heterogeneous domains. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. AAAI Press, New York, NY.

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

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

P. Machamer, L. Darden, and C. F. Craver. 2000. Thinking about mechanisms. *Philosophy of Science*, 67(1): 1–25.

J. L. Mackie. 1980. *The Cement of the Universe: A Study of Causation*. Clarendon Press, Oxford.

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

T. Maudlin. 2019. The why of the world. *Boston Review*. http://bostonreview.net/science-nature/tim-maudlin-why-world.

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

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

J. Pearl. 1993. Aspects of graphical models connected with causality. *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.

J. Pearl. 2000. *Causality: Models, Reasoning, and Inference*, 2nd. Cambridge University Press, New York, NY, USA.

J. Pearl. 2001. Bayesianism and Causality, or, Why I am Only a Half-Bayesian. In D. Corfield and J. Williamson, eds., *Foundations of Bayesianism, Applied Logic Series, Volume 24*, pp. 19–36. Kluwer Academic Publishers, the Netherlands.

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

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*, pp. 151–179. John Wiley and Sons, Ltd, Chichester, UK.

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

J. Pearl. 2018a. A personal journey into Bayesian networks. Technical Report R-476, <http://ftp.cs.ucla.edu/pub/stat_ser/r476.pdf>, Department of Computer Science, University of California, Los Angeles, CA.

J. Pearl. 2018b. Does obesity shorten life? or is it the soda? on non-manipulable causes. *Journal of Causal Inference*, 6(2).

J. Pearl. 2019. On the interpretation of $do(x)$. *Journal of Causal Inference*, 7(1).

J. Pearl and E. Bareinboim. 2019. A note on 'generalizability of study results'. *Journal of Epidemiology*, 30: 186–188.

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 P. Besnard and S. Hanks, eds., *Uncertainty in Artificial Intelligence 11*, pp. 444–453. Morgan Kaufmann, San Francisco.

D. C. Penn and D. J. Povinelli. 2007. Causal cognition in human and nonhuman animals: A comparative, critical review. *Annual Review of Psychology*, 58: 97–118.

E. Perkovic, J. Textor, M. Kalisch, and M. H. Maathuis. 2017. Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs. *J. Mach. Learn. Res.*, 18(1): 8132–8193.

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.

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. *Mathematical Modeling*, 7: 1393–1512.

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.

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. *Journal of Educational Psychology*, 66(5): 688–701.

W. C. Salmon. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, Princeton.

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*, volume 2, pp. 1219–1226.

H. A. Simon. 1953. Causal ordering and identifiability. In W. C. Hood and T. C. Koopmans, eds., *Studies in Econometric Method*, pp. 49–74. Wiley and Sons, Inc., New York, NY.

S. A. Sloman and D. Lagnado. 2015. Causality in thought. *Annual Review of Psychology*, 66(3): 1–25.

M. E. Sobel. 1990. Effect Analysis and Causation in Linear Structural Equation Models. *Psychometrika*, 55(3): 495–515.

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

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

L. J. Stockmeyer. 1977. The polynomial-time hierarchy. *Theoretical Computer Science*, 3: 1–22.

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

P. Suppes and M. Zanotti. 1981. When are probabilistic explanations possible? *Synthese*, 48: 191–199.

R. S. Sutton and A. G. Barto. 2018. *Reinforcement Learning: An Introduction*, second. 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)*, pp. 567–573. AAAI Press/The MIT Press, Menlo Park, CA.

J. Tian and J. Pearl. 2002b. On the Testable Implications of Causal Models with Hidden Variables. *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, pp. 519–527.

M. J. Van Der Laan and D. Rubin. 2006. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1).

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

J. Woodward. 2002. What is a mechanism? a counterfactual account. *Philosophy of Science*, 69: 366–377.

J. Woodward. 2003. *Making Things Happen*. Oxford University Press, New York, NY.

J. Woodward. 2016. Causation and manipulability. In E. N. Zalta, ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016.

G. H. von Wright. 1971. *Explanation and Understanding*. Cornell University Press.

J. Zhang. 2008. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17): 1873–1896.

J. Zhang. 2013. A Lewisian logic of causal counterfactuals. *Minds and Machines*, 23: 77–93.

J. Zhang and E. Bareinboim. 2018a. Equality of opportunity in classification: A causal approach. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds., *Advances in Neural Information Processing Systems 31*, pp. 3671–3681. Curran Associates, Inc.

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

J. Zhang and E. Bareinboim. 2018c. Non-parametric path analysis in structural causal models. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, pp. 653–662. AUAI Press.