

Causal Machine Learning: A Survey and Open Problems

Jean Kaddour^{*,1}, Aengus Lynch^{*,1}, Qi Liu², Matt J. Kusner¹, Ricardo Silva¹

*Equal contribution. ¹University College London. ²University of Oxford.

{jean.kaddour.20, aengus.lynch.17}@ucl.ac.uk.

22 July 2022.

Abstract

Causal Machine Learning (CAUSALML) is an umbrella term for machine learning methods that formalize the data-generation process as a structural causal model (SCM). This perspective enables us to reason about the effects of changes to this process (interventions) and what would have happened in hindsight (counterfactuals). We categorize work in CAUSALML into five groups according to the problems they address: (1) causal supervised learning, (2) causal generative modeling, (3) causal explanations, (4) causal fairness, and (5) causal reinforcement learning. We systematically compare the methods in each category and point out open problems. Further, we review data-modality-specific applications in computer vision, natural language processing, and graph representation learning. Finally, we provide an overview of causal benchmarks and a critical discussion of the state of this nascent field, including recommendations for future work.

Contents

1	Introduction	1
2	Causality: A Minimal Introduction	4
2.1	Bayesian Networks	4
2.2	Causal Bayesian Networks	7
2.3	Structural Causal Models	11
2.4	Causal Representation Learning	14
2.5	Spurious Relationships due to Confounding	15
2.6	Causal Estimand Identification	16
2.7	Causal Influence	17
3	Causal Supervised Learning	18
3.1	Invariant Feature Learning	19
3.2	Invariant Mechanism Learning	32
3.3	Open Problems	35
4	Causal Generative Modeling	37
4.1	Structural Assignment Learning	38
4.2	Causal Disentanglement	43
4.3	Open Problems	47
5	Causal Explanations	50
5.1	Feature Attribution Explanations	51
5.2	Contrastive Explanations	55
5.3	Open Problems	58
6	Causal Fairness	61
6.1	Two more detailed examples	63
6.2	Counterfactual Fairness Criteria	64
6.3	Interventional Fairness	66
6.4	Fairness under Distribution Shifts	67

6.5 Open Problems	68
7 Causal Reinforcement Learning	70
7.1 Isn't RL already "Causal"?	70
7.2 Causal Bandits	72
7.3 Model-Based RL	75
7.4 Multi-Task RL	78
7.5 Off-Policy Policy Evaluation	82
7.6 Imitation Learning	86
7.7 Credit Assignment	90
7.8 Counterfactual Data Augmentation	92
7.9 Agent Incentives	94
7.10 Open Problems	98
8 Modality-specific Applications	101
8.1 Causal Computer Vision	101
8.2 Causal Natural Language Processing	110
8.3 Causal Graph Representation Learning	118
9 Causal Benchmarks	122
9.1 Reinforcement Learning	122
9.2 Computer Vision	124
9.3 Natural Language Processing	125
10 The Good, the Bad and the Ugly	127
10.1 The Good	127
10.2 The Bad	128
10.3 The Ugly	130
11 Related Work	132
11.1 Other Surveys	132
11.2 Machine Learning for Causal Inference	133
12 Conclusion	138

1

Introduction

Today, machine learning (ML) techniques excel at finding associations in independent and identically distributed (i.i.d.) data. A few fundamental principles, including empirical risk minimization, backpropagation, and inductive biases in architecture design, have led to enormous advances for solving problems in fields like computer vision, natural language processing, graph representation learning, and reinforcement learning.

However, new challenges have arisen when deploying these models to real-world settings. These challenges include: (1) large reductions in generalization performance when the data distribution shifts [1], (2) a lack of fine-grained control of samples from generative models [2], (3) biased predictions reinforcing unfair discrimination of certain sub-populations [3, 4], (4) overly abstract and problem-independent notions of interpretability [5], and (5) unstable translation of reinforcement learning methods to real-world problems [6].

Multiple works have argued that these issues are partly due to the lack of causal formalisms in modern ML systems [7, 8, 9, 10, 11]. Subsequently, there has been a surge of interest in the research community on *causal machine learning* (CAUSALML), which are machine learning methods that utilize causal knowledge about the to-be-modeled system.¹ This survey covers how causality can address open ML problems.

In a nutshell, causal inference provides a language for formalizing structural knowledge about the data-generating process (DGP) via *Structural Causal Models* (SCMs) [12]. With SCMs, one can estimate what will happen to data after changes (called *interventions*) are made to its generating process. Going one step further, they also allow us to model the consequences of changes in hindsight while taking into account what happened (called *counterfactuals*). We introduce these concepts in more detail in Chapter 2, assuming no prior knowledge of causality.

Despite the extensive work on designing various types of CAUSALML algorithms, a clear categorization of its problems and methodology is lacking. We believe this is partly explained by the fact that CAUSALML usually involves assumptions about the data unfamiliar to large parts of ML. These assumptions are often tricky to relate across different problem setups (and untestable), making it difficult to measure progress and applicability. These issues are what motivates this survey.

¹In parallel, there is a growing trend of using modern machine learning techniques to estimate causal quantities, e.g., treatment effects, which is *not* the focus of this survey, see Sec. 11.2 for more details on this line of work.

Contributions

1. We give a minimal introduction to key concepts in causality that is completely self-contained (**Chapter 2**). We do not assume any prior knowledge of causality. Throughout, we give examples of how these concepts can be applied to help further ground intuition.
2. We taxonomize existing CAUSALML work into **causal supervised learning** (Chapter 3), **causal generative modeling** (Chapter 4), **causal explanations** (Chapter 5), **causal fairness** (Chapter 6), **causal reinforcement learning** (Chapter 7). For each problem class, we compare existing methods and address avenues for future work.
3. We review data-modality-specific applications in **computer vision**, **natural language processing**, and **graph representation learning** (Chapter 8), and **causal benchmarks** (Chapter 9).
4. We discuss *the good, the bad, and the ugly* (Chapter 10): what benefits CAUSALML provides compared to non-causal ML methods (*the good*), what challenges the field faces at the time of this writing (*the bad*), and what inevitable price one has to pay for using CAUSALML techniques (*the ugly*).

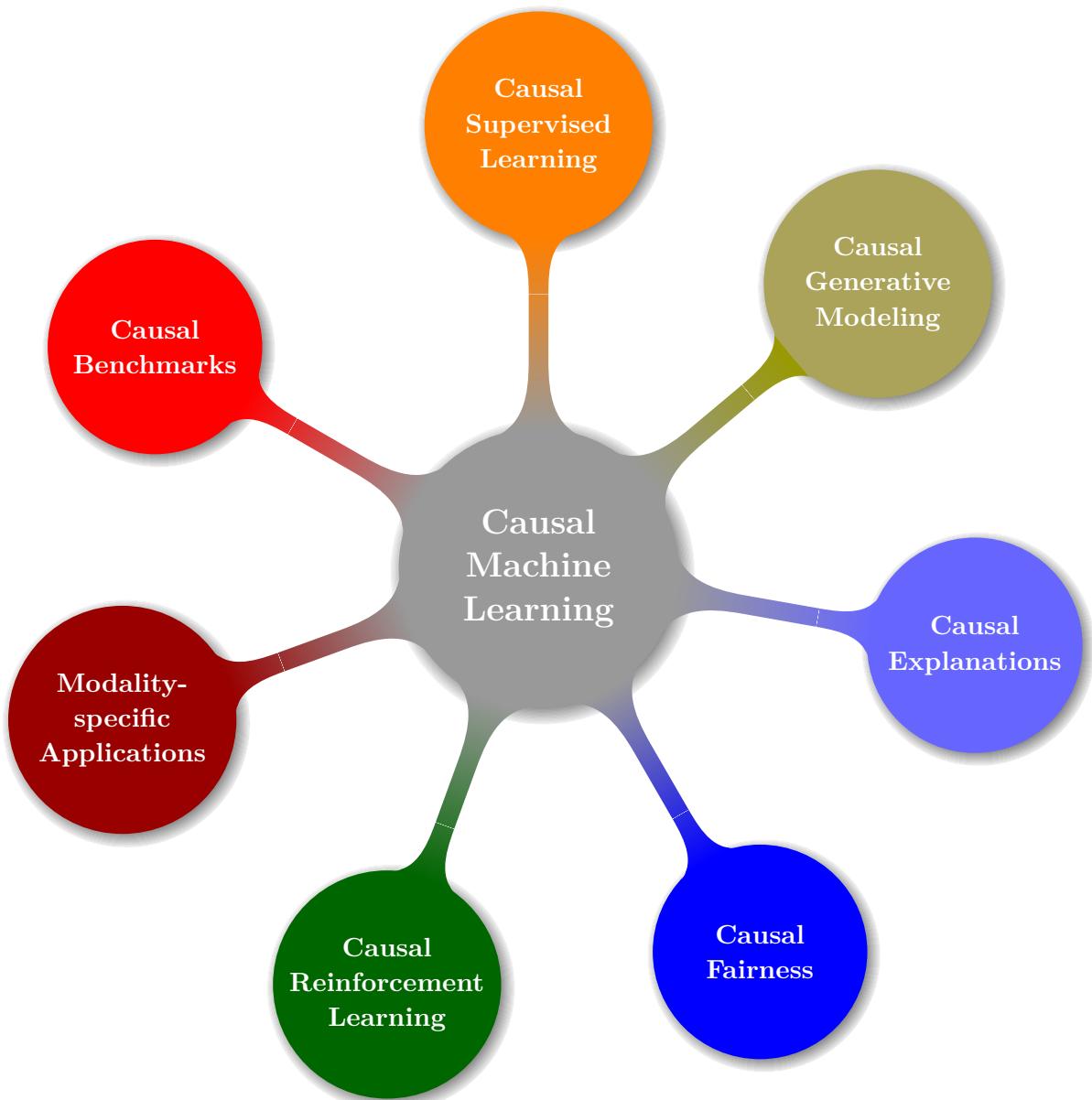


Figure 1.1: Taxonomy of Causal Machine Learning.

2

Causality: A Minimal Introduction

This section informs the reader of the concepts of causality used in CAUSALML research. We leave out many formal statements, e.g., proofs, and direct readers interested in such further details to [7, 12, 13, 14].

Notation

\mathcal{G}	Graph
$\text{de}(X_i), \text{de}_i$	Descendants of X_i
$\text{an}(X_i), \text{an}_i$	Ancestors of X_i
$\text{pa}(X_i), \text{pa}_i$	Causal parents of X_i
$\text{do}(\cdot)$	do-operator
\mathcal{M}	Structural Causal Model

2.1 Bayesian Networks

To reason about the causal effects of some random variables on others, we need to formalize causal relations. The canonical representation of causal relations is a *causal directed acyclic graph* (causal DAG), also called a *causal diagram*. It can encode a priori assumptions about the causal structure of interest (e.g., from expert knowledge).

Before we define causal DAGs, we introduce some terminology to describe a DAG and *Bayesian Networks* (BNs): a probabilistic graphical model representing probabilistic relationships between random variables. From there, we motivate modeling causal relationships, as they complement BNs with the ability to reason about interventions and counterfactuals.

2.1.1 Graphs

A *graph* \mathcal{G} is a collection of *nodes*, and *edges* that connect (some of) the nodes. In a *directed* graph, the edges are directed: pointing from one node to another. Visually, arrows indicate this direction; notation-wise, we call nodes connected by one edge *adjacent*.

A *path* in a graph is any sequence of adjacent nodes, regardless of the direction of the edges that join them. For example, $A \leftarrow B \rightarrow C$ is a path, but not a *directed path*.

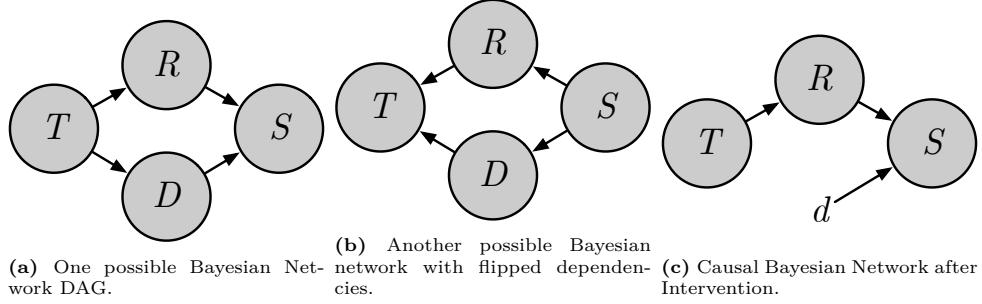


Figure 2.1: DAGs of Bayesian and Causal Bayesian Networks.

The latter is a path consisting of directed edges all directed in the same direction, e.g., $A \rightarrow B \rightarrow C$.

A *directed cycle* is a directed path that starts from a node A and ends in A . A *directed acyclic graph* (DAG) is a directed graph with no directed cycles.

In a DAG, edges point from a *parent* node into a *child* node. We denote the parents of a node X with $\text{pa}(X)$; and X an *ancestor* of Y (denoted by $X \in \text{an}(Y)$), and Y a *descendant* of X (denoted by $Y \in \text{de}(X)$) if there is a directed path that starts at node X and ends at node Y .

We denote a random node variable as X , assume that all distributions possess a mass or density function, and write $p(x)$ to represent its distribution.

2.1.2 Graphs as Joint Distribution Factorizations

We are interested in modeling probability distributions over random variables in both probabilistic and causal modeling. One reason why graphs are helpful for that is that they allow one to conveniently express how a joint distribution over a set of random variables factorizes. Specifically, we will show how graphs allow one to encode (conditional) independence relationships. To explain this, we adopt an introductory example given by Pearl [12].

Imagine that you are a farmer. You want to understand better the relationship between the time of the year T , duration of enabled lawn sprinkler D , amount of rain R , and pavement slipperiness S . To be able to conduct controlled simulations, you want to model the joint distribution $p(t, d, r, s)$.

The chain rule of probability (or product rule) always permits us to factorize a joint distribution over $N = 4$ variables as a product of N conditional distributions:

$$p(t, d, r, s) = p(t | d, r, s)p(d | r, s)p(r | s)p(s). \quad (2.1)$$

More generally, it states that

$$p(x_1, \dots, x_N) = p(x_1) \prod_i p(x_i | x_1, \dots, x_{i-1}). \quad (2.2)$$

Note how modeling this naive factorization takes an exponential number of parameters and quickly becomes intractable as N increases. Neal [13] illustrates this as

follows: Suppose that each x_i is binary and let us denote parameters by θ . For modeling the conditional distribution $p(x_n \mid x_{n-1}, \dots, x_1)$, we only need to model $\theta_n = p(X_n = 1 \mid x_{n-1}, \dots, x_n)$ because

$$p(X_n = 0 \mid x_{n-1}, \dots, x_n) = 1 - p(X_n = 1 \mid x_{n-1}, \dots, x_n). \quad (2.3)$$

However, this implies that we need 2^{n-1} parameters to model the full joint distribution. For example, if $n = 4$, to model $p(x_4 \mid x_3, x_2, x_1)$, we now need to store

$$\theta_1 = p(x_4 \mid X_3 = 0, X_2 = 0, X_1 = 0), \quad (2.4)$$

$$\theta_2 = p(x_4 \mid X_3 = 0, X_2 = 0, X_1 = 1), \quad (2.5)$$

$$\vdots \quad (2.6)$$

$$\theta_8 = p(x_4 \mid X_3 = 1, X_2 = 1, X_1 = 1). \quad (2.7)$$

Fortunately, some of your farmer colleagues have developed a DAG that describes how the variables are related, shown in Fig. 2.1a. This DAG is helpful because we can use it to describe conditional independence relationships. For example, since the time of the year is independent of all the other variables, we only need to model the marginal $p(t)$ instead of $p(t \mid d, r, s)$. Overall, we can simplify the joint distribution such that, in sum, the conditionals depend only on four instead of six random variables (compared to Eq. (2.1)):

$$p(t, d, r, s) = p(t)p(d \mid t)p(r \mid t)p(s \mid d, r). \quad (2.8)$$

This simplification formally relies on the Markov condition.

Def.: 2.1.1: Markov Condition [12]

Given a graph \mathcal{G} of nodes \mathbf{X} with joint distribution $p(\mathbf{x})$, the Markov Condition states that the parents \mathbf{pa}_i of every node X_i make X_i independent of its non-descendants $\mathbf{X} \setminus \mathbf{de}_i$, i.e.,

$$p(x_i \mid \mathbf{pa}_i) = p(x_i \mid \mathbf{X} \setminus \mathbf{de}_i).$$

This condition immediately implies the following factorization of the joint distribution

$$p(\mathbf{x}) = \prod_i p(x_i \mid \mathbf{pa}_i).$$

This joint factorization is the product of all variables conditioned on their parents in the graph (if any). The core idea behind Bayesian Networks is to decompose a (potentially large) joint distribution $p(\mathbf{x})$ into several small conditional ones according to the assumed DAG relations.

ML Perspective: 2.1.1: Autoregressive Distribution Factorizations

Autoregressive models are one example of modern ML techniques exploiting joint distribution factorizations based on conditional independence assumptions. For example, in generative modeling, some methods [15, 16] represent the distribution over an image \mathbf{x} with a DAG factorization under a particular pixel by pixel order, even if no independence is imposed:

$$p(\mathbf{x}) = \prod_{i=1}^N p(x_i | \mathbf{x}_{<i}), \quad (2.9)$$

where $\mathbf{x}_{<i} = [x_1, x_2, \dots, x_{i-1}]$ denotes the vector of pixels with index less than i . The advantage of doing so is, e.g., to make sampling easier or enable to learn conditionals in parallel.

Similarly, for sequence modeling tasks (e.g., language modeling with recurrent neural networks [17]), we often assume that the joint distribution over all tokens can be decomposed into conditionals that only depend on hidden states \mathbf{h} produced by a learnable function f_θ with parameters θ , i.e.,

$$p(\mathbf{x}) = \prod_{t=1}^T p(x_t | \mathbf{h}_{t-1}), \quad \mathbf{h}_t = f_\theta(x_t, \mathbf{h}_{t-1}). \quad (2.10)$$

2.2 Causal Bayesian Networks

2.2.1 Interventions

BNs give rise to utilizing structural knowledge about the distribution of interest to represent it more efficiently. However, BNs remain oblivious to *interventions* on the distribution's underlying process, as we will see in this subsection.

As previously described, the random variable D quantifies the duration of the sprinkler turned on. Imagine that you are interested in reasoning about what would happen if we fix it to value d , given that, e.g., there might be a non-linear relationship between the sprinkler and the slipperiness that you seek to understand better.

Using the Bayesian network toolbox, perhaps the most obvious strategy to pursue is first to infer the conditional distribution

$$p(t, r, s | d) = p(t)p(d | t)p(r | t)p(s | d, r). \quad (2.11)$$

Then, if we are only interested in $p(s | d)$, we can marginalize the other variables out using the sum rule. Alternatively, you may train an estimator for $\hat{p}(s | d)$.

However, what does $p(s | d)$ mean? It is the distribution of S given that we **observe** variable D for a fixed value, e.g., $d = 10$ minutes. In other words, we restrict our focus to those observations in which the sprinkler was set to d .

Let us think about the implications of examining this **observational** distribution: $p(s | d)$ accounts for the probability of observing d depending on the season of the year T via $p(d | t)$ and the amount of rain $p(r | t)$, as can be seen in Eq. (2.11).

Now, switching to the causal inference toolbox, our quantity of interest is typically an *intervention*, which we denote by the do-operator. Following the above example, we are interested in estimating $p(s | \text{do}(d))$. The difference in its meaning is that $p(s | \text{do}(d))$ denotes the distribution of S **intervened** upon the value of D , i.e., if we set it to d . This means that we are not considering a certain sub-population for which we observe $D = d$, but we reason about what happens to the (total) population after taking the **action** $\text{do}(d)$.

Returning to our example, Fig. 2.1c shows the new network representation after the action took place. Naturally, we remove the arrow $T \rightarrow D$ from Fig. 2.1a, because after intervening on D by setting it to a constant value¹, it does not depend on any other variables anymore.

Formally, there is a small yet significant difference compared to the conditional observational distribution in Eq. (2.11):

$$p(t, r, s | \text{do}(d)) = p(t) \underbrace{p(d | t)}_{=1} p(r | t) p(s | d, r) \neq p(t, r, s | d). \quad (2.12)$$

We no longer care about the relationship between the year's season and sprinklers before the action because that relationship is no longer in effect while we perform the action. Once we physically turn the sprinkler on, a new mechanism in which the season has no say determines the state of the sprinkler.

Theorem: 2.2.1: Truncated Factorization [12]

We assume that p and \mathcal{G} satisfy the Markov assumption and modularity. Given a set of intervention nodes \mathcal{S} , if \mathbf{x} is consistent with the intervention, then

$$p(x_1, \dots, x_n | \text{do}(s)) = \prod_{i \notin \mathcal{S}} p(x_i | \mathbf{pa}_i) \quad (2.13)$$

Otherwise, $p(x_1, \dots, x_n | \text{do}(s)) = 0$.

However, to yield a valid *Causal* BN that enables the estimation of the above interventional distributions, we have to make stronger assumptions than for a Bayesian network, which rest on causal (and not associational) knowledge about the underlying system. We do not cover these here, as we will soon introduce another formalism of causation in Sec. 2.3.

Our takeaway is that Causal BNs differ from regular BNs, e.g. because the conditional independence assumptions in regular BNs do not necessarily imply causality.

¹There also exists the idea of *soft* interventions in which D can still depend on other variables, but for the sake of exposition, we focus on the simpler case of a *hard* intervention here.

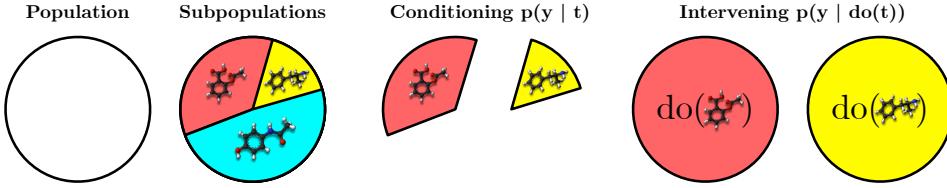


Figure 2.2: Conditioning versus intervening. Adapted from [13, 18].

However, their implied factorization will be valid for any recursive set of independencies and any ordering of the variables.

To illustrate the difference between the two, recall our initial example shown in Fig. 2.1a. Here, we conveniently made independence assumptions that seem causal already. Alternatively, we could have come up with an independence structure shown in Fig. 2.1b, where the edges are flipped

$$p(t, d, r, s) = p(t | d, r)p(d, r | s)p(s). \quad (2.14)$$

Based on common sense, the corresponding DAG does not seem causal. According to commonly known laws of nature, we would not expect the pavement's slipperiness to cause the amount of rain. Yet, for the purpose of *statistical* inference, it is perfectly adequate to factorize the joint distribution that way.

Another, more simplistic perspective might be that causal assumptions lie at the highest abstraction level for any modeling problem, above common statistical assumptions like the model's flexibility. In pure statistical inference (including most ML setups), we often ignore this level and do not impose strong assumptions on the causality between variables. Therefore, we do not yield the ability to compute interventions, etc.

2.2.1.1 Examples of Interventions

The following examples provide intuition on how conditional and interventional distributions differ.

Espresso Machine [19]: Imagine that Y is the pressure in an espresso machine's boiler, and X is the reading of the built-in barometer. Given a functioning barometer, $p(y | x)$ is a unimodal distribution centered around X , with randomness due to measurement noise. However, if we forcefully break the barometer and set it to 0, it will not affect the pressure in the tank. Therefore, $p(y | \text{do}(x)) = p(y) \neq p(y | x)$.

Medical Treatment [18]: Imagine a dataset where each observation $(\mathbf{x}_i, \mathbf{t}_i, y_i) \in \mathcal{D}$ represents a hospital patient's medical history record \mathbf{x}_i , prescribed drug treatment in form of a molecular graph \mathbf{t}_i , and health outcome y_i . Fig. 2.2 illustrates how in this setup the intervention $p(y | \text{do}(\mathbf{t}))$ refers to the scenario in which all patients described with medical history features \mathbf{x} take treatment \mathbf{t} versus the conditional distribution $p(y | \mathbf{t})$ restricting our focus to the subpopulation of \mathbf{X} that received \mathbf{t} .

2.2.2 Counterfactuals

Recall our farmer example from Sec. 2.1.2 and imagine that one day, you observe that the slipperiness S was very high, and even worse, a colleague of yours slipped and broke their arm. You want to determine under what circumstances S would have been reduced on that day. You come up with a hypothesis that you want to test: “*If we had turned the sprinkler off, the slipperiness would have been low.*”

Your hypothesis - an “*if*” statement in which the “*if*” portion is unrealized - is a *counterfactual* statement. It incorporates the factual data (D was non-zero on that day) and an intervention (setting D to 0), in which parts of the environment remain unchanged (T, R). Thereby, it paves the way to comparing two outcomes under the same conditions, differing only in one aspect.

To formalize a counterfactual statement, we must go beyond just the do-operator. For example, suppose that on that day, we observed slipperiness s , time of year t , rain r , and wetness w . Then, simply writing $p(s \mid \text{do}(d), s, t, r)$ leads to a clash between the hypothetical slipperiness and the actual slipperiness observed [20].

One way to clarify the distinction is to label the two outcomes of interests - the factual and counterfactual - with different subscripts. We denote the actual slipperiness by S and the counterfactual slipperiness under the intervention $D = d$ by S_d , such that our estimand becomes $p(s_d \mid s_0, t, r)$.

This notation is convenient to express our quantity of interest but does not operationalize the estimation of it. We require *Structural Causal Models* (SCMs) to do the latter. We will introduce these in Sec. 2.3 and conclude the current one with two more examples of counterfactuals.

2.2.2.1 Examples of Counterfactuals

Intuitively, counterfactuals are hypothetical retrospective interventions given an observed outcome. They help to explain the data since we can analyze the changes resulting from manipulating each variable.

After comparing conditional and interventional distributions in Sec. 2.2.1.1, we now contrast interventional and counterfactual ones.

Medical Treatment: Let us revisit the medical treatment example in Sec. 2.2.1.1, where we described $p(y \mid \mathbf{x}, \text{do}(\mathbf{t}))$ as the scenario in which all patients described with medical history features \mathbf{x} take treatment \mathbf{t} . Perhaps the most obvious counterfactual quantity we can think of in this scenario is $p(y_{\mathbf{t}'} \mid \mathbf{x}, \mathbf{t}, y)$.

How do $p(y \mid \mathbf{x}, \text{do}(\mathbf{t}'))$ and $p(y_{\mathbf{t}'} \mid \mathbf{x}, \mathbf{t}, y)$ semantically differ? The latter can be interpreted as *imagining* the former “after the fact” that $\mathbf{x}, \mathbf{t}, y$ occurred. So instead of asking “what happens if we give treatment \mathbf{t} to patient \mathbf{x} ?”, the latter is retrospective, asking “what would have happened if we had given treatment \mathbf{t}' to patient \mathbf{x} instead of \mathbf{t} ?“.

Layer	Activity	Semantics	Example
(1) Associational $p(y x)$	Seeing 	How would seeing x change my belief in Y ?	What does a symptom tell us about the disease?
(2) Interventional $p(y \text{do}(x), z)$	Doing 	What happens to Y if I do x ?	What if I take aspirin, will my headache be cured?
(3) Counterfactual $p(y_{x'} x, y)$	Imagining 	Was it x that caused Y ?	Was it the aspirin that stopped my headache?

Table 2.1: Pearl’s Ladder of Causation [12] (also called *Pearl’s Causal Hierarchy* [21]). Distributions on one layer virtually underdetermine information at higher layers, e.g., counterfactuals subsume interventional and associational ones.

Image Editing: Let $\mathbf{X} \in \mathbb{R}^D$ be an image. Commonly, we presume that we can describe \mathbf{X} by lower-dimensional, semantically meaningful factors of variation. For example, assume we observe an image \mathbf{x} and infer its (latent) primary object of interest \mathbf{o} and background features \mathbf{b} . We can then sample images with edited backgrounds in a controlled manner by sampling from the counterfactual distribution $\tilde{\mathbf{x}} \sim p(\mathbf{x}_{\mathbf{b}'} | \mathbf{x}, \mathbf{o}, \mathbf{b})$.

How does $p(\mathbf{x} | \mathbf{o}, \text{do}(\mathbf{b}'))$ differ from $p(\mathbf{x}_{\mathbf{b}'} | \mathbf{x}, \mathbf{o}, \mathbf{b})$? If we sample images from the former distribution, depending on its variance, we may get a very diverse set of images with object and background features \mathbf{o}, \mathbf{b} , respectively. If we sample from the latter, we expect sampled images that look identical except for their background.

2.2.3 Pearl’s Ladder of Causation

Table 2.1 shows Pearl’s three-layer ladder of causation [12], summarizing the differences between associational (or *observational*), interventional and counterfactual distributions. It is also called the *Causal Hierarchy*, because questions at level $i \in \{1, 2, 3\}$ can only be answered if information from level $j \geq i$ is available. Counterfactuals subsume interventional and associational questions, so they sit at the top of the hierarchy. Models that can answer counterfactual questions can also answer questions about observations and interventions, as we will shortly see in Sec. 2.3.

2.3 Structural Causal Models

In Sec. 2.2, we learned how causal BNs allow us to move from associational distributions in regular BNs to interventional ones. However, we were not able to construct counterfactual distributions with causal BNs. In this section, we learn another causation formalism that permits counterfactual analysis.

This formalism is a Structural Causal Model (SCM), sometimes called *Structural Equation Model* or *Functional Causal Model* [12]. In SCMs, we express causal relationships through deterministic, functional equations. This formalism reflects Laplace’s conception of natural laws being deterministic and randomness being a purely epistemic notion [12]. Hence, we introduce stochasticity in SCMs based on the assumption that certain variables in the equations remain unobserved.

Def.: 2.3.1: Structural Causal Model

An SCM $\mathcal{M} := (\mathbf{S}, p(\boldsymbol{\epsilon}))$ consists of structural assignments $\mathbf{S} = \{f_i\}_{i=1}^N$,

$$x_i := f_i(\epsilon_i; \mathbf{pa}_i), \quad (2.15)$$

where \mathbf{pa}_i is the set of parents of x_i (its direct *causes*), and a joint distribution $p(\boldsymbol{\epsilon}) = \prod_{i=1}^N p(\epsilon_i)$ over mutually independent exogenous noise variables^a (i.e. unaccounted sources of variation).

For every SCM, we yield a DAG \mathcal{G} by adding one vertex for each X_i and directed edges from each parent in \mathbf{pa}_i (the causes) to child X_i (the effect).

^aThere is no formal need for these variables to be independent, but for the purposes of the survey it will suffice to assume so.

Eq. (2.15) means that, in any SCM, we have that each variable X_i is caused by parent variables, and unobserved exogenous “noise” variables ϵ_i . To denote that the noise variables may play a major role causally, sometimes we use the more standard random variable notation U_i to represent them. However, exogenous variables are present in every SCM, and thus, we often omit them from causal graphs for brevity.

Since every SCM induces a (causal) graph, it also implies the previously introduced Markov condition in Def. 2.1.1. We call this the *causal* Markov condition, as the DAG contains the causal relationships among variables.

Theorem: 2.3.1: Causal Markov Condition [12]

Every SCM \mathcal{M} entails a joint density $p_{\mathcal{M}}(\mathbf{x})$ such that each variable X_i is independent of all its non-descendants given its parents \mathbf{pa}_i in \mathcal{G} .

2.3.1 Interventions

We already learned the idea of estimating an intervention using a causal BN in Sec. 2.2.1. An SCM allows us to predict the effects of interventions, too: we substitute one or multiple of its structural assignments with the intervention’s value.

The SCM view further highlights the difference between interventions $p(y | \text{do}(\mathbf{x}'))$ and counterfactuals $p(y_{\mathbf{x}'} | \text{do}(\mathbf{x}'), \mathbf{x})$: interventions operate at the population level in the sense that, if we construct the interventional distribution through an SCM, its exogenous noise terms still consist of the prior distribution $p(\boldsymbol{\epsilon})$, and not from the posterior $p(\boldsymbol{\epsilon} | \mathbf{x})$, which incorporates our knowledge of what already happened.

The solution to a counterfactual query is an individual-level answer to a “what would have happened if” question where all the exogenous variable sources had been controlled, leaving information to pass through a (modified) set of structural equations. Information about the exogenous variables can be obtained from observable variables already realized.

2.3.2 Counterfactual Inference

To compute counterfactuals, we can manipulate an existing SCM and turn it into a counterfactual one. To do so, we estimate the exogenous noise terms $p(\epsilon | \mathbf{x})$ given the observed datum \mathbf{x} .

Def.: 2.3.2: Counterfactual SCM [7]

Consider an SCM $\mathcal{M} = (\mathbf{S}, p(\epsilon))$ over nodes \mathbf{X} . Given observations \mathbf{x} , we define a counterfactual SCM by replacing the prior distribution of noise variables $p(\epsilon)$ with the posterior $p(\epsilon | \mathbf{x})$:

$$\mathcal{M}_{\mathbf{x}} := (\mathbf{S}, p(\epsilon | \mathbf{x})). \quad (2.16)$$

Given a counterfactual SCM $\mathcal{M}_{\mathbf{x}}$, we yield counterfactual distributions by intervening on its structural assignments \mathbf{S} . To illustrate this, let $\tilde{\mathbf{S}}$ denote the modified structural assignments with intervention $\text{do}(x_i = \tilde{x}_i)$. Then, we denote the modified, counterfactual SCM as $\tilde{\mathcal{M}} := \mathcal{M}_{\mathbf{x}; \text{do}(\tilde{x}_i)}$. Finally, $\tilde{\mathcal{M}}$ yields the counterfactual distribution $p_{\tilde{\mathcal{M}}}(\mathbf{x})$.

We summarize this *counterfactual inference* procedure in the following.

Def.: 2.3.3: Counterfactual Inference [12]

We infer counterfactual queries through a three-step procedure:

1. **Abduction:** Infer $p(\epsilon | \mathbf{x})$, the state of the world (the exogenous noise ϵ) that is compatible with the observations \mathbf{x} .
2. **Action:** Replace the equations (e.g., $\text{do}(\tilde{x}_i)$) corresponding to the intervention, resulting in a modified SCM $\tilde{\mathcal{M}} = \mathcal{M}_{\mathbf{x}; \text{do}(\tilde{x}_i)} = (\tilde{\mathbf{S}}, p(\epsilon | \mathbf{x}))$.
3. **Prediction:** Use the modified model to compute $p_{\tilde{\mathcal{M}}}(\mathbf{x})$.

2.3.3 Independent Mechanism

A nice property of SCMs is the *Principle of Independent Mechanisms* (sometimes also called *autonomy* or *modularity*). It is analogous to the truncated factorization property that we previously discussed in the context of Causal Bayesian Networks (Theorem 2.2.1). Its basic premise is that interventions are local, and intervening on a variable X_i only changes the causal mechanism for X_i , leaving the other mech-

anisms invariant. This principle allows us to encode many different interventional distributions in a single graph [12].

Def.: 2.3.4: Principle of Independent Mechanisms [12, 13]

An SCM consists of autonomous modules $p(x_i | \mathbf{pa}_i)$,

$$p(\mathbf{x}) = p(x_1, \dots, x_D) = \prod_{i=1}^D p(x_i | \mathbf{pa}_i). \quad (2.17)$$

This principle implies that if we intervene on a subset of nodes $S \subseteq \{1, \dots, D\}$, then for all i , we have that

1. If $i \notin S$, then $p(x_i | \mathbf{pa}_i)$ remains unchanged.
2. If $i \in S$, then $p(x_i | \mathbf{pa}_i) = 1$ if x_i is the value that X_i was set to by the intervention; otherwise $p(x_i | \mathbf{pa}_i) = 0$.

2.4 Causal Representation Learning

The goal of representation learning is to retrieve low-dimensional representations \mathbf{Z} that summarize our high-dimensional data \mathbf{X} , where $\dim(\mathbf{Z}) \ll \dim(\mathbf{X})$. The learned representations then facilitate solving downstream tasks, as the features of interest (e.g., an object in the image) are typically not given explicitly in the granular input data (e.g., pixels). However, these representations often rely on spurious associations and yield entangled dimensions that are hard to interpret [22, 23, 24].

In contrast, *Causal representation learning* (CRL) assumes that an SCM over high-level causal variables generates the data \mathbf{X} . Representations \mathbf{Z} correspond to instances of these typically latent causal variables.

With access to the SCM, we can estimate the data distribution after interventions on these variables or infer counterfactuals for specific data points. Unfortunately, learning the entire SCM is difficult without extensive supervision or domain knowledge. This task consists of three components, summarized in Def. 2.4.1.

Def.: 2.4.1: Causal Representation Learning [11]

In *causal representation learning*, we aim to learn a set of causal variables \mathbf{Z} that generate our data \mathbf{X} , s.t. we have access to the following:

1. *Causal Feature Learning*: an injective mapping $g : \mathcal{Z} \rightarrow \mathcal{X}$ s.t. $\mathbf{X} = g(\mathbf{Z})$
2. *Causal Graph Discovery*: a causal graph $G_{\mathbf{Z}}$ among the causal variables \mathbf{Z}
3. *Causal Mechanism Learning*: the generating mechanisms $p_{G_{\mathbf{Z}}}(z_i | \mathbf{pa}(z_i))$ for $i = 1, \dots, \dim(\mathbf{Z})$

where $\mathbf{pa}(Z_i) \subset \{Z_j\}_{j \neq i} \cup \epsilon_i$ and ϵ_i is the exogenous causal parent of Z_i .

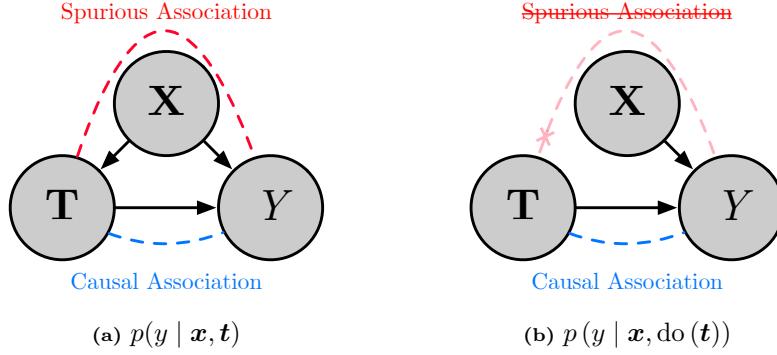


Figure 2.3: Spurious Relationship between \mathbf{T} and Y due to \mathbf{X} being an observed confounder. We adapt the Figures from [13].

2.5 Spurious Relationships due to Confounding

Recall the medical treatment example from Sec. 2.2.1.1, where we have a dataset where each observation $(\mathbf{x}_i, \mathbf{t}_i, y_i) \in \mathcal{D}$ represents a hospital patient’s medical history record \mathbf{x}_i , prescribed drug treatment \mathbf{t}_i , and health outcome y_i . In real-world scenarios, the patient’s pre-treatment health conditions \mathbf{x}_i influence both the doctor’s treatment prescription and outcome, thereby **X confounds** the effect of the treatment **T** on the outcome **Y** (and we call **X** a *confounder* or *confounding variable*). Moreover, we say that **T** and **Y** are confounded (by **X**), or *spuriously associated*.

Fig. 2.3 visualizes how associations flow in the observational $p(y | \mathbf{x}, \mathbf{t})$ and interventional distribution $p(y | \mathbf{x}, \text{do}(\mathbf{t}))$. In Fig. 2.3a, we see that $p(y | \mathbf{x}, \mathbf{t})$ entails both **causal** and **spurious** associations from **T** to **Y**, while $p(y | \mathbf{x}, \text{do}(\mathbf{t}))$ isolates the **causal** association from **T** to **Y**, as shown in Fig. 2.3b. The **causal** effect flows along directed paths, while **spurious** associations flow along all *unblocked* paths. To determine whether a path is unblocked, certain criteria need to be checked, which we do not cover here, but one can find in [12].

Here, we want to highlight the **spurious** association between **T** and **Y**: Imagine a doctor whose policy is to give expensive treatments to very ill patients with low recovery chances and cheap treatments to very healthy patients with high recovery chances. **Y** is a scalar that denotes the post-treatment health outcome; the higher, the better. Assuming that cheap and expensive treatments are equally effective, cheap and costly treatments are nonetheless positively and negatively correlated with health outcomes, respectively. This correlation is **spurious** because it is due to the doctor’s policy, based on the patient’s pre-treatment health conditions **X**, and not the treatment’s actual **causal** effect on the outcome.

When do spurious relationships become problematic? A simplified answer is whenever the confounder is unobserved (also called *hidden confounding*). The reason is that without further knowledge about the data-generating process, a sophisticated ML model will likely rely on spurious associations in the training dataset, which may not occur anymore when the model is in production. This reliance is a feature, not a bug, though: it would be wasteful if the model would not utilize spurious asso-

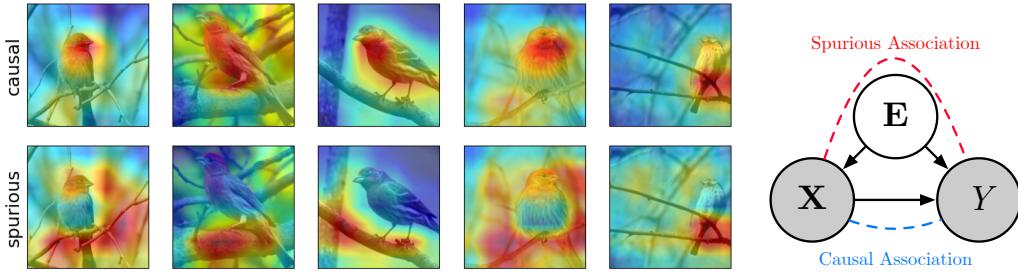


Figure 2.4: Spurious relationships due to hidden confounding in ImageNet [25, 26]. The hidden confounder animal environment \mathbf{E} caused images of birds to include trees and boughs. The heatmaps highlight causal and spurious associations between images \mathbf{X} and bird labels Y .

cations if we do not enforce it to avoid them. ML Perspective 2.5.1 illustrates how hidden confounding may harm classification models in a computer vision context. We explore remedies for this issue in Sec. 2.4.

ML Perspective: 2.5.1: Spurious Relationships in ImageNet [25, 26]

In ML benchmark datasets, unobserved confounding may occur too. Consider the image classification scenario with image \mathbf{X} , a label Y , and an unobserved variable \mathbf{E} that dictates the background of the classification object of interest, e.g., the environment of animals. It causes pictures of birds to include trees and boughs, as illustrated by Singla and Feizi [26] in Fig. 2.4. The authors find that classifiers trained on the ImageNet dataset rely on confounded relationships to classify objects, i.e., the same object in a different background environment is more likely to be misclassified.

2.6 Causal Estimand Identification

So far, we have discussed the semantic difference between observational and interventional distributions and the reasons for computing the latter. When is it feasible to estimate the latter? *Identification* of causal estimands refers to the process of moving from the causal estimand (e.g., $p(y \mid \text{do}(\mathbf{x}))$) to an equivalent statistical estimand (e.g., $p(y \mid \mathbf{x})$), which we can then estimate from data [13].

We call a causal estimand *identifiable* if it is possible to compute it from a purely statistical quantity. If it is not identifiable, then regardless of how much data we have, we will not be able to isolate the causal association of interest in our data.

In the absence of hidden confounders and as long as we know the causal graph, the causal estimand is identifiable. For example, suppose our estimand of interest is the *average treatment effect* $p(y \mid \text{do}(\mathbf{t}))$ in the scenario of Fig. 2.3a, i.e., the causal effect of treatment \mathbf{t} averaged over all possible patient features \mathbf{X} . Here, the *backdoor criterion* is satisfied, which makes \mathbf{X} a *valid adjustment set* [12]. The mathematical procedure outputting a statistical estimand is called *backdoor adjustment*.

Generally speaking, given any causal DAG (potentially including unobserved confounders), graphical tests allow us to determine the identifiability of specific causal estimands [27, 28]. Besides the backdoor criterion, a *frontdoor* criterion exists. We do not cover these criteria here; for this survey, it is sufficient to know about their existence, and that finding them can be automated [29]. We direct readers interested in learning more about them to excellent explanations in [7, 12, 13, 14, 30].

2.7 Causal Influence

Besides interventional and counterfactual queries, another common quantity of interest in causal inference is one variable's causal influence on another. For example, in Chapter 5, given an input vector \mathbf{x} and a black-box ML model $f_{\theta}(\cdot)$ with parameters θ , we will look at quantifying the causal influence an input feature x_i has on the model prediction $\hat{y} = f_{\theta}(\mathbf{x})$. Another example will occur in Sec. 7.7, where we measure the social influences of agents in multi-agent systems, e.g., how the action one agent takes influences the subsequent action of another agent.

Janzing et al. [31] postulate a set of natural, intuitive requirements that a measure of causal influence should satisfy. Then, they evaluate various information-theoretic measures on whether they meet these desiderata. Finally, they conclude that the KL-divergence is a suitable measure.

Def.: 2.7.1: Causal influence measured by KL-divergence [31]

Given an SCM \mathcal{M} , the causal influence $\mathcal{M}_{k \rightarrow l}$ from node k to node l is

$$\begin{aligned}\mathcal{M}_{k \rightarrow l} &= \sum_{\mathbf{pa}_l} D_{\text{KL}} \left[p_{\mathcal{M}}(x_l | \mathbf{pa}_l) \| p_{\widetilde{\mathcal{M}}}(x_l | \mathbf{pa}_l \setminus x_k) \right] p_{\mathcal{M}}(\mathbf{pa}_l) \\ &= D_{\text{KL}} \left[p_{\mathcal{M}}(x_l | \mathbf{pa}_l) \| p_{\widetilde{\mathcal{M}}}(x_l | \mathbf{pa}_l \setminus x_k) \right],\end{aligned}\quad (2.18)$$

where $D_{\text{KL}}(\cdot \| \cdot)$ denotes the KL-divergence and $p_{\widetilde{\mathcal{M}}}$ the interventional distribution after removing the edge from node k to node l .

3

Causal Supervised Learning

The goal of supervised learning is to learn the conditional distribution $p(y | \mathbf{x})$ by training on data of the form $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where \mathbf{X} and Y denote covariates and label, respectively. One of the most fundamental principles in supervised learning is to assume that our data \mathcal{D} is *independent and identically distributed* (i.i.d.). This assumption has strong implications. On the one hand, it allows us to split a set of observations into train, validation, and test dataset, opening up an easy way to perform model training, selection, and evaluation. On the other hand, it implies that the test set and, perhaps more importantly, unseen inputs occurring when the model is in production follow the same distribution as the training set. Put differently, the i.i.d. assumption says that “the past is indicative of the future”.

With no surprise, the validity of this assumption has been challenged [1, 6]; it has been famously called “the big lie in machine learning” [32]. Whenever we deploy our models in the real world, we have little to no control over the distribution we observe; e.g., variables can change in frequency (see Fig. 3.1), and novel feature combinations can occur that are not contained in the training set. Put simply, if the i.i.d. assumption breaks down, models relying on it will perform poorly [33].

Let us consider the example in Fig. 3.1, inspired by ML Perspective 2.5.1. In the training dataset, pictures of cows typically exhibit alpine pasture backgrounds, a spurious association caused by the cow’s natural habitat. If we train a model under the i.i.d. assumption, the model will rely on this spurious association for future predictions: this is a feature, not a bug. Naturally, problems will arise when we use the model in test settings where cow pictures do not include grass backgrounds. Then, our model is at risk of misclassifying cows.

This example illustrates how the i.i.d. assumption breaks down whenever the test distribution differs from the training distribution. As an alternative to the i.i.d. assumption, we can assume that our data is sampled from interventional distributions governed by an SCM. For a given dataset generated across a set of environments \mathcal{E} , $\{(\mathbf{x}_i^e, y_i^e)\}_{i=1}^N\}_{e \in \mathcal{E}}$, we view each environment $e \in \mathcal{E}$ as being sampled from a separate interventional distribution. Regarding the example in Fig. 3.1, we view the test dataset as being generated by interventions on the latent variables for background features.

How can we estimate $p(y | \mathbf{x})$ in a principled manner? In the following sections, we will discuss two classes of methods that aim to learn domain-robust, transferable *features* or *mechanisms*: in *invariant feature learning*, we learn a content representation



Figure 3.1: Image classifiers are prone to spurious relationships in the dataset [34]. Cows in ‘common’ contexts (e.g., Alpine pastures) are detected and classified correctly (A), while cows in uncommon contexts (beach, waves, and boat) are not detected (B). In such an instance, we can view (B) as a sample from a distribution with intervention on the background. We show the top five labels and confidence produced by ClarifAI.com.

C of the causal parents of Y , $\text{pa}(Y)$, such that $Y \sim p(y \mid c)$ across all environments. In *invariant mechanism learning*, we identify a set of mappings \mathcal{F} that allow us to predict Y from \mathbf{X} across a range of interventional distributions.

Notation

X	Observed covariates
<i>D</i>	$\dim(\mathbf{X})$
<i>Y</i>	Prediction target (label)
S	Spurious (or style) variables for prediction
C	Content variables for prediction ($\text{pa}(Y)$)
ϵ	Independent exogenous causal parents
<i>E</i>	Environment index
\mathcal{E}	Set of environments
U	Unobserved confounders
A	Protected attributes

3.1 Invariant Feature Learning

Invariant feature learning (IFL) is the task of identifying features of our data \mathbf{X} that are predictive of Y across a range of environments \mathcal{E} . From a causal perspective, the causal parents $\text{pa}(Y)$ are always predictive of Y under any interventional distribution except where Y itself has been intervened upon. This is because of the *Principle of Independent Mechanisms* [7] (Def. 2.3.4).

IFL methods often simplify the governing SCM to focus on identifying the causal parents of Y . We can abstract a complex SCM into a simple SCM by collecting the causal parents of Y into one variable, while the other variables are collected into

Available resources	Method	Key Idea	Ref.
Content-Invariant Transformations	Deconfound data	Perform data augmentations and train on the augmented data	Sec. 3.1.1.1
	Deconfound intermediate representations	Deconfound representations obtained from a pre-trained model	Sec. 3.1.1.2
	Deconfound model during training	Regularize the model to enforce invariances	Sec. 3.1.1.3
	Deconfound post-training predictions	Deconfound predictions by identifying and subtracting the confounder bias	Sec. 3.1.1.4
Multiple Environments	Invariant risk minimization	Identify content features through constrained optimization	Sec. 3.1.2.1
	Causal Matching	Match content features across environments	Sec. 3.1.2.2
	SCD as latent variables	Unconstrained optimization through latent variable modeling	Sec. 3.1.2.3
	Compositional Recognition	Leverage additional object and attribute labels in data	Sec. 3.1.2.4

Table 3.1: Overview of Invariant Feature Learning (IFL) Methods.

another. The most general abstraction we found in the literature is the *Style and Content Decomposition* (SCD) in Fig. 3.2 [35, 36, 37].

Def.: 3.1.1: Style and Content Decomposition [35, 36]

The *style and content decomposition* (SCD) is a causal graph of a data generating process (DGP) for \mathbf{X} and Y . We call \mathbf{S} the *style* variables and \mathbf{C} the *content* variables, where both are assumed to be latent. The content variables group all of the causal parents of Y , $\text{pa}(Y)$, while the style variables group the rest of the variables. The generations of \mathbf{X} and Y follow the distributions

$$\mathbf{X} \sim p(\mathbf{x} | \mathbf{s}, \mathbf{c}), \quad Y \sim p(y | \mathbf{c}). \quad (3.1)$$

Assuming that Def. 3.1.1 holds, we define Invariant Feature Learning as follows.

Def.: 3.1.2: Invariant Feature Learning [32, 38]

Invariant Feature Learning (IFL) aims to identify the content features \mathbf{C} that cause both \mathbf{X} and Y , and a mapping $p(y | \mathbf{c})$, such that

$$\mathbf{C} = \Phi(\mathbf{X}) \text{ s.t. } Y \sim p(y | \mathbf{c}) \quad (3.2)$$

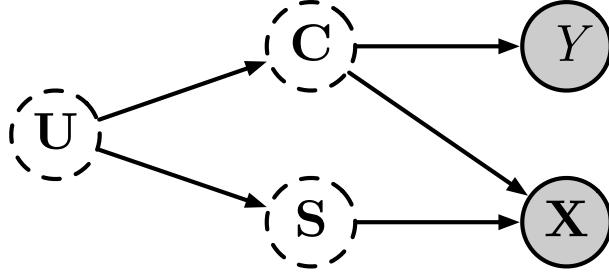


Figure 3.2: General set up of the Style and Content Decomposition: \mathbf{S} and \mathbf{C} both generate our data \mathbf{X} , but only \mathbf{C} generates \mathbf{Y} . The mapping $p(y | c)$ is assumed to be invariant across environments generated by interventions on \mathbf{S} . Often we observe spurious associations between \mathbf{S} and \mathbf{C} . This is explained by the presence of the unobserved confounder \mathbf{U} .

3.1.1 Content-invariant transformations

This section reviews IFL methods that extract content features by applying content-invariant transformations to the training dataset, intermediate representations, or predictions.

3.1.1.1 Deconfounding data

Data Augmentation (DA) is ubiquitous in modern machine learning pipelines involving high-dimensional datasets. The idea is to identify a set of transformations that can be applied to any sample in our dataset yet will not change its semantics (i.e., its content features). It requires domain knowledge from a practitioner about the types of permissible transformations that avoid diluting the content information. Its motivation is to enforce invariances in a model and, therefore, to improve generalization.

This section reviews works that motivate data augmentation through a causal perspective [39, 40, 41]. This perspective views data augmentations as constructing counterfactuals of instances, where we intervene on style features, akin to the Style and Content Decomposition (Fig. 3.2).

Causal Perspective: 3.1.1: Data augmentation [39, 40]

Data augmentation can be viewed as a set of interventions on the *style* variable in the *style and content decomposition* of Fig. 3.2, which breaks any spurious associations between style variables and \mathbf{Y} .

The augmentations in this section are hand-crafted. Later in Chapter 4, we will review how to generate counterfactual data augmentations using causal generative models.

Ilse et al. [40] explains that DA weakens the spurious association between observed domains and task labels. To show that, they introduce the concept of *intervention-augmentation* equivariance, formalizing the relationship between data augmentation and interventions on features caused by the domain. If intervention-augmentation

equivariance holds, one can use data augmentation to simulate new environments e using only observational data. This simulation removes the spurious association $E \leftarrow \mathbf{U} \rightarrow Y$ caused by a hidden confounder \mathbf{U} , which allows E to affect Y through the back door. Based on this insight, they derive an algorithm that can select data augmentation techniques from a list of transformations, leading to better domain generalization.

Kaushik et al. [42] propose a method of data augmentation via a human-in-the-loop process in which, given some documents and their (initial) labels, humans must revise the text with edits sufficient to flip the label and thus generate a counterfactual sample. Notably, the method prohibits edits not sufficient to flip the applicable label. Hence, they generate negative and positive samples, allowing for contrastive learning as an approach to representation learning.

Teney et al. [43] develop this idea further by incorporating samples with contrasting labels but similar semantic information (i.e. similar content variables), i.e. counterfactual samples, as a means of improving the training of model $f(\cdot)$ (see Chapter 4 for further comment). They propose a regularization term that enforces alignment between local gradients $\nabla_{\mathbf{x}} f(\mathbf{x}_i)$ and a ground truth gradient. The ground truth gradient mimics the translation in the input space necessary to switch the model output between two contrasting samples and their corresponding labels.

Specifically, for two counterfactual samples $\{(\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j)\}$ with $y_i \neq y_j$, define $\mathbf{g}_i = \nabla_{\mathbf{x}} f(\mathbf{x}_i)$ and $\hat{\mathbf{g}}_i = \mathbf{x}_j - \mathbf{x}_i$. Then their proposed regularization term enforces similarity between $\hat{\mathbf{g}}_i$ and \mathbf{g}_i . The method can be applied wherever a practitioner has access to counterfactual samples. The authors demonstrate improvements in generalization for vision and language tasks such as visual question answering.

Mao et al. [44] derive an optimal intervention strategy using data augmentations to mimic causal interventions and propose a loss function, GENINT, that incorporates data augmentations generated by GANs. They propose to empirically identify a suite of data augmentations that do not interfere with the object labels according to the method proposed in GANSPACE [45]. They intervene directly on latent factors in the original data, such as background and angle of view, as seen in Fig. 3.3. Extensive experiments demonstrate superior OOD classification accuracy on ImageNet-C using AlexNet [46] architecture compared to rival data augmentation strategies.

3.1.1.2 Deconfounding intermediate representations

In this setting, we have access to samples of \mathbf{X} and pre-trained representations \mathbf{R} . Mao et al. [47] propose a method (see Algorithm 1) that improves how the pre-trained representations, learned by self-supervised learning or otherwise, are leveraged for classification models. They assume that the data is generated by the causal structure shown in Fig. 3.4, with $\mathbf{X} \sim p(\mathbf{x} \mid \mathbf{c}, \mathbf{s})$ generated by style and content features, and \mathbf{R} the representation of \mathbf{X} that a pre-trained model produces. Importantly, they assume the presence of unobserved confounding $\mathbf{U}_{S,Y}$ between \mathbf{S}

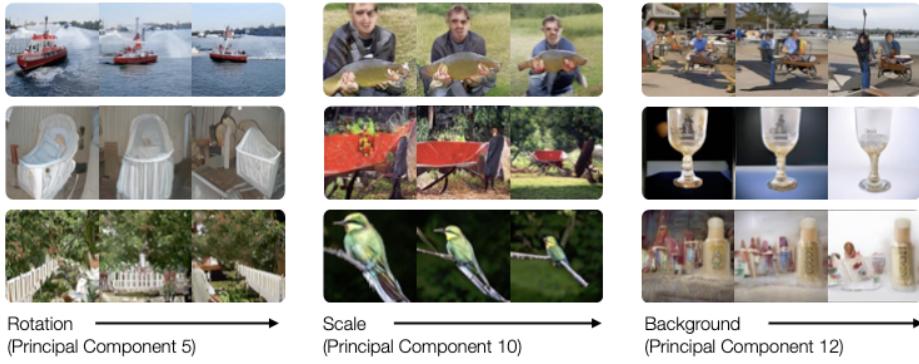


Figure 3.3: Mao et al. [44] use GANSpace [45] to generate a variety of data augmentations: some of the data augmentations are seen here. The data augmentations are interpreted as interventions on style features that our predictor $p(y | \mathbf{x})$ must be invariant to.

and Y , which naturally explains the cause of spurious associations between \mathbf{X} and Y in any training dataset.

Algorithm 1 Causal-Transportability Model Training [47]

Input: Training set D over $\{(\mathbf{X}, Y)\}$.
Phase 1: Compute $\hat{p}(\mathbf{r} | \mathbf{x})$ from representation of VAE or pre-trained model.
Phase 2: Sample $\mathbf{x}_i, \mathbf{r}_i, y_i \sim \mathcal{D}' := (\mathbf{X}, \mathbf{R}, Y)$
for $i = 1, \dots, K$ **do**
 Random sample \mathbf{x}'_i from the same category as \mathbf{x}_i
 Train $\hat{p}(y | \mathbf{x}', \mathbf{r})$ via minimizing $\mathcal{L}_{\text{class}}$
end for
Output: Model $\hat{p}(\mathbf{r} | \mathbf{x})$ and $\hat{p}(y | \mathbf{x}, \mathbf{r})$

Since \mathbf{S} and Y are confounded by $\mathbf{U}_{\mathbf{S}, Y}$, and $\mathbf{S} \rightarrow \mathbf{C}$ in the assumed causal graph (Fig. 3.4), $p(y | \mathbf{c})$ is confounded too. However, both $p(y | \mathbf{c}, \text{do}(\mathbf{s}))$ and $p(y | \text{do}(\mathbf{c}), \mathbf{s})$ are unconfounded (i.e. $p(y | \text{do}(\mathbf{x}))$ is unconfounded). Thus, Mao et al. [47] suggest that $p(y | \mathbf{c}, \text{do}(\mathbf{s}))$ is an invariant predictor. To simulate this intervention for a given sample \mathbf{x} , they obtain the representation via $p(\mathbf{r} | \mathbf{x})$, and a batch of corrupted images sampled randomly from the dataset. The corruption of the image is intended to destroy high level information and thus simulate interventions on \mathbf{S} features.

Mao et al. [47] introduce a training algorithm (Sec. 3.1.1.2) that learns $p(y | \mathbf{x}, \mathbf{r})$ by minimizing a classification loss $\mathcal{L}_{\text{class}}$, and a separate evaluation algorithm that estimates the causal quantity $p(y | \text{do}(\mathbf{x}))$. The output of the classification algorithm reads as $\hat{y} = \text{argmax}_y p(y | \text{do}(\mathbf{x}))$.

In experiments, the authors utilize pre-training methods such as SIMCLR [48] and SWAV [49] to yield \mathbf{R} . They then compare against the ERM loss for classification on OOD benchmarks such as ImageNet9 [50], demonstrating clear improvements. Their method performs best on COLOREDMNIST [32], where it outperforms IRM (Sec. 3.1.2.1) and GENINT [44] (Sec. 3.1.1.1).

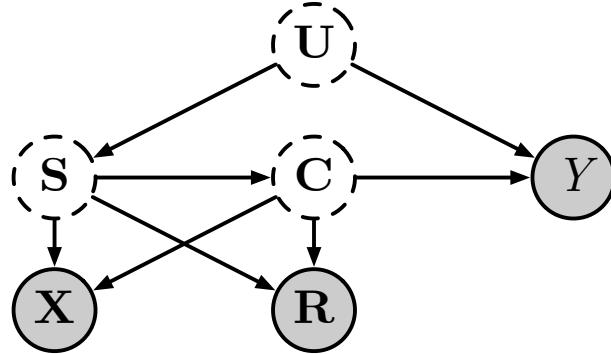


Figure 3.4: Causal Transportability [47] identifies causal parents from a set of representations: Mao et al. [47] propose an algorithm to make better use of pre-trained representations \mathbf{R} by considering the shown causal graph (Algorithm 1). \mathbf{S} and \mathbf{C} are the causal parents of \mathbf{X} . \mathbf{R} is a set of representations that have been identified using a feature learner, such as a pre-training algorithm, that carry at least all the information in \mathbf{C} . \mathbf{S} and \mathbf{Y} are confounded by a hidden variable \mathbf{U}_{XY} .

3.1.1.3 Deconfounding models during training

Counterfactual Invariance: *Counterfactual invariance* is a framework for constructing predictors whose predictions are invariant to certain perturbations on \mathbf{X} [51], using practitioner specifications. To define invariance, we specify an additional variable \mathbf{A} that captures information that should not influence predictions. However, \mathbf{A} may causally influence the covariates \mathbf{X} .

Consider the image dataset in Fig. 3.1 as a motivating example. We find that our classifier seems unable to identify cows when the background is changed from the mountains to a beach. In this instance, we identify *background* as \mathbf{A} , since changing our background should not affect the model prediction, but it does have a causal effect on the covariates \mathbf{X} .

Let \mathbf{X}_a denote the counterfactual \mathbf{X} we would have seen had \mathbf{A} been set to a , leaving all else fixed.

Def.: 3.1.3: Counterfactual Invariance [51]

A predictor $f(\cdot)$ is counterfactually invariant to \mathbf{A} if for any given sample \mathbf{x} , counterfactual $\mathbf{x}_a \sim p(\mathbf{x}_a | \mathbf{x})$, and $\forall a \in \mathcal{A}$, we have $f(\mathbf{x}) = f(\mathbf{x}_a)$.

This framework requires the practitioner to identify i) the causal direction $\mathbf{X} \rightarrow \mathbf{Y}$ or $\mathbf{X} \leftarrow \mathbf{Y}$, ii) the sensitive attributes \mathbf{A} and iii) whether the association between \mathbf{A} and \mathbf{Y} is due to confounding or selection bias in the data collection. See Fig. 3.5 for a graphical illustration. With this information, the practitioner can appropriately choose between one of two regularizers that enforce a counterfactual invariance signature (a necessary but not sufficient condition). This approach was restricted to binary \mathbf{A} in the implementation in Veitch et al. [51], but conceptually, one may extrapolate the idea to higher dimensions with appropriate conditioning on \mathbf{A} .

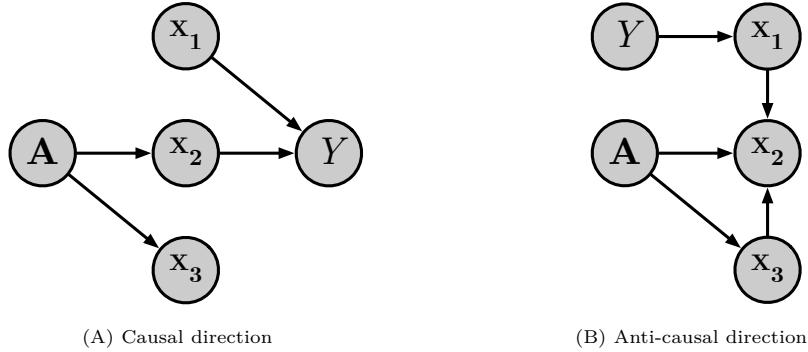


Figure 3.5: Counterfactual invariance causal graphs [51]: The counterfactual invariance framework accommodates the above causal models (Sec. 3.1.1.3). In both $\mathbf{X} \rightarrow Y$ and $\mathbf{X} \leftarrow Y$, \mathbf{A} is not a causal parent of Y and we aim to remove any spurious influence \mathbf{A} may have on our predictor $p(y | \mathbf{x})$. Thus, we aim to predict Y solely using \mathbf{X}_1 features.

If \mathbf{A} is a protected attribute (e.g., sex or race) that we would like to be an invariant input to the predictor, this definition is connected to *counterfactual fairness*, introduced in Def. 6.2.1.

Asymmetry learning: Similar to the counterfactual invariance framework, Mouli and Ribeiro [52] present *asymmetry learning* for learning classifiers which are counterfactually invariant to certain distribution shift settings. The practitioner specifies a set of equivalence relations in the data generating process to which a model shall be invariant. Mouli and Ribeiro [52] introduce a theoretical paradigm for out-of-distribution generalization that models the data generating process and specifies the types of test distributions we can generalize to. Asymmetry learning takes \mathbf{X} , \mathbf{Y} , the set of provided equivalence relations as input.

As a motivating example, consider the pendulum experiment included in [52]. Two global properties $\rho_1(\cdot), \rho_2(\cdot)$ are identified for any pendulum swing x , one of which is the initial potential energy of the system, $\rho_1(\cdot)$. Then, they define an equivalence relation \sim_1 such that for any two samples $x^{(1)}, x^{(2)}$ which have the same initial potential energy, $\rho_1(x^{(1)}) = \rho_1(x^{(2)})$, we say $x^{(1)} \sim_1 x^{(2)}$.

Each equivalence relation \sim_i induces a set of object transformations \mathcal{T}_i that preserve equivalence class membership. Specifically, for any $\mathbf{x} \in [\mathbf{x}]_i$ and for any $t \in \mathcal{T}_i$, we have $t \circ \mathbf{x} \in [\mathbf{x}]_i$. These transformations are viewed as defining how the data is generated, such that $\mathbf{X}^{\text{tr}} := T^{\text{tr}} \circ \mathbf{X}^\dagger$ and $\mathbf{X}^{\text{te}} := T^{\text{te}} \circ \mathbf{X}^\dagger$, with $P(T^{\text{tr}}) \neq P(T^{\text{te}})$. Each T is viewed as a concatenation of transformations associated with a given equivalence relation, $T = t_1 \circ \dots \circ t_r$ for arbitrary r . The differences between T^{tr} and T^{te} are found in a subset of the transformations, for instance we may have $T^{\text{tr}} = t_1 \circ t_{\text{tr}} \circ t_3$ and $T^{\text{te}} = t_1 \circ t_{\text{te}} \circ t_3$.

The task of OOD generalization is to make assumptions about which equivalence relation-induced transformations will remain invariant between test and training data and enforce the invariances in learning our prediction model. The training data needs sufficient variation among the global properties for the learning procedure to identify invariances.

Self-supervised learning using SCD: While the previous methods aimed to improve prediction for one task, the following methods explore how the principles of IFL can be exploited for self-supervised learning (SSL), where representations are learned to assist with a suite of downstream tasks. This strategy has paved the way to utilize a vast amount of unlabeled data in settings for which the availability of labeled data is often limited, e.g., for large-scale language modeling [53], medical image analysis [54], or molecular property prediction [55]. We can divide prevailing approaches primarily into whether they are reconstructive [56] or discriminative [48, 57, 58]. Within the discriminative regime, we aim to enforce proximity between representations of similar objects.

Contrastive learning is an approach within the discriminative regime that utilizes positive and negative sampling to enforce proximity between semantically equivalent samples and farness between semantically different samples in representation space. Many contrastive learning approaches rely on some form of data augmentation to generate positive and negative samples, as well as observing data from different environments but with similar labels (e.g., *multi-view contrastive learning* [59]). Kügelgen et al. [41] prove that data augmentations for SSL can indeed identify the invariant content features for representation learning in both the reconstructive and discriminative setting. From a causal perspective, they interpret contrastive objectives as comparing counterfactuals under soft style interventions.

Mitrovic et al. [39] argue that contrastive pre-training approaches implicitly assume the SCD causal structure illustrated in Fig. 3.2. Therefore, it teaches an encoder to disentangle \mathbf{C} and \mathbf{S} causally. The authors examine the setting where we have access to data augmentations to offer positive and negative samples for the contrastive loss and propose an objective named RELIC.

Conventional data augmentations can be interpreted as interventions on the style variable \mathbf{S} (see Causal Perspective 3.1.1). Mitrovic et al. [39] frame the task of self-supervised learning as proxy task prediction, where they propose a list \mathcal{Y} of proxy labels y_t . To explicitly enforce proxy target prediction invariance under data augmentations, they formalize a criterion and propose to enforce it during training by adding a KL-divergence term to the objective. Using Principle of Independent Mechanisms (Def. 2.3.4), they conclude that under SCD, performing interventions on \mathbf{S} does not change the conditional distribution $p(y_t | \mathbf{c})$, i.e. manipulating the value of \mathbf{S} does not influence this conditional distribution. Thus, $p(y_t | \mathbf{c})$ is invariant under changes in style \mathbf{S} , i.e. for all $\mathbf{s}^{(1)}, \mathbf{s}^{(2)} \in \mathcal{S}$,

$$p(y_t | \mathbf{c}, \text{do}(\mathbf{s}^{(1)})) = p(y_t | \mathbf{c}, \text{do}(\mathbf{s}^{(2)})) \quad (3.3)$$

As follow-up work, Tomasev et al. [60] propose RELICv2, which differs from RELIC in the selection of appropriate sets of positive and negative points and how to combine resulting views of data in the objective function.

3.1.1.4 Deconfounding post-training predictions

A variety of methods [61, 62, 63] propose to remove the influence of unobserved confounders \mathbf{U} from predictions through *counterfactual regularization*, after model training has ended. This involves estimating the confounding effect of \mathbf{U} on the prediction \tilde{Y} and then subtracting it, thus deconfounding the prediction. For a prediction over sample \mathbf{x} , \mathbf{x}' is generated such that it carries none of the causal information in \mathbf{x} . Then, the prediction is deconfounded using the difference

$$\tilde{Y}_{\text{causal}} = \tilde{Y}_{\mathbf{x}} - \tilde{Y}_{\mathbf{x}'} \quad (3.4)$$

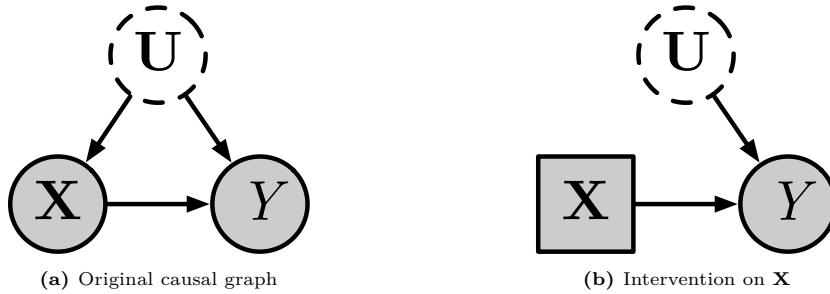


Figure 3.6: Counterfactual Regularization to remove unobserved confounding: These graphs explain how counterfactual regularization can eliminate the effect of unobserved confounding on model predictions, as explained in Sec. 3.1.1.4. An intervention on \mathbf{X} eliminates the influence of \mathbf{U} on it, and deconfounds the causal effect $\mathbf{X} \rightarrow \mathbf{Y}$.

Chen et al. [61] apply an intervention to \mathbf{X} by setting it to a zero vector (or random noise), identifying the resulting prediction of \hat{Y} , and implementing Eq. (3.4). The authors propose to use this for domain generalization in trajectory prediction. Rao et al. [63] propose improving attention mechanisms for visual categorization, and Niu et al. [62] propose removing language bias in visual question answering.

3.1.2 Multiple Environments

It is often possible to obtain datasets collected from multiple environments \mathcal{E} in the form $\left\{(\mathbf{x}_i^e, y_i^e)_{i=1}^N\right\}_{e \in \mathcal{E}}$. For example, data may be records obtained between hospitals under varying protocols or images of houses collected in different seasons. As a more concrete example, the WILDS benchmark [64] offers a curated collection of multiple environment data from real-world scenarios.

As many machine learning algorithms operate on i.i.d. data, practitioners may shuffle the data collected from multiple environments. Arjovsky et al. [32] suggest that information about the data generating process is lost when we shuffle, and spurious associations can occur in the data after doing so. Furthermore, by shuffling the data, we lose any distinction between associations across and within environments. For example, a spurious association exists between grass features and the cow label in Fig. 3.1, yet this association does not hold across environments if one environment is alpine pastures and another is the beach.

How can we use this information effectively? From a causal perspective, we can treat each environment as being generated from a set of interventions on style variables

in the SCD (Fig. 3.2). Each environment encodes a specific set of interventions, and a greater variety of environments can reveal more style variables.

Causal Perspective: 3.1.2: Data collected from multiple environments

For data in the form $\{\mathbf{X}^e, Y^e\}_{e \in \mathcal{E}}$ where \mathcal{E} is a set of environments, it is often reasonable to view each environment as being generated by a set of interventions on the *style* variables in the SCD (Fig. 3.2). Of course, this perspective fails when the underlying SCM differs between environments.

The methods in this section propose identifying the predictive features across all environments and learning an invariant mapping from the invariant features to the output variable Y . We interpret the invariant features as content features in the SCD or causal parents of Y .

3.1.2.1 Invariant Risk Minimization

Peters et al. [38] introduce *Invariant Causal Prediction* (ICP), an algorithm to find the *causal feature set*, the minimal set of features which are causal predictors of a target variable. They exploit the invariance property given by the Principle of Independent Mechanisms (Def. 2.3.4)

Invariant Risk Minimization (IRM) [32] is an extension of ICP that, instead of selecting variables, learns representations free of spurious associations. It posits the existence of a feature map Φ such that the optimal linear classifier $\hat{w} : \Phi(\mathbf{X}) \rightarrow \mathcal{Y}$ which maps these features to the output is the same for every environment $e \in \mathcal{E}$. The feature map and classifier compose to form a prediction function $f(\mathbf{x}) = w \circ \Phi(\mathbf{x})$. Arjovsky et al. [32] define the optimally-invariant classifier as one that minimizes the worst-case environment-specific empirical risk $\mathcal{R}^e := \mathbb{E}_{\mathbf{X}^e, Y^e} [\ell(f(\mathbf{X}^e), Y^e)]$ over all environments.

The authors argue that such a function will use only invariant features since non-invariant features will have varying associations with the label between different environments. We interpret this as an instance of the Style and Content Decomposition Fig. 3.2, where $\Phi(\mathbf{X})$ represents our content variable \mathbf{C} and \mathbf{S} variability is observed over a variety of environments.

Def.: 3.1.4: Invariant Risk Minimization [32]

To learn $\Phi(\mathbf{X})$ and an invariant classifier function $w_\beta : \Phi(\mathbf{X}) \rightarrow \mathcal{Y}$ parameterized by β , the IRM objective is the following constrained optimization problem:

$$\min_{\Phi, \hat{\beta}} \sum_{e \in \mathcal{E}} \mathcal{R}^e(\Phi, \hat{\beta}) \quad \text{s.t.} \quad \hat{\beta} \in \arg \min_{\beta} \mathcal{R}^e(\Phi, \beta) \quad \forall e \in \mathcal{E}. \quad (3.5)$$

In practice, this bilevel program is highly non-convex and difficult to solve. To find an approximate solution, the authors consider a Lagrangian form, whereby the sub-

optimality with respect to the constraint is expressed as the squared norm of the gradients of each of the inner optimization problems:

$$\min_{\Phi, \hat{\beta}} \sum_{e \in \mathcal{E}} \left[\mathcal{R}^e(\Phi, \hat{\beta}) + \lambda \left\| \nabla_{\hat{\beta}} \mathcal{R}^e(\Phi, \hat{\beta}) \right\|_2^2 \right] \quad (3.6)$$

Assuming the inner optimization problem is convex, achieving feasibility is equivalent to the penalty term being equal to 0. Thus, Eq. (3.5) and Eq. (3.6) are equivalent if we set $\lambda = \infty$.

Unfortunately, both Rosenfeld et al. [65] and Kamath et al. [66] show that IRM often performs no better than standard *empirical risk minimization* (ERM). IRM achieves its best results when the underlying causal relations are linear, and there is sufficient heterogeneity observed in the training environments, such that enough *degrees of freedom* are eliminated [65]. However, if either of these conditions is not met, then IRM can achieve worse results than ERM. Ahuja et al. [67] pose the IRM objective as finding the Nash equilibrium of an ensemble game among several environments.

Ahuja et al. [68] point out that while IRM-like approaches provably generalize OOD in linear regression tasks, they do not necessarily in linear classification tasks, which require much stronger restrictions in the form of support overlap assumptions on the distribution shifts. They establish that augmenting the invariance principle with *information bottleneck*, [69] constraints resolves some of these issues.

Similar to IRM, Krueger et al. [70] propose RISK EXTRAPOLATION (REx), a domain generalization method which also uses a weaker form of invariance than ICP. However, while IRM specifically aims for invariant prediction, REx seeks robustness to whichever forms of a distributional shift are present. The authors prove that variants of REx can recover the causal mechanisms of the targets while also providing some robustness to covariate shift.

Wang and Jordan [24] note that invariant features should be *non-spurious* but also *efficient*. To assess non-spuriousness and efficiency simultaneously, they require the representations to satisfy a probability of necessity and sufficiency (PNS) condition, which captures the following behavior.

A representation is non-spurious and efficient if the label responds to the feature both ways: non-spuriousness refers to the effectiveness of turning on the label by turning on the feature captured by the representation. Efficiency is the effectiveness of turning off the label by turning off this feature. In other words, if the feature is enabled, the label will be enabled; if it is disabled, the label will be disabled. PNS calculates the probability of this condition. The authors turn this condition into a constrained optimization objective, which they call CAUSAL-REP.

Jiang and Veitch [71] study anti-causal domain shifts, where $\mathbf{X} \leftarrow Y$, instead of $\mathbf{X} \rightarrow Y$. Similar to IRM, the authors propose a training objective for extracting an invariant predictor. Recall that IRM aims to build predictors that only rely on the causal parents of Y . If the observed covariates \mathbf{X} may be all caused by Y , the causal

parents of Y are the empty set. In this case, the naive causally invariant predictor is vacuous. Jiang and Veitch [71] address these scenarios.

Wang et al. [72] introduce *Invariant-feature Subspace Recovery* (ISR), addressing the fact that IRM and its extensions cannot generalize to unseen environments with less than $d_s + 1$ training environments, where d_s is the dimension of the spurious-feature subspace. They propose two algorithms, ISR-MEAN and ISR-Cov. By identifying the subspace spanned by invariant features, ISR-MEAN achieves provable domain generalization with $d_s + 1$ training environments. By using information of second-order moments, ISR-Cov reduces the required number of training environments to $\mathcal{O}(1)$.

3.1.2.2 Causal Matching

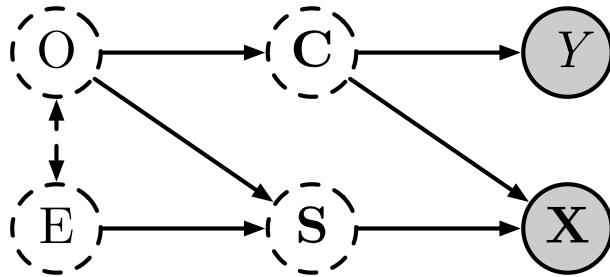


Figure 3.7: Modified Style and Content Decomposition [73]: When S and C are correlated, we would like to describe the nature of this association using a causal graph. Causal Matching [73] introduces the environment index variable E and object variable O in order to extend the causal graph (See Sec. 3.1.2.2). The algorithm aims to identify objects from content features C .

A progression beyond IRM is to analyze in greater detail how S and C interact across environments. While IRM assumes that S and C are spuriously associated in certain environments, Mahajan et al. [73] consider a data generating process (DGP) where S and C are confounded by an *object* variable O and where our environment E is a causal parent of S only. In contrast to IRM, the dependence between S and C is explicitly modelled with the causal graph Fig. 3.7.

Mahajan et al. [73]'s contribution is twofold. First, given prior knowledge of object locations in the data encoded in a matching set Ω , they propose a regularized objective that learns a representation $\Phi(\mathbf{X}) : \mathcal{X} \rightarrow \mathcal{C}$ using Ω , and a classifier $w_\beta : \Phi \rightarrow \mathcal{Y}$. Second, without prior knowledge of the object locations, the authors propose a matching algorithm to identify the object locations. Then, the estimated matching set $\tilde{\Omega}$ is deployed in the classifier objective.

Specifically, they propose that for any two data points with the same label but generated in different environments $e \neq e'$, $(\mathbf{x}^{(e)}, y), (\mathbf{x}^{(e')}, y)$, their underlying content features C should be similar. This idea motivates their matching algorithm MATCHDG, which alternates between i) building a matching set and ii) learning a representation $\Phi(\mathbf{X})$.

3.1.2.3 Style and Content as Latent Variables

As opposed to learning representations for style and content, there are a group of methods that model style and content as latent variables and learn $\{p^e(\mathbf{x}, y)\}_{e \in \mathcal{E}}$ across environment $e \in \mathcal{E}$. Through this perspective of probabilistic style and content variables \mathbf{C} and \mathbf{S} in the SCD, we change the optimization objective from a constrained one to an unconstrained one.

These methods sample \mathbf{C} and \mathbf{S} from a learned distribution $p^e(\mathbf{s}, \mathbf{c} \mid \mathbf{x})$, which varies depending on the environment $e \in \mathcal{E}$. To make a prediction, they sample from density $Y \sim p(y \mid \mathbf{c})$, which is invariant to the environmental changes by the Principle of Independent Mechanisms (Def. 2.3.4).

Sun et al. [74] propose LACIM, which exploits training data from multiple environments and assumes the data is generated according to a similar structure (Fig. 3.7) to the MATCHDG method seen in Sec. 3.1.2.2. However, LACIM solves an unconstrained optimization objective. A prior $p_\theta^e(\mathbf{c}, \mathbf{s})$ is learned for each environment e to capture the varying dependence between \mathbf{C} and \mathbf{S} .

The prior model $p_\theta(\mathbf{c}, \mathbf{s} \mid e)$, inference model $p_\theta(\mathbf{c}, \mathbf{s} \mid \mathbf{x}, e)$, generative model $p_\theta(\mathbf{x} \mid \mathbf{c}, \mathbf{s})$ and predictive model $p_\theta(y \mid \mathbf{c})$ are separately learned for each environment, but crucially, the parameters of $p_\theta(\mathbf{x} \mid \mathbf{c}, \mathbf{s})$ and $p_\theta(y \mid \mathbf{c})$ are shared across environments. As such, the causal variables \mathbf{C} are identified such that $p_\theta(y \mid \mathbf{c})$ achieves low loss.

In contrast, Liu et al. [75] exploit training data from a single environment only, but assume $\mathbf{C} \perp\!\!\!\perp \mathbf{S}$. \mathbf{C} is identified by exploiting the Principle of Independent Mechanisms (Def. 2.3.4). Their approach, called the *Causal Semantic Generative model* (CSG), operates similarly to LACIM.

Lu et al. [76] propose to perform the same task as LACIM but with a three stage training procedure instead of the end-to-end approach of LACIM and CSG. Their method is called iCARL. In phase 1, they learn \mathbf{Z} , the set of latent generative variables for \mathbf{X} . In phase 2, they use the PC algorithm [77] and conditional independence testing to isolate $\mathbf{C} \subset \mathbf{Z}$, the causal parents of Y . In phase 3, a classifier learns $p(y \mid \mathbf{c})$. iCARL outperforms ERM and IRM on the Colored Fashion MNIST [32] OOD classification benchmark.

3.1.2.4 Compositional Recognition

Compositional recognition is the problem of learning to recognize new combinations of known components. Note that this method is trained on data with both attribute \mathbf{A} and object O labels, as opposed to additional environment labels, which the previous methods in Sec. 3.1.2 exploited. Atzmon et al. [78] argue that deep discriminative models fail at compositional recognition because of two reasons: (i) distribution-shift and (ii) entanglement of representations. Hence, they propose constructing a representation based on a causal graph that models images as being caused by attributes \mathbf{A} and objects O , as illustrated in Fig. 3.8. Unlike objects or attributes by themselves, Atzmon et al. [78] argue that combinations of objects and attributes generate

the same distribution over images in train and test sets. Hence, they propose considering images of unseen combinations as generated by interventions on the attribute and object labels, casting zero-shot inference as the problem of finding which intervention caused a given image. They learn the distribution $p(\mathbf{x} | \mathbf{a}, o)$, which is assumed to be more stable across environments than either of $p(\mathbf{x} | \mathbf{a})$ or $p(\mathbf{x} | o)$. Then, one can learn a classifier using these robust representations.

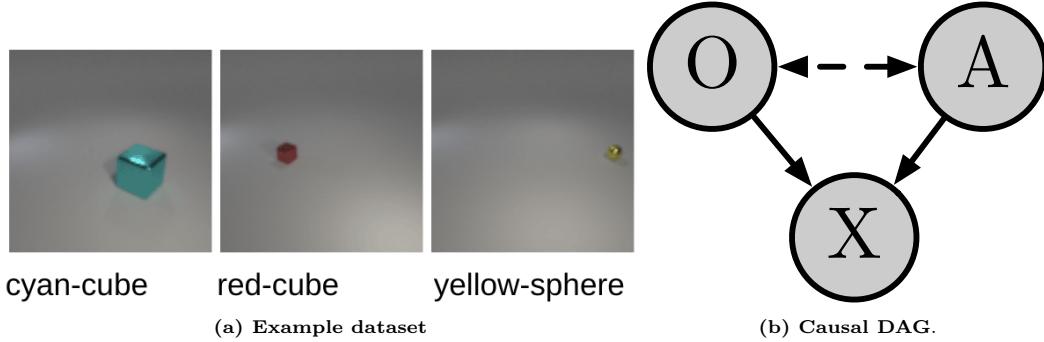


Figure 3.8: Composition of attributes leads to robust generation [78]: While objects and attributes features are unreliable features for prediction when used by themselves, combinations of such features are assumed to be reliable (Sec. 3.1.2.4). Hence, they learn $p(\mathbf{x} | \mathbf{a}, o)$ which will be invariant across interventional distributions. Examples of the dataset required are shown in Fig. 3.8a, collected from the AO-CLEVR dataset [79].

3.2 Invariant Mechanism Learning

Method	Key Idea	Ref.
Separate Networks	Networks can be combined in novel domains and account for independent interventions	Sec. 3.2.1
Domain Mappings	Mappings from target domain to source domain which each account for independent interventions	Sec. 3.2.2

Table 3.2: Overview of IML Methods.

In the previous section, we exploited the utility of invariant features to remove spurious associations from our models. Now we look at various methods with fundamentally different objectives compared to invariant feature learning.

Consider how a human may hear someone speaking quietly or loudly but can distinguish between what is said and how loudly it is. The content of the speech corresponds to a feature of the data. The speech volume corresponds to the mechanism that maps the information to the observer. This observation follows the Principle of Independent Mechanisms (Def. 2.3.4).

Invariant mechanism learning (IML) aims to identify a suite of data-generating mechanisms representing different interventional distributions. For independent latent confounders \mathbf{U} of \mathbf{X} and Y , we view each interventional distribution as being generated by interventions on a subset of the confounders (see Fig. 3.9). Then, for

the collection of learned mechanisms \mathcal{F} , we employ a subset of the mappings \mathcal{F} to predict Y from \mathbf{X} . We emphasize that IML does not learn feature maps of the confounders \mathbf{U} .

Def.: 3.2.1: Invariant Mechanism Learning (IML)

The task of IML is to identify a set of mappings \mathcal{F} which capture an interventional distribution generated by independent unobserved confounders $\mathbf{U} = \{U_1, \dots, U_D\}$:

$$\mathcal{F} = \{f_i : \mathcal{X} \rightarrow \mathcal{Y} \mid Y = f_i(\mathbf{X}) \Leftrightarrow Y \sim P(Y \mid \mathbf{X}, \text{do}(u_i))\}_{i=1}^D \quad (3.7)$$

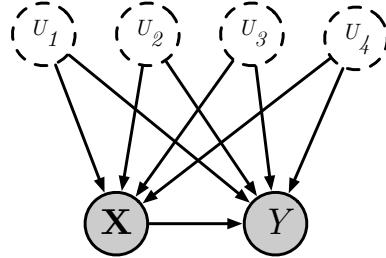


Figure 3.9: Graphical motivation for invariant mechanism learning: We consider $\mathbf{X} \rightarrow Y$ to be confounded by a set of independent confounders \mathbf{U} , such that an intervention on any U_i generates an interventional distribution.

3.2.1 Independent Invariant Mechanisms as Separate Networks

Parascandolo et al. [80] propose a method that aims to identify a suite of competing data transformations ('mechanisms') that are trained to specialize among themselves in recovering distinct underlying structure from a sample.

Goyal et al. [81] develop on the previous work by applying independent mechanisms to sequential data, specifically video and text. They call their proposed architecture *Recurrent Independent Mechanisms* (RIMs). At each time step, a soft attention layer selects the top k out of N competing mechanisms to exploit for processing of the input, and a second attention layer allows for 'sparse communication' between the mechanisms at each time step to aid contextual understanding.

The benefits observed over other sequential architectures such as the LSTM (Hochreiter and Schmidhuber [82]) and transformers (Vaswani et al. [83]) are (i) reduced degradation of information over long-dormant phases in the input, and (ii) improved trajectory prediction for multiple objects.

Madan et al. [84] use the RIMs proposed by Goyal et al. [81] for a meta-learning algorithm that quickly adapts the parameters of the modules and slowly adapts the attention mechanism parameters. Their fast and slow learning dynamic demonstrates improvements over LSTM and RIMs in meta-learning benchmarks.

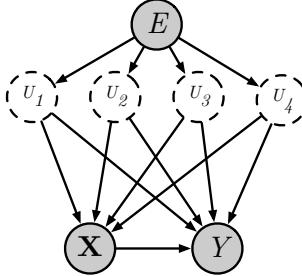


Figure 3.10: Invariant mappings between domains: Yue et al. [85] propose to learn a set of mappings between different environments generated by the above causal graph (Sec. 3.2.2), where \mathbf{U} is confounding the causal effect $\mathbf{X} \rightarrow \mathbf{Y}$. Hence, they propose to learn a set of mappings between source and test domain, that each account for an isolated intervention on independent parts of \mathbf{C} .

3.2.2 Invariant Mechanisms as Mappings Between Domains

Yue et al. [85] propose to view unsupervised domain adaptation for classification as a problem of inferring a set of disentangled causal mechanisms which generate mappings from the source domain to the test domain.

They model the data generating process (DGP) using the causal graph shown in Fig. 3.10. E represents an environment index and \mathbf{U} a set of unobserved domain aware confounding variables on \mathbf{X} and \mathbf{Y} . Yue et al. [85] leverage the following result from transportability theory [86]:

$$p(y \mid \text{do}(\mathbf{x}), e) = \sum_{\mathbf{u}} p(y \mid \mathbf{x}, \mathbf{u})p(\mathbf{u} \mid e). \quad (3.8)$$

This result shows that when we have access to the confounders \mathbf{U} , we can identify the effect of an intervention $\text{do}(\mathbf{X})$ in a separate domain E on Y . Given that \mathbf{U} are unobserved, the authors propose to learn a set of proxy variables $\hat{\mathbf{U}}$, and identify disentangled counterfactual mappings between domains that each correspond to an isolated intervention on U_i which keeps all other $U_{j \neq i}$ constant. They learn $\{(M_i, M_i^{-1})\}_{i=1}^k$ in an unsupervised fashion, where $M_i : X_{\text{tr}} \rightarrow X_{\text{te}}$ and $M_i^{-1} : X_{\text{te}} \rightarrow X_{\text{tr}}$.

For a given input in the test domain, they propose to map the test domain input to the source domain (i.e., to what the input would have been had it been generated in the source domain) and then predict based on this counterfactual input.

Similarly, Teshima et al. [87] propose to identify an invariant mechanism that generates \mathbf{X} from a set of independent components \mathbf{Z} which vary across domains. The method takes in training data from multiple domains and uses nonlinear ICA to *identify features* \mathbf{Z} and an invariant mapping $f : \mathcal{Z} \rightarrow \mathcal{X}$ across domains. The learned mapping \hat{f} identifies the values of \mathbf{Z} in the target domain and generates pseudo samples which mimic the target domain. A standard supervised learning algorithm can be trained on this generated data.

3.3 Open Problems

3.3.1 Lack of targeted benchmarks for invariance learning

Previous works in invariance learning often evaluate their suggested methods on novel toy experiments instead of a standardized test bed. As a result, practitioners cannot quickly determine the best approach for their problem. We summarize some existing benchmarks that may serve as inspiration for future works.

One common approach to evaluating causal supervised learning models is introducing spurious associations in the training data and evaluating performance in a test domain where the association has changed. For instance, Arjovsky et al. [32] introduce the COLOREDMNIST benchmark, which adds colors to the digits and changes the strength of association between color and digit label between test and training data.

Wang and Jordan [24] varies spurious associations between face attributes of interest and irrelevant attributes of the CELEBA dataset [88]. For example, the training data contains a spurious association between *black hair* and *necklaces*. Further, they propose to compare the generalization ability under changes to the specified dimension of the representations in their models.

3.3.2 Connections between Invariance Learning, Adversarial Robustness and Meta-learning

We note two active ML research areas akin to invariance learning: adversarial robustness and meta-learning. We put forward to investigate how causal assumptions may benefit these areas, as it is currently underexplored.

In adversarial robustness, we are interested in learning classifiers that are robust to adversarial perturbations. An adversarial perturbation is an additive random variable Δ that makes a model fail to classify $\tilde{\mathbf{X}} = \mathbf{X} + \Delta$, while the model correctly classifies an image \mathbf{X} . Typically, $\tilde{\mathbf{X}}$ and \mathbf{X} are indistinguishable to the human eye. Naturally, we can see the problem of AR from a causal perspective and interpret a robust model as one being invariant against perturbations, as we will later see in Sec. 8.2.1.2.

In meta-learning, we often seek to learn shared structure across tasks and task-specific parameters that can be quickly adapted to unseen tasks [89, 90, 91]. Here, we can interpret the shared structure as being invariant across tasks. Future work must examine if causal assumptions can be used to learn better task-invariant representations. For example, instead of conditioning on task variables, it could be interesting to explore interventions on them.

3.3.3 Exploiting additional supervision signals for IFL

The previously discussed methods in invariant feature learning (Sec. 3.1) require two distinct forms of extra supervision beyond labels Y : i) content-invariant transformations (Sec. 3.1.1), or ii) environment indices (Sec. 3.1.2). These additional supervisory signals serve as means to separate data among different interventions on the spurious features, which helps a model avoid any predictive dependence on them.

We suggest three directions for future research. First, we can combine these two forms of signals to exploit a combination of domain knowledge of the data generating process and data collected from multiple environments. One straightforward idea is to imagine data augmentations as themselves generating new environments and thus use existing algorithms to handle data from multiple environments, with data augmentation techniques multiplying the number of environments available.

Second, we can attempt to exploit high-dimensional environment information, as opposed to environment indices $e \in \mathcal{E}$. For instance, one could complement the dataset collected for the cow images (Fig. 3.1) with an aerial photo of the pictures' location landscape, which serves as an environment variable that causes the data and label. Kaddour et al. [91] show examples of additional task descriptors, such as one image per robotics environment.

Third, we can demand labels on the data beyond just a category, such as *object* and *attribute* labels in Sec. 3.1.2.4. Identifying and exploiting other forms of labels that inform invariant feature learning is an open problem.

4

Causal Generative Modeling

The goal of generative modeling is to produce samples that mimic the characteristics of our training data. The field of *controllable generation* refers to techniques that allow us to enforce a set of attributes that novel samples should satisfy. Non-causal controllable generation samples from a conditional distribution $p(\mathbf{x} \mid \mathbf{a})$ where \mathbf{A} is an attribute specification [92, 93]. Such observational distributions restrict us to sample properties as seen in the training data. However, we may want to generate samples with novel combinations of attributes unobserved in our training data or edit specific samples by specifying only the attributes we want to change, with downstream causal effects automatically incorporated.

Causal Generative Modeling (CGM) offers a causal perspective on controllable generation and sample editing by estimating an interventional or counterfactual distribution, respectively. For controllable generation, given a causal representation \mathbf{Z} of our data, we sample from $p(\mathbf{x} \mid \text{do}(\mathbf{a}))$, where $\mathbf{A} \subset \mathbf{Z}$ is the set of attributes we wish to enforce. If we wish to edit sample \mathbf{x} according to attributes \mathbf{a} , our target estimand is $p(\mathbf{x}_{\mathbf{A} \rightarrow \mathbf{a}} \mid \mathbf{x})$.

One unique use case of CGM is **scientific investigation** of complex causal mechanisms: Following Pawlowski et al. [94], imagine a medical imaging setup in which we are interested in how a person’s anatomy would change if particular traits were different. By incorporating expert domain knowledge in the form of a causal graph, counterfactual distributions enable us to model the appearance of brain MRI scans under a counterfactual scenario where the person’s biological sex is different. By analyzing corresponding counterfactual samples, we gain better insight into understanding the physical manifestation of biological sex on the brain. Fig. 4.1 shows some of the counterfactual samples produced by Pawlowski et al. [94].

Another unique CGM use case is **counterfactual data augmentation**. Recall the example of training an image classifier in Fig. 3.1. Given training data on cows commonly found in alpine pastures, the classifier fails to generalize to unfamiliar backgrounds, where cows are observed on the beach, for example. As a result, a classifier may learn a spurious association between grass backgrounds and the cow label in such an instance. However, we can eliminate the spurious association from the dataset by augmenting the data with images of cows in varying backgrounds. For example, Sauer and Geiger [95] propose to learn a generative model that disentangles background and foreground attributes. Then, the model can generate images where cow images are intervened upon to have new backgrounds, such as a beach.

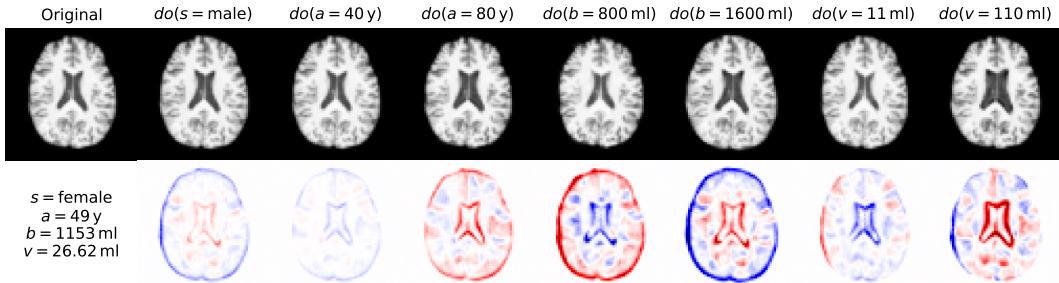


Figure 4.1: Brain image counterfactual samples [94]: One way to study the effects of varying demographics on the brain structure is to generate counterfactual samples, as done here by DEEP-SCM (Sec. 4.1.1). The intervention variables are a) age, s) sex, b) brain volume, and v) ventricle volume. **Top row:** Counterfactual samples with interventions specified for each column. **Bottom row:** Difference maps.

How can we learn such interventional and counterfactual distributions? Given that it is impossible to learn disentangled representations unsupervisedly from data without extensive domain knowledge [96, 97], we explore methods in this section that vary according to the supervision and domain knowledge they require. On the one hand, we discuss techniques demanding some domain knowledge of the underlying causal graph in *structural assignment learning* (Sec. 4.1). On the other hand, methods exist that relax this requirement, which we refer to as *causal disentanglement* (Sec. 4.2).

Notation

\mathbf{Z}	Generative variables
K	$\dim(\mathbf{Z})$
\mathbf{A}	Attributes / intervention variables
I	Index set of intervention variables s.t. $\mathbf{Z}_I = \mathbf{A}$
P	$\dim(I)$
ϵ	Independent exogenous causal parents
\mathbf{G}	Graph adjacency matrix

4.1 Structural Assignment Learning

We first highlight some methods that require a practitioner to specify the underlying causal graph \mathcal{G} of the data generating process (DGP) over observed generative variables \mathbf{Z} , but not the full SCM \mathcal{M} . Instead, for $z_j := f_j(\mathbf{pa}(z_j), \epsilon_j), j = 1,.., K$, these methods may learn the structural assignments $\{f_j\}_{j=1}^K$ from the data, where $K = |\mathcal{G}|$. Note that these methods rely on the absence of any *hidden confounders* (Sec. 2.5).

We train these models to learn the structural assignments of the underlying SCM and identify the exogenous noise values $\{\epsilon_i\}_{i=1}^K$ of a given sample. After training, we can generate counterfactuals $p(z_{\mathbf{A} \rightarrow \mathbf{a}} | z)$ for \mathbf{A} the intervened variables.

Method	Key Idea	Ref.
DEEPSMC [94]	Learn each structural assignment independently	Sec. 4.1.1
VACA [98]	Learn structural assignments jointly with a GNN	Sec. 4.1.2
DCEVAE [99]	Group structural assignments in the SCM with latent variable modeling	Sec. 4.1.3
DEAR [100]	Learn all structural assignments in the SCM with latent variable modeling	Sec. 4.1.4
DIFF-SCM [101]	Learn exogenous noise with diffusion model	Sec. 4.1.5
CGN [95]	Learn generating mechanisms of user-specified independent factors of variation	Sec. 4.1.6
ADA-GVAE [102]	Given intervention sample pairs, learn independent latent variables and the generative mechanisms	Sec. 4.1.7

Table 4.1: Overview of Structural Assignment Learning methods.

4.1.1 Independently learned structural assignments

Pawlowski et al. [94] introduce DEEPSMC, a principled approach for estimating interventional and counterfactual distributions given the underlying causal DAG of the data-generating process. They propose to instantiate its SCM by learning the functional assignments f_i for each variable $z_i := f_i(\epsilon_i; \mathbf{pa}(z_i))$ from its parents $\mathbf{pa}(z_i)$ and mutually independent noise terms ϵ_i , using normalizing flows [103] and variational inference [104].

To estimate a given sample’s counterfactuals, we need access to its exogenous noise values. Obtaining access to the noise values observed in a given sample is called the abduction step, as we covered in Def. 2.3.2. To perform this step, Pawlowski et al. [94] propose to learn a mapping for each observed z_i to its respective noise term, $\epsilon_i = f_i^{-1}(z_i; \mathbf{pa}(z_i))$ through normalizing flows in the low dimensional setting, and through variational inference in the high dimensional setting. After learning this mapping, one can perform a counterfactual query by modifying the variables of choice in the causal graph and evaluating a prediction from the SCM with fixed noise values.

Fig. 4.1 highlights some examples of the counterfactual samples produced. We note that this method has been applied with causal graphs of order 5. Scaling up this approach to larger causal graphs is an open research problem.

4.1.2 Structural assignments with a GNN

Sanchez-Martin et al. [98] develop on the ideas of DEEPSMC (Sec. 4.1.1) and investigate how Graph Neural Networks (GNNs) can be used to solve the same task. Unlike DeepSCM, their approach simultaneously learns all (potentially non-

linear) structural assignments during training. They refer to their approach as VACA(VAriational Causal graph Autoencoder).

They express the causal graph as an adjacency matrix \mathbf{G} embedded in each layer of the GNN. Both the encoder $p_{\theta}(\mathbf{w} | \mathbf{z}, \mathbf{G})$ and decoder $p_{\theta}(\mathbf{z} | \mathbf{w}, \mathbf{G})$ are GNNs that take \mathbf{G} as input. They consider \mathbf{Z} as representing the set of endogenous causal variables and identify each latent variable W_i as capturing all the information about Z_i that cannot be explained by $\text{pa}(Z_i)$. Note that \mathbf{W} does not necessarily correspond to the exogenous variables ϵ , and $p(\mathbf{w}) \neq p(\epsilon)$.

One limitation of this model is its expressivity w.r.t. the size of the underlying causal graph. VACA captures causal interventions if and only if the number of hidden layers in its decoder is greater than or equal to $\gamma - 1$, with γ being the length of the longest path between any two endogenous nodes in the true causal graph. As GNNs' performance tends to decrease sharply with depth, VACA will struggle to perform well over large causal graphs.

4.1.3 Group structural assignments

Kim et al. [99] develop a generative model, DCEVAE, that produces counterfactual samples by segmenting the causal DAG and learning three different segments (see Fig. 4.2). These segments must depend upon the interventions we want to perform. They propose to cluster the causal graph based on which features undergo interventions, which may carry two benefits: i) it can reduce the problem of error propagation along Markov factorizations that VACA (Sec. 4.1.2) addressed, and ii) it allows for counterfactual sampling over higher-order causal graphs than the previous methods.

This method outputs an image and takes observed concept labels \mathbf{Z} , such as *moustache* or *gender* in CelebA data [88], as inputs. Given interventions on $\mathbf{A} \subset \mathbf{Z}$, the new interventional distribution $p(\mathbf{z}^I | \text{do}(\mathbf{a}))$ is inferred, and \mathbf{x}^I is generated from \mathbf{z}^I . The practitioner specifies which concept labels they want to intervene on, \mathbf{A} , and the causal descendants of the intervened variables \mathbf{Z}_d . The remaining attributes are $\mathbf{Z}_r = \mathbf{Z} \setminus \{\mathbf{A} \cup \mathbf{Z}_d\}$. The exogenous variables are split too, so ϵ_d, ϵ_r are the exogenous causal parents of $\mathbf{Z}_d, \mathbf{Z}_r$ respectively. Given a sample \mathbf{x} , the model produces a counterfactual sample from interventions on \mathbf{A} by identifying the generative variables \mathbf{z} and the sample exogenous factors ϵ_d and ϵ_r , and generating counterfactual \mathbf{x}_a .

The learning task is to estimate the encoder that maps the observed causal variables \mathbf{z} to the exogenous variables, $p_{\theta}(\epsilon_r, \epsilon_d | \mathbf{a}, \mathbf{z}_d, \mathbf{z}_r, \mathbf{x})$ and the decoder $p_{\theta}(\mathbf{z}_d, \mathbf{z}_r, \mathbf{x}, \epsilon_d, \epsilon_r | \mathbf{a})$ which generates the counterfactual sample. Due to their disentanglement of $\mathbf{Z} = \mathbf{A} \cup \mathbf{Z}_d \cup \mathbf{Z}_r$ and $\epsilon = \epsilon_d \cup \epsilon_r$, the encoders and decoders factorize neatly.

Liu et al. [88] applied their method to the CelebA dataset and showed that it improves upon causal agnostic models in counterfactual sampling with downstream effects of interventions accounted for. However, their shown images are often blurry.

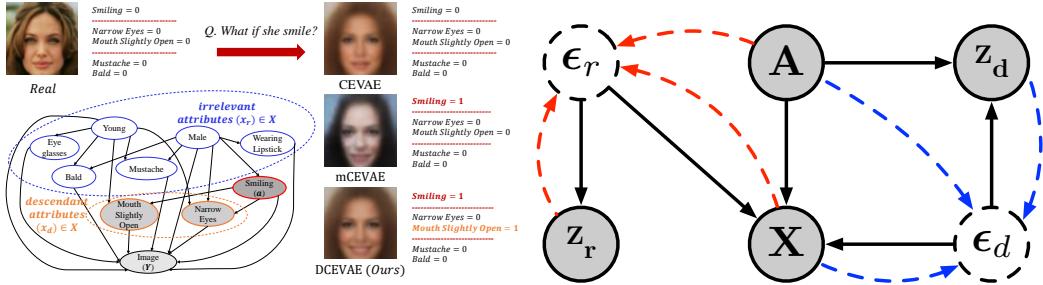


Figure 4.2: Disentangled Causal Effect Variational Autoencoder (DCEVAE) [99], see Sec. 4.1.3.

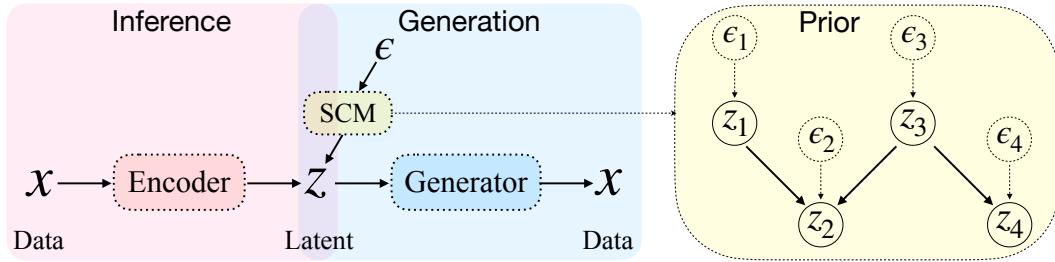


Figure 4.3: DEAR [100]: the prior $p_\beta(z)$ encodes the SCM among latents \mathbf{Z} .

4.1.4 GAN-based latent variable Causal

Similarly to DCEVAE, Shen et al. [100] propose a latent variable model for counterfactual sampling called DEAR (Disentangled gGenerative cAusal Representation). In contrast to DCEVAE, they propose to learn all of the structural assignments among the latent variables. They employ a GAN [105], which removes the constraint of a Gaussian prior in VAE-based methods. The practitioner must have access to the causal ordering of the variables of interest, and labels for the causal variables in each sample, as found in the CelebA dataset [88]. Given access to the causal ordering, a supergraph (i.e., a graph containing all subgraphs, akin to a superset) of the adjacency matrix \mathbf{G} is initialized in the prior distribution $p_\beta(z)$, from which they identify \mathbf{G} during training.

The authors propose to learn the generative model $p_\theta(x | z)p_\beta(z)$ and the encoder $q_\phi(z | x)$ with parameters θ, β and ϕ , and a discriminator D_ϕ through solving a bi-level optimization objective. For given exogenous factors ϵ , z is generated from the prior $p_\beta(z)$ by a function $F_\beta(\epsilon)$ of the form

$$z = F_\beta(\epsilon) := f\left((\mathbf{I} - \mathbf{G}^\top)^{-1}h(\epsilon)\right), \quad (4.1)$$

where $\beta = \{f, h, \mathbf{G}\}$. They specify f to be invertible, enabling a modification of Eq. (4.1) permitting simulation of interventions. Also, they enforce the encoder

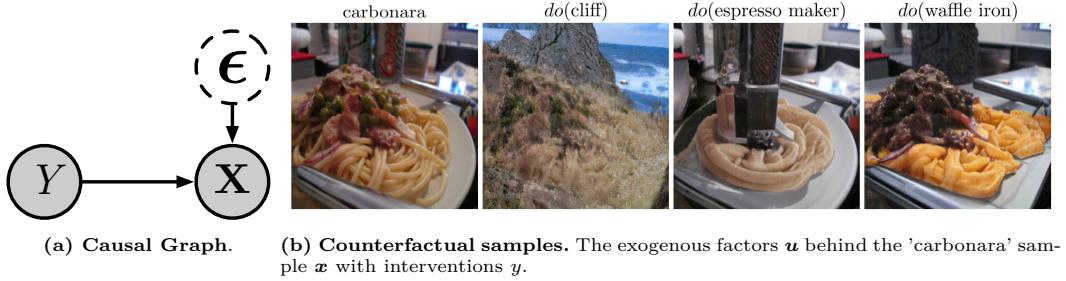


Figure 4.4: Diffusion-SCMs [101]: It let us generate counterfactuals \mathbf{x}_y by intervening on Y and inferring abducted noise \mathbf{u} . The diffusion process encodes for the exogenous factors ϵ , and the classifier guidance is used to simulate interventions on Y . See Sec. 4.1.5.

$q(\mathbf{z} \mid \mathbf{x})$ to satisfy alignment between sample labels, which indicate causal variables of interest, and the latent variables encoded.

4.1.5 Diffusion-based counterfactual estimation

Diffusion models have recently emerged as a highly effective generative modeling framework for images [106], and the classifier guidance framework proposed by Dhariwal and Nichol [107] permits conditional image generation. Sanchez and Tsafaris [101] propose DIFF-SCM, a framework for counterfactual sampling that employs Denoising Diffusion Models [93, 106, 107]. They use a bi-variable causal model in their implementation, where $Y \rightarrow \mathbf{X}$ as in Fig. 4.4a.

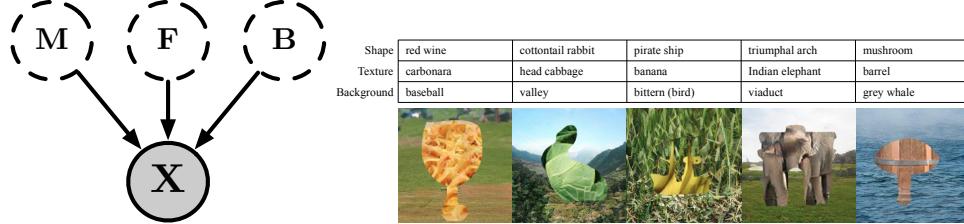
Following the score based framework of Song et al. [93] and the classifier guidance framework of Dhariwal and Nichol [107], DIFF-SCM uses an *anti-causal predictor* as a means of classifier guidance. The anti-causal predictor steers the generation towards the counterfactual distribution, and the extent to which the guidance is balanced with the score matching objective is managed by a parameter s . Sanchez and Tsafaris [101] view the forward diffusion as the encoding for the exogenous variables ϵ (abduction step) and view classifier guidance as a means to simulate an intervention in the generation process.

In their work, they simulate interventions on the class label Y , thus producing counterfactual samples which share some qualitatively high-level similarities yet differ in classification label (see Fig. 4.4b).

4.1.6 Disentangled mechanisms among variables of interest

Sauer and Geiger [95] propose *Counterfactual Generative Networks* (CGN), a generative network motivated by invariant causal mechanisms (see sec. 3.2). The authors employ three parallel BIGGAN [108] mechanisms that identify and modify different factors of variation (FoV): object shape, object texture, and background. The mechanisms are trained to specialize in one FoV and are composed through a composition module C, illustrated in Fig. 4.5.

CGN relies upon the independence of the given FoVs, thus assuming a simple causal graph where each FoV causes \mathbf{X} , as shown in Fig. 4.5a. CGN takes noise vector ϵ and attributes \mathbf{A} as input and outputs either an intervention sample (where the noise values ϵ vary between the mechanisms for given attributes a) or a counterfactual sample (where the noise values and attributes are shared for each mechanism).



(a) **Causal DAG** with mask M , (b) **ImageNet Counterfactuals**. CGN learns the disentangled foreground F , and background B . FoVs and enables the generation of permutations thereof.

Figure 4.5: Counterfactual Generative Networks [95]: The factors of variation (FoVs) are object shape, object background and object texture (see Sec. 4.1.6).

4.1.7 Disentangled mechanisms using intervention sample pairs

Similar to the goal of CGN, Locatello et al. [102] propose ADA-GVAE, a generative model that identifies independent representations. This method demands two data conditions: (i) the data is a collection of paired samples before and after an unspecified intervention, and (ii) we know the number of intervened latent generative variables. Specifically, we have a collection of tuples $\{(\mathbf{x}_i, \mathbf{x}_i^I, P)\}_{i=1}^n$ where P is the number of intervened generative variables, and \mathbf{x}^I denotes a sample after an intervention. In contrast to CGN, they do not require specification of the latent variables of interest, such as *object shape*, but instead sample pairs and latent dimension specification.

For latent generative variables \mathbf{Z} , Locatello et al. [102] assume that they are independent, and admit a prior factorization $p(\mathbf{z}) = \prod_{i=1}^K p(z_i)$ (thus a similar causal graph to that assumed in Fig. 4.5a). For $I \subset [K]$ the index set of intervened generative variables, $|I| = P$, and $\mathbf{Z}_I := \mathbf{A}$ the intervened generative variables, we have

$$p(\mathbf{x}, \mathbf{x}^I, \mathbf{z}, \mathbf{z}_I, I) = p(\mathbf{x} | \mathbf{z})p(\mathbf{x}^I | \mathbf{z}_I, \mathbf{z}_I)p(\mathbf{z})p(\mathbf{z}_I)p(I) \quad (4.2)$$

To identify the intervened variables \mathbf{Z}_I , the unintervened variables $\mathbf{Z}_{\bar{I}}$ are chosen to be the $K - P$ variables that induce the smallest values of $D_{\text{KL}}(p_{\theta}(z_i | \mathbf{x}) \| p_{\theta}(z_i | \mathbf{x}^I))$. After identification, the posterior distributions for unintervened variables are set to equality between the sample pairs, while the posteriors for the intervened variables are left untouched. A maximum likelihood objective is then optimized to learn the generative model.

4.2 Causal Disentanglement

Next, we study methods that do not require the specification of any underlying causal graph. Instead, they identify both the underlying graph and the structural

Method	Data Requirement	Ref.
CAUSALVAE [109]	Causal variable labels	Sec. 4.2.1
ILCM [110]	Intervention sample pairs	Sec. 4.2.2
CITRIS [111]	Sequential data with intervention target labels	Sec. 4.2.3

Table 4.2: Overview of Causal Disentanglement methods, differing by their data requirements.

assignments between the variables, thus learning a set of causally disentangled representations [11, 14].

Def.: 4.2.1: Causal Disentanglement [11]

We say a set of representations \mathbf{Z} , s.t. $\mathbf{X} = g(\mathbf{Z})$ for some mapping g , are *causally disentangled* if they permit the factorization

$$p(z_1, \dots, z_K) = \prod_{i=1}^K p(z_i \mid \mathbf{pa}(z_i)). \quad (4.3)$$

where $\mathbf{pa}(Z_i) \subset \{Z_j\}_{j \neq i} \cup \epsilon_i$ and ϵ_i is the exogenous causal factor of Z_i .

These methods do not require access to the complete causal graph \mathcal{G} , instead requiring practitioner knowledge about the generative variables \mathbf{Z} of interest. They learn how to reproduce observational distribution $p(\mathbf{x})$, interventional distribution $p(\mathbf{x} \mid \text{do}(\mathbf{a}))$, and counterfactual distributions $p(\mathbf{x}_{\mathbf{A} \rightarrow \mathbf{a}} \mid \mathbf{x})$.

4.2.1 Exploit causal variable labels

Yang et al. [109] propose CAUSALVAE, a method that learns a causal model over latent variables from data and generates counterfactual samples. The dataset must contain labels on the latent causal variables of interest for each sample, following the identifiability framework outlined in [97]. These labels represent the generative variables \mathbf{Z} .

Like VACA (Sec. 4.1.2), CAUSALVAE expresses the SCM as an adjacency matrix \mathbf{G} . However unlike VACA, CAUSALVAE learns \mathbf{G} as well as linear structural assignments as in Eq. (4.4),

$$\mathbf{z} = \mathbf{G}^\top \mathbf{z} + \boldsymbol{\epsilon}, \quad (4.4)$$

while VACA permits nonlinear structural assignments too.

We generate a counterfactual sample \mathbf{x}_a according to attributes \mathbf{a} as follows: for a given sample \mathbf{x} , the encoder identifies the exogenous noise $\boldsymbol{\epsilon}$. The exogenous noise determines the causal latent variables \mathbf{z} via the SCM, Eq. (4.4), to obtain $\mathbf{z} = (\mathbf{I} - \mathbf{G}^\top)^{-1} \boldsymbol{\epsilon}$. Now that we have an encoding for \mathbf{z} , we simulate an intervention on attributes \mathbf{a} . To obtain counterfactual \mathbf{z}_a , \mathbf{z} is input into Eq. (4.4) except with



Figure 4.6: Simulated interventions by ILCM [110] compared to ground truth: ILCM predicts the effects of a set of interventions I on a set of causal variables $\{c_i\}_{i \in I}$, using the *Causal3DIdent* dataset [41]. Given a sample \mathbf{X} , the model generates a simulation of the sample under intervention, as seen in the bottom row of the figure. See Sec. 4.2.2 for explanation.

modifications on \mathbf{G} that reflect the interventions \mathbf{a} . Then counterfactual \mathbf{z}_a is input into the decoder to generate a counterfactual sample \mathbf{x}_a .

This method was applied to the CelebA dataset [88] where they constructed an SCM over four latent variables.

4.2.2 Exploit intervention sample pairs

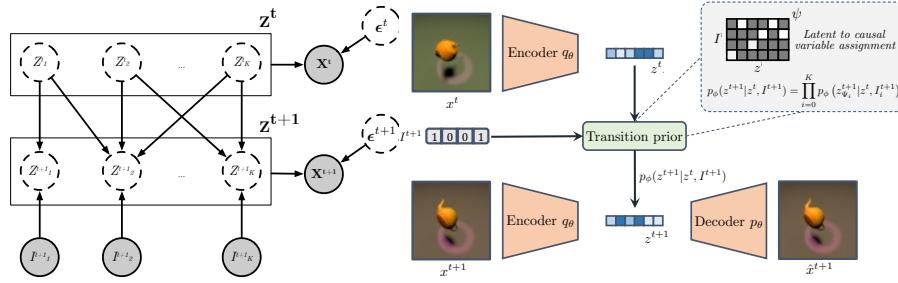
Brehmer et al. [110] improve upon ADA-GVAE by i) performing causal graph discovery and ii) removing the requirement that a practitioner has access to the number of intervened variables between sample pairs, proposing to learn a generative model over a collection of tuples $\{(\mathbf{x}_i, \mathbf{x}^I)\}_{i=1}^n$. In this setting, they propose two types of models: *Explicit or Implicit Latent Causal Models*. For the former, ELCM, they introduce a prior over latent variables $p(\mathbf{z})$ that encodes the structure of a specified causal graph. For the latter, which they call ILCM, they propose to learn a *noise encoder* $p_\theta(\epsilon, \epsilon^I | \mathbf{x}, \mathbf{x}^I)$ that maps the data $(\mathbf{X}, \mathbf{X}^I)$ to the exogenous noise values (ϵ, ϵ^I) in the latent causal model, before and after interventions. Note that ILCM relies upon the interventions I only changing one noise value ϵ_i , between ϵ and ϵ^I . A prior $p(\epsilon, \epsilon^I, I)$ is specified, which assumes independence of exogenous causal parents, $\epsilon_i \perp\!\!\!\perp \epsilon_j$ for all $i, j \in [K]$. No practitioner specifications of the underlying causal graph, the causal variables of interest, or the structural mechanisms are required for ILCM. What is required is the specification of the dimension of \mathbf{Z} , K , which determines the size of the causal graph.

ILCM implicitly learns the generative variables \mathbf{Z} since the exogenous noise values determine the variables in an SCM. To obtain the generative variables in explicit form, they learn *solution functions* $s(., .)$ s.t. $z_i = s(\epsilon_i, \epsilon_{-i})$. A causal graph is learned by applying intervention-based causal discovery algorithms over the generative variables \mathbf{Z} .

ILCM learns by maximizing an ELBO approximation to the maximum likelihood, and a noise decoder $p(\mathbf{x} | \epsilon)$ generates a sample from the inferred exogenous variables. Brehmer et al. [110] reproduce interventional distributions $\mathbf{X}^I \sim p(\mathbf{x} | \text{do}(\mathbf{a}))$ as demonstrated in Fig. 4.6.

4.2.3 Exploit sequential data with intervention labels

While ILCM exploits independence of exogenous noise values ϵ_i for causal graph identification, Lippe et al. [111] propose to exploit independence of the time indexed causal variables, conditioned on the previous time step and assuming no instantaneous effects. Their method, called *Causal Identifiability from Temporal Intervened Sequences* CITRIS, exploits data from a sequence $\{(\mathbf{x}^t, I^t)\}_{t=0}^T$, where each intervention target $I^t \in \{0, 1\}^K$ labels the causal variable intervened upon to generate sample \mathbf{x}^t . Given such data, CITRIS learns causal latent variables \mathbf{Z} and a transition prior $p_\beta(z^{t+1} | z^t, I^{t+1})$, which encodes the governing SCM and is trained using an ELBO objective.



(a) **Causal DAG** with latent variables Z_i , exogenous factors ϵ^t , latents \mathbf{Z}^t and \mathbf{Z}^{t+1} .
(b) **Architecture**: the transition prior encodes the SCM between intervention targets I_i^t and observations \mathbf{x}^t .

Figure 4.7: CITRIS [111]: The transition prior $p(z^{t+1} | z^t, I^{t+1})$ captures the temporal causal relations from \mathbf{Z}^t to \mathbf{Z}^{t+1} .

CITRIS also accommodates multi-dimensional causal variables, i.e., it groups latent variables of the same causal factor, such that the transition prior factorizes based on these multi-dimensional variables. For example, this property helps to learn one position variable to capture 3D coordinates instead of three independent axis variables because the ground truth axes are not always well-defined. To learn multi-dimensional causal variables, the authors encourage practitioners to specify the dimension of the latent space \mathcal{Z} to be larger than the number K of causal variables.

CITRIS learns an assignment function $\psi : [\dim(\mathbf{Z})] \rightarrow [K]$ which factorizes the latent space into causal variables and exogenous noise, such that $\mathbf{Z} = \{Z_{\psi_1}, \dots, Z_{\psi_K}, \epsilon\}$. Hence, the transition prior factorizes as

$$p_\beta(z^{t+1} | z^t, I^{t+1}) = p_\beta(\epsilon^{t+1} | z^t, I_i^{t+1}) \cdot \prod_{i=1}^K p_\beta(z_{\psi_i}^{t+1} | z^t, I_i^{t+1}). \quad (4.5)$$

We obtain the parents of any latent variable Z_{ψ_i} , and ultimately the causal graph, by identifying which of the latents $\{Z_{\psi_1}^t, \dots, Z_{\psi_K}^t\}$ influences $Z_{\psi_i}^{t+1}$ in the transition prior.

Further, CITRIS learns an invertible mapping $g_\theta : \mathcal{X} \rightarrow \mathcal{Z}$ between the observational data \mathbf{X} and the latents \mathbf{Z} . CITRIS is also capable of using representations from a pre-trained encoder g_θ , by learning a normalizing flow model which maps

the representations to disentangled representations \mathbf{Z} which the transition prior can admit.

Since CITRIS assumes no instantaneous effects in the data, it fails to discover the underlying causal graph when instantaneous effects are present. For instance, [112] highlight the example where both (i) flicking a switch, and (ii) its effect on a light bulb, can occur within one time step of a sequence. In this case, both the variable *light bulb* and *switch* are not conditionally independent given the previous time step information.

Lippe et al. [112] propose to address this shortcoming with iCITRIS (*instantaneous* CITRIS), employing a similar framework to CITRIS, and exploiting the same type of data. In contrast to CITRIS, which accommodated soft interventions, the authors assume that the interventions are perfect, rendering an intervened variable independent of its parents. The transition prior is augmented with instantaneous causal graph structure information, and the authors propose to learn the graph and generative model parameters by solving a bi-level optimization objective.

4.3 Open Problems

4.3.1 Different levels of abstraction

Causal representation learning (CRL, Def. 2.4.1) aims to find an SCM among latent generative factors \mathbf{Z} of our data \mathbf{X} , where the generative factors are assumed to be *causal variables*. This problem is ill-posed, however, unless we specify the correct level of abstraction in the SCM. For instance, we may choose to construct a causal model between each pixel in an image ($\dim(\mathbf{Z}) = \dim(\mathbf{X})$) or between three types of possible *objects* that may be observed in an image ($3 = \dim(\mathbf{Z}) \ll \dim(\mathbf{X})$). Both models may deliver equal performance (given enough data and compute), yet the latter is more efficient and interpretable for a human.

It is not obvious how practitioners should determine a reasonable level of abstraction for causal representation learning, much less do so automatically [113, 114, 115, 116]. All of the approaches in Sec. 4.2 required specifications that determined the abstraction of the SCM, for example. CAUSALVAE (Sec. 4.2.1) required variable labels in the data, while ILCM (Sec. 4.2.2) specified $\dim(\mathbf{Z})$ and required samples pairs before and after interventions. Can we identify other sensible specifications or even relax them? For instance, ILCM relaxed the specifications of ADA-GVAE (Sec. 4.1.7).

Important theoretical results determine the limits and the opportunities of how much we can automate causal representation learning. Locatello et al. [96] demonstrate that unsupervised disentangled representation learning is ill-posed without the proper model inductive biases and some extra supervisory information. Khemakhem et al. [97] prove that it is possible to identify the underlying joint distribution between observed and latent variables, as long as a practitioner specifies a factorized prior of the latent variables and provides additional supervisory signals. Gresele

et al. [117] derive identifiability results for identifying independent generative variables when considered from multiple environments, even under nonlinear mixing in the generative process. Finally, Cohen [118] identifies that disentangled representations typically permit impossible combinations of causal variables and proposes a framework for interventions, incorporating relations between causal variables that govern what is physically possible.

Thus, what kinds of supervisory signals are optimal for CRL, what trade-offs exist, and which signals are reasonable to require, are all open problems.

4.3.2 Scaling structural assignment learning to larger graphs

The structural assignment learning methods reviewed in Sec. 4.1 generate compelling counterfactual samples. However, they either operate on relatively low-order causal graphs or amortize the structural assignments and focus on a set of plausible interventions.

In applied sciences, we often want to model higher-order causal graphs and flexibly generate counterfactual samples from them to capture real-life dynamics, such as in cell biology [119, 120]. Yet, the approaches reviewed in Sec. 4.1 do not demonstrate results over causal graphs of order larger than 5, except for DCEVAE (Sec. 4.1.3) which amortizes the graph assignments. VACA (Sec. 4.1.2) relies on GNNs, which struggle to scale to large-scale graphs due to over-smoothing [121]. Hence, how to tackle the scalability of the underlying causal graphs in structural assignment learning remains an open problem.

4.3.3 Understanding counterfactual data augmentation

A growing trend has been counterfactual data augmentation (CFDA), as we saw in Sec. 3.1.1.1 and will see later in Secs. 7.8, 8.1.3.1, 8.2.3 and 8.3.2. Here, we focus on CFDA using causal generative models instead of hand-crafted transformations. While many augmentations can be hand-crafted, such as color transformations and rotations over images [122], we can also use generative models to create data augmentations. For instance, GENINT (Sec. 3.1.1.1) use a GAN [108] to create data augmentations.

However, while GENINT uses a causal agnostic generative model, it is possible to use causal generative modeling to generate CFDAs. We explored the causal perspective on data augmentation in Sec. 3.1.1, where we viewed an augmentation as an intervention on a spurious data feature. Sauer and Geiger [95] Sec. 4.1.6 compare the efficacy of training a classifier on datasets augmented with 1) the counterfactual samples and 2) non-counterfactual samples generated by a GAN and find that counterfactual samples improved the classifier’s performance more.

Exploring when and by how much CFDA using causal generative models improves generalization remains an open problem. For hand-crafted data augmentations, some understanding of their efficacy already exists. For example, Hernández-García and

König [123] conclude that it is more effective than explicit regularization, such as weight decay or dropout. Chen et al. [124] show that conventional data augmentation flattens the loss surface and that it can achieve similar performance gains as flat-minima optimizer [125]. Cubuk et al. [126, 127] offer automated augmentation generators. For CFDA, works and results like those mentioned above are missing.

Further, we may use CFDA as an evaluation proxy for comparing causal generative methods by their capability of improving an external predictor's generalization performance. For a fixed prediction model and training dataset, methods in this chapter can be compared on the prediction algorithm test loss when one trains the model using data augmentations generated by the counterfactual sampling method. Such a benchmarking method is under-explored.

5

Causal Explanations

The goal of *AI explainability* (or *explainable AI*) is to output *explanations* that make the decisions that a model constructs understandable to humans and provide an answer to *why* the output was predicted [128]. From the point of view of *model interpretability* [5], which refers to the degree to which an observer can understand the cause of a model prediction, an explanation falls under the category of *post-hoc* interpretability: the model (prediction) is analyzed after training, while *intrinsic* interpretability may refer to models that are restricted in their complexity and therefore sacrifice predictive performance [129]. Many explanation methods have been proposed, and we refer the reader to [130, 131, 132, 133] for excellent surveys.

In this section, we focus on two classes of explanation techniques: *feature attribution* and *contrastive explanations*. We will explain how causality enters the picture in the corresponding subsections. All discussed methods provide *local* explanations of individual predictions based on a single input, and are *model-agnostic*, i.e., they can be used across different model types. In contrast, there also exist methods providing *global* explanations, which describe the model’s average reliance on each feature across the whole dataset, and model-specific methods, such as saliency maps for neural networks trained on images [134, 135]. These latter two are not covered in this survey.

To motivate the usefulness of explanations, let us consider the case of someone who applied for a loan and was rejected by a financial institution’s loan distribution model. In most cases, the individual wants to understand the model’s reasoning in a bid to strengthen their next application. Fixing this scenario as a running example, we will introduce multiple techniques to produce explanations in the upcoming subsections.

Notation

x^F	Factual individual with undesired outcome
x^{CE}	Counterfactual explanation individual with desired outcome
x^{SCF}	Structural counterfactual individual with desired outcome (w.r.t. causal graph)

Class	Method	Key Idea	Ref.
Feature Attributions	CXPLAIN	Quantify causal influences of input features to model accuracy	Sec. 5.1.1
	GCE	Learn latent factors which cause change in the output	Sec. 5.1.2
	ASVs	Relaxes asymmetry axiom of Shapley values w.r.t. given causal knowledge	Sec. 5.1.3
	CSVs	Intervenes upon features instead of conditioning on them	Sec. 5.1.4
Contrastive Explanations	Counterfactual Explanations	\mathbf{x}^{CE} satisfies minimal distance to \mathbf{x}^F	Sec. 5.2.1
	Algorithmic Recourse via Minimal Interventions	\mathbf{x}^{SCF} satisfies minimal cost set of interventions on \mathbf{x}^F	Sec. 5.2.2

Table 5.1: Method Overview of Causal Explanations.

5.1 Feature Attribution Explanations

Attribution-based explanations assign a ranking to features, representing the marginal contribution of each feature to the output of the model. In our running example, such explanations may show that for the individual whose loan application was denied, the most important feature was their overall income, while their credit card debt did not contribute to the model’s prediction. While this explanation does not provide an instruction of what the individual should do to get the loan applied, it informs the applicant about which features of their application may have caused the rejection to which degree.

Regarding our running example, we illustrate why a loan applicant would prefer *causal* feature attribution explanations over associational ones, as follows. Imagine that the financial institution primarily cares about the applicant’s yearly income: the higher the income, the higher the chances of providing a loan. Other applicant features may be spuriously correlated with the income: e.g., their education level, job, industry, whether they are self-employed, etc.

A traditional, non-causal attribution method may assign uniformly high scores across all these correlated features. However, if the applicant knew that only their income matters, they likely would focus on improving it more directly: instead of even considering a career change, working directly on getting a promotion at their current job might be more effective. Hence, in this scenario, a causal attribution method would help the applicant to achieve their desired outcome more efficiently.

5.1.1 CXPlain

Schwab and Karlen [136] transform the task of producing feature importance estimates for an existing predictive model into one of supervised learning. They train separate supervised *causal explanation* (CXPLAIN) models to explain the predictive model. To train the explanation model, they use a causal influence (Sec. 2.7) function that quantifies both the attribution of each input feature and groups of input features to the accuracy of the prediction model.

The causal feature attribution $a_i \in \mathbb{R}$ measures to what extent the i th input feature causally contributed to the predictive model's output \hat{Y} , as the decrease in error, measured by a standard classification loss \mathcal{L} . In other words, a_i denotes the causal influence of adding that feature to the set of input features. Pre-computing the feature attributions for N training samples $\{\mathbf{x}_i\}_{i=1}^N$ takes $N(D + 1)$ evaluations of the target predictive model at training time, where D is the number of input features.

They also provide confidence interval estimates of the level of uncertainty associated with each feature importance estimate \hat{a}_i produced by a CXPLAIN model. By using bootstrap ensemble methods, M explanation models are trained to estimate uncertainty. Each model is trained on N training samples from $\{\mathbf{x}_i\}_{i=1}^N$, sampled randomly with replacement from the original training set.

5.1.2 Generative Causal Explanations

O'Shaughnessy et al. [137] introduce the *generative causal explanations* (GCE) framework for post-hoc explanations based on learned disentangled latent factors that produce a change in the classifier output distribution. For example, consider a black-box color classifier, for whose color classifications one wants to construct explanations. In this setup, one may learn a latent encoder (which they call the *local explainer*) that learns a low-dimensional representation (α, β) describing the color and shape of inputs. Changing α (color) changes the output of the classifier, which detects the color of the data sample, while changing β (shape) does not affect the classifier output. Fig. 5.1 shows a pictorial depiction of this architecture.

To construct causal explanations, the authors propose two components: (i) a method for representing and moving within the distribution of the data and (ii) a metric that quantifies the causal influence (Sec. 2.7) of various aspects of the data on the classifier outputs. To ensure that the learned disentangled representations represent the data distribution, while encouraging a small subset of latent factors to exert a large causal impact on the classifier output, they formulate a corresponding optimization objective.

5.1.3 Asymmetric Shapley Values

Shapley values provide a principled, model-agnostic way to explain model predictions. They are able to capture all the interactions between features that lead to a prediction by relying on coalitional game theory: the idea is to fairly distribute the

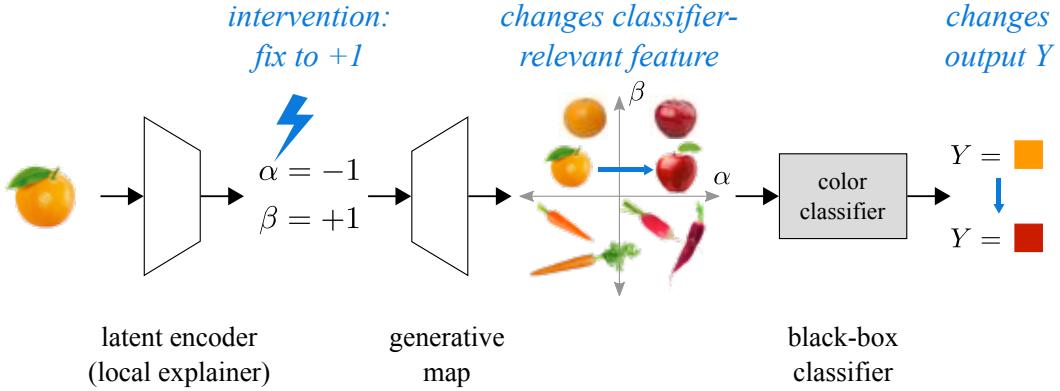


Figure 5.1: Generative Causal Explanations [137]: Changing learned latent factors produces a change in the classifier output statistics. To quantify the attribution of the latent factors, O’Shaughnessy et al. [137] use causal influence measures of variables in a SCM (Sec. 2.7).

“payout” among the features, which naturally quantifies which features contribute to a prediction [129, 138]. They are principled because they uniquely satisfy four intuitive mathematical axioms.

However, Frye et al. [139] argue that Shapley values suffer from a significant limitation: they ignore all causal structure in the data. One of the axioms is *Symmetry*: it places all features on equal footing in the model explanation, by requiring attributions to be equally distributed over features that are identically informative (i.e. redundant). Frye et al. [139] argue that when redundancies exist, we might instead seek a sparser explanation by relaxing this axiom.

For example, two features might be bijectively related to one another if one is known to be the deterministic causal ancestor of the other. In this case, it makes sense to attribute all the importance to the ancestor and none to the descendant, in opposition to the Symmetry axiom. Otherwise, the explanations may obfuscate known causal relationships in the data.

To relax the Symmetry axiom, the authors propose *Asymmetric Shapley values* (ASVs), which uniquely satisfy the other axioms and reduce to Shapley values if the distribution over the ordering in which features are fed to the model is uniform. This relaxation allows the practitioner to place a non-uniform distribution over the ordering that incorporates causal understanding into the explanation. For example, one may place nonzero weight only on permutations in which an ancestor precedes its (known) descendants. In other words, only feature permutations that are consistent with the causal structure between the features have non-zero probability. This approach favors explanations in terms of distal (i.e. root) causes, rather than explanations towards proximate (i.e. immediate) causes.

Frye et al. [139] discuss that ASVs span a continuum between maximally-data agnostic Shapley values, and causality-based explanation methods that often require the exact causal process underlying data. Hence, ASVs allow any knowledge about the data-generating process, however incomplete, to be incorporated into an explanation of its model, without the often-prohibitive requirement of full causal inference. For

example, if causal knowledge is limited, even a single known causal ancestor can be ordered first, with permutations over remaining features uniformly weighted.

Def.: 5.1.1: Distal distribution (ASVs) [139]

Assume that there are D input features, where Π denotes the set of all permutations of them, and $\pi(j) < \pi(i)$ means that feature j precedes feature i under ordering π . Let $\Delta(\Pi)$ be the set of probability measures on Π , so that each $w \in \Delta(\Pi)$ is a map $w : \Pi \rightarrow [0, 1]$ satisfying $\sum_{\pi \in \Pi} w(\pi) = 1$.

Asymmetric Shapley values replace the uniform distribution $w \in \Delta(\Pi)$ of Shapley values [138] with:

$$w_{\text{distal}}(\pi) \propto \begin{cases} 1 & \text{if } \pi(i) < \pi(j) \text{ for any known} \\ & \text{ancestor } i \text{ of descendant } j \\ 0 & \text{otherwise} \end{cases} . \quad (5.1)$$

As a consequence of using Def. 5.1.1, the ASVs of known causal ancestors indicate the effect these features have on predictions while their descendants remain unspecified. The ASVs of the descendants then represent their incremental effect upon specification.

5.1.4 Causal Shapley Values

Janzing et al. [140] argue that there is a misconception in previous work like Lundberg and Lee [138] on Shapley values because these methods use observational conditional distributions rather than interventional ones. While these proposals are conceptually flawed, in practice, their software implementation still works because of its approximate nature.

Heskes et al. [141] remedy this conceptual flaw by proposing *causal Shapley values* (CSVs) that explain the total effect of features on the prediction, taking into account their causal relationships. To incorporate causal knowledge, they replace the conventional conditioning by observation from conditional Shapley values with conditioning by intervention.

In contrast to ASVs, they point out that there is “*no need to resort to asymmetric Shapley values to incorporate causal knowledge*”. Relaxing the Symmetry axiom is orthogonal to their approach. One can additionally get *asymmetric causal Shapley values* that implement both ideas. As for ASVs, a practitioner needs to provide only a partial causal order and a way to interpret dependencies between features that are on an equal footing.

One benefit of CSVs is that they permit a decomposition of the total effect that a feature has on a model’s prediction into direct and indirect effects. The direct effect measures the expected change in prediction when the stochastic feature X_i is fixed to x_i , without altering the other “out-of-coalition” features. The indirect effect

measures the difference in expectation when the distribution of the other “out-of-coalition” features changes due to the additional intervention $\text{do}(X_i = x_i)$.

Chen et al. [142] argue that neither observational (“*true to the data*”) nor interventional (“*true to the model*”) conditional probabilities are preferable in general, but that the choice is application dependent. They present two real data examples from the domains of credit risk modeling and biological discovery to show how a different choice of value function performs better in each scenario, and how possible attributions are impacted by the choice of probability distribution.

Similar to CSVs, Jung et al. [143] introduce do-Shapley values, which do not have the restriction that the model is *accessible*, which means that it can be evaluated for arbitrary input features. Instead, their method is compatible with *inaccessible* models, and uses semi-Markovian causal graphs (DAGs with bidirected edges).

5.2 Contrastive Explanations

Sociological studies have shown that human explanations are typically *contrastive*: they emphasize the causal factors that explain why an event occurred *instead* of another event [144]. By pruning the space of all causal factors, such explanations facilitate easier communication and reduce the cognitive load for both explainer and explainee [145].

Contrastive explanations are typically *counterfactual* in the sense that they estimate an altered version of an observation that would have changed the model’s prediction, given our knowledge of the model’s prediction from the original datapoint. The basic implementation of this idea, which we will henceforth refer to as *counterfactual explanation*, is conceptually simple and does not require much causal machinery. Furthermore, we will look at *algorithmic recourse* (AC): instead of providing an *understanding* of the least *distant* point in feature space that results in the desired prediction, AC aims at producing the minimal *cost* set of actions for the individual, taking the dependencies among observed variables into account in the form of causal knowledge.

5.2.1 Counterfactual Explanations

Counterfactual explanations (CE) explain a prediction by computing a (usually minimal) change of an individual’s features that would cause the underlying model to classify it in a desired class [146]. By showing feature-perturbed versions of the same person who would have received the loan, *counterfactual explanations* can provide actionable information about what to do in the future to secure a better outcome. CE are *counterfactual* in that they consider the alteration of an entity in the history of the event P , where P is the undesired model output [147].

For example, a feature instantiation that would have changed the prediction in the above example would be “you would have received the loan if your income was higher by \$10k”. In contrast to explanation methods that rely on approximating the

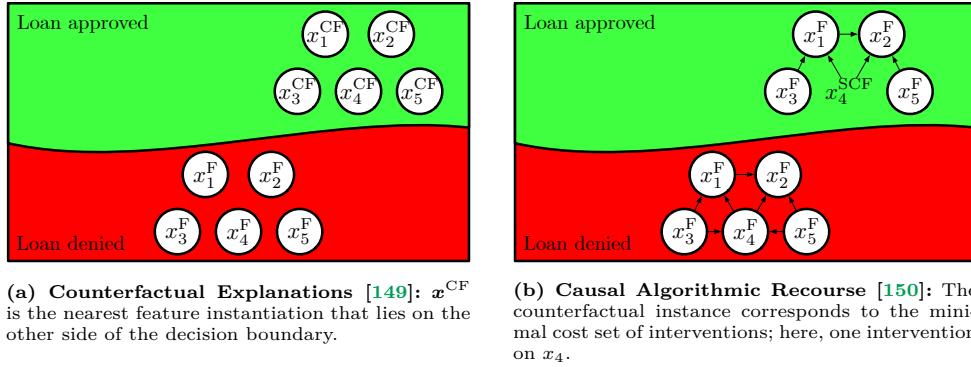


Figure 5.2: Counterfactual explanations vs. Causal Algorithmic Recourse, illustrated by an individual who got their loan application denied, represented by features $\mathbf{x}^F = \{x_1^F, \dots, x_5^F\}$.

classifier's decision boundary [148], counterfactual explanations are always truthful w.r.t. the underlying model by using actual predictions of the algorithm.

Def.: 5.2.1: Counterfactual Explanation [149]

A counterfactual explanation \mathbf{x}^{CE} (or *nearest contrastive explanation*) for an individual \mathbf{x}^F is given by a solution to the following optimization problem:

$$\mathbf{x}^{\text{CE}} \in \arg \min_{\mathbf{x} \in \mathcal{X}} \text{dist}(\mathbf{x}, \mathbf{x}^F) \text{ s.t. } h(\mathbf{x}) \neq h(\mathbf{x}^F), \mathbf{x} \in \mathcal{P} \quad (5.2)$$

where $\text{dist}(\cdot, \cdot)$ is a similarity metric on \mathcal{X} , and \mathcal{P} is an optional set of plausibility constraints to reflect feasibility, or diversity of the obtained counterfactual explanations [151].

Naively explaining predictions may interfere with constructive recourse, if they violate plausibility and actionability (feasibility) constraints for suggested feature changes. For example, actions such as asking the individual to reduce their age or change their race are not feasible. Therefore, common plausibility constraints include (i) domain-consistency, (ii) density-consistency, and (iii) prototypical-consistency, while actionability constraints distinguish between (i) actionable and mutable, (ii) mutable but non-actionable, and (iii) immutable (and non-actionable) [133].

There are many flavors and variants of CE methods. Schut et al. [152] propose generating CEs in a white-box setting without an auxiliary model, by using the predictive uncertainty of the classifier. Abid et al. [153] outlines an approach for producing counterfactual explanations in terms of human-understandable concepts. Mahajan et al. [154] propose to preserve causal relationships among input features through feasibility constraints. For a more detailed review of CE methods, we refer the reader to Verma et al. [132].

5.2.2 Causal Algorithmic Recourse via Minimal Interventions

Previously, we looked at counterfactual explanations, which demonstrate “how the world would have (had) to be different for a desirable outcome to occur” [149]. However, these explanations may not always be translated into optimal *recourse actions*, a recommendable set of actions to help an individual to achieve a favorable outcome while respecting the costs of the actions [155]. The field of *Causal Algorithmic Recourse* (or *consequential recommendations*) deals with generating such “what should be done in the future” recommendations while respecting both causal relations between features and as costs of actions [150].

From a causal perspective, actions correspond to interventions (Sec. 2.3.1). The modeling goal is to predict the effect of such interventions on one individual’s situation to ascertain whether or not the desirable outcome is achieved [147].

Def.: 5.2.2: Algorithmic Recourse via Minimal Interventions (ARMI) [147]

Consider an SCM $\mathcal{M} = \{\mathbb{F}, \mathbb{X}, \mathbb{U}\}$ consisting of observed variables $\mathbb{X} \in \mathcal{X}$, exogenous variables $\mathbb{U} \in \mathcal{U}$, and structural equations $\mathbb{F} : \mathcal{U} \rightarrow \mathcal{X}$. ARMI seeks the minimal cost set of **interventions** that results in a counterfactual instance yielding the favourable output from $h(\cdot)$:

$$\mathbf{a}^* \in \arg \min_{\mathbf{a} \in \mathcal{A}} \text{cost}(\mathbf{a}; \mathbf{x}^F) \quad \text{s.t. } h(\mathbf{x}^{\text{SCF}}) \neq h(\mathbf{x}^F) \quad (5.3)$$

$$\mathbf{x}^{\text{SCF}} = \mathbb{F}_A(\mathbb{F}^{-1}(\mathbf{x}^F)), \quad \mathbf{x}^{\text{SCF}} \in \mathcal{P}, \quad \mathbf{a} \in \mathcal{A} \quad (5.4)$$

where $\mathbf{a}^* \in \mathcal{A}$ directly specifies the set of feasible actions to be performed for minimally costly recourse, with $\text{cost}(\cdot; \mathbf{x}^F) : \mathcal{A} \times \mathcal{X} \rightarrow \mathbb{R}_+$, and $\mathbf{x}^{*\text{SCF}} = \mathbb{F}_{\mathbf{a}^*}(\mathbb{F}^{-1}(\mathbf{x}^F))$ denotes the resulting structural counterfactual.

Although \mathbf{x}^{SCF} is a counterfactual instance, it does not need to correspond to the nearest counterfactual explanation \mathbf{x}^{CE} , resulting from Eq. (5.2). Importantly, using the formulation in Eq. (5.4), Karimi et al. [147] prove formally that consequential recommendations generated by optimizing w.r.t the decision boundary may be sub-optimal, as they do not benefit from the causal effect of actions towards changing the prediction.

Note that Def. 5.2.2 requires a causal model. In practice, we may learn this model from data under the assumption of no hidden confounding, but this assumption can be unrealistic for some scenarios. von Kügelgen et al. [156] relax this assumption and propose a partial identification approach that is compatible with unobserved confounding and arbitrary structural equations. Their approach bounds the expected counterfactual effect of recourse actions.

5.3 Open Problems

5.3.0.1 Unifying Feature Attribution and Explanations

Attribution and explanation methods can be complementary. For example, as Verma et al. [132] point out that explanations need to emphasize what should remain the same *in addition* to what should change in order to achieve a desired outcome.

Imagine an ML model used for loan prediction that uses “income” and “years of employment” as inputs. The ML model rejected the loan request of an individual and recommended an increase in “income”. Therefore, the individual changed their job and their feature “years of employment” was reset to zero. Despite the increase in “income” the model still rejected the loan request since it did not specify that the other feature should not change.

Having access to causal attribution scores may prevent such situations: a low score for the feature ‘years of employment’ would inform the individual that changing its value should not cause any interference for the desired outcome. First steps towards combining both approaches can be found in [157, 158, 159, 160, 161].

Fokkema et al. [162] interpret a counterfactual explanation \mathbf{x}^{CE} as attributions that describe a perturbation of \mathbf{x}^{F} . Since the differences between \mathbf{x}^{F} and \mathbf{x}^{CE} can be interpreted as the changes needed to flip the class, we can regard $\varphi_f = \mathbf{x}^{\text{CE}} - \mathbf{x}^{\text{F}}$ as the attribution vector.

5.3.0.2 Scalability and Throughput

Computing explanations comprises solving optimization problems that are often very expensive and difficult to solve. On the attribution side, the computation time often increases exponentially with the number of features, e.g., in the case of Shapley values, making their exact solution computationally intractable when we have more than a few features. On the contrastive explanations side, incorporating plausibility and actionability constraints makes the problem NP-hard or even NP-complete when solving for [133], e.g., integer-based variables [163], neural networks, or quadratic objectives and constraints.

To make explanations deployable in large-scale systems with many users, approximate methods are highly needed. On the attribution side, one solution might be to compute contributions for only a few samples of the possible coalitions [129]. Lundberg and Lee [138] propose the approximate Kernel SHAP method, which assumes feature independence. As first steps towards that goal, Mahajan et al. [164] learn a VAE which can generate multiple counterfactuals for any given input at once.

5.3.0.3 Dynamics

Most explanation methods assume a static black-box model that does not change over time. However, distribution shifts occur frequently in many real-life ML do-

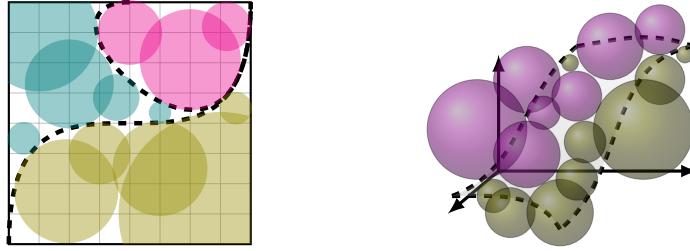


Figure 5.3: Model Stealing through Contrastive Explanations [133]: Karimi et al. [133] illustrate the model stealing process in 2D and 3D using hypothetical non-linear decision boundaries. “How many optimal contrastive explanations are needed to extract the decision regions of a classifier?” can be formulated as “How many factual balls are needed to maximally pack all decision regions?”

mains: user behavior may change over time, models get updated, and a decision-making system’s utility function may be modified. Naturally, these issues arise for model explanations too, and there exists only limited work to address them [165, 166, 167].

5.3.0.4 Security and Privacy

Consider a deployed ML system that provides explanations to its users, e.g., through access to an API. In such settings, an adversary may combine predictions and explanations to extract an approximate model as well as information about the data used to train it.

For example, in the case of contrastive explanations, three possible attack strategies are *model extraction* [168] by (i) augmenting the surrogate’s model dataset with counterfactual instances/labels, and (ii) approximating its decision boundaries through optimal contrastive explanations, as visually illustrated in Fig. 5.3 and formalized by Karimi et al. [133], and (iii) *membership inference attacks* which quantify the information explanations leak about the presence of a datapoint in the training set for a model [169].

In the case of models using causal knowledge about the DGP (Secs. 5.1.3, 5.1.4 and 5.2.2), we hypothesize that it is important to guard this causal knowledge against causal-discovery-like model extraction attacks too. An attacker who is allowed to perform enough queries may use an active learning approach to actively intervene on input points and yield the causal graph of the explanation model [170].

Slack et al. [171] describes the vulnerabilities of counterfactual explanations and show that they may converge to drastically different counterfactuals under a small perturbation, indicating that they are not robust. To this end, they propose an objective to train seemingly fair models where counterfactual explanations find lower cost recourse under perturbations.

There has been very limited work on guarding explanation methods against such concerns. Shokri et al. [169] propose differentially private (DP) model explanations. They design an adaptive DP gradient descent algorithm that finds the minimal pri-

vacy budget required to produce accurate explanations. Naidu et al. [172] investigate the interpretability of DP-trained medical imaging models.

5.3.1 Robustness vs. Recourse Sensitivity

Two desirable properties of explanation methods are *robustness* and *recourse sensitivity* [162]. Robustness refers to changes in the inputs: if we reason that similar users should get similar options for recourse, then small changes in the input x should not cause large jumps in the explanation $\varphi_f(x)$, i.e. φ_f should be continuous. A recourse sensitive attribution method permits that the user can always achieve sufficient utility by moving in the direction of the vector $\varphi_f(x)$.

Fokkema et al. [162] point out that attribution methods and counterfactual explanations cannot be robust and recourse sensitive at the same time. Their main conclusion is that for any way of measuring utility, there exists a model f for which no attribution method φ_f can be both recourse sensitive and continuous. Future work should therefore consider workarounds to circumvent Fokkema et al. [162]'s impossibility result.

6

Causal Fairness

Machine learning models increasingly assist in life-changing decisions like parole hearings, loan applications, and university admissions. Decisions in these areas may have ethical or legal implications, so a model practitioner must consider the societal impact of their work.

If the data used to train an algorithm contains demographic disparities against certain races, genders, or other groups, the algorithm will too. Instead of just maximizing measures that quantify desirable statistical properties of a predictor, *algorithmic fairness* aims to provide criteria that can be used to assess a model's fairness and mitigate harmful disparities.

Causality plays a significant role in investigating a model's fairness because it typically depends on the causal structure of the data: for some causal graphs, it can be fair to include certain input features, while for others, it is not. Unfortunately, statistical-based fairness measures are oblivious to distinguishing between different causal relationships among input variables and quickly fail to detect discrimination in the presence of statistical anomalies such as Simpson's paradox [173].

For example, imagine that instead of considering the causal structure of the data, we would down-weight or discard sensitive attributes. Doing so may not result in a fair procedure as the sensitive attribute often correlates with other attributes. Especially in large feature spaces, sensitive attributes are often redundant given the other features. The classifier may learn a redundant encoding for the sensitive features in such settings.

Barocas et al. [174] illustrate this issue of learning classifiers using sensitive attributes without explicitly being asked to with the following example: Consider a fictitious start-up that sets out to predict your income from your genome. DNA can predict income better than random guessing because DNA encodes information about ancestry, which correlates with income in some countries, such as the United States. Hence, the learned classifier may use ancestry in an entirely implicit manner. Unfortunately, removing redundant encodings of ancestry from the genome is a difficult task that cannot be accomplished by removing a few individual genetic markers.

In contrast, by embracing the sensitive attributes and their causal relationships with the other input variables, we can intervene on them and analyze how changing their values affects the model's predictions. A fair predictor would be invariant w.r.t. such interventions. In the DNA example, this means that – all other things being equal –

a fair predictor remains invariant w.r.t. ancestry by ensuring that interventions on the genetic markers do not influence the predictions.

In this chapter, we discuss two classes of causality-based fairness criteria: *counterfactual* (CFF) and *interventional* fairness (IF). CTF criteria assess the unfair influence of protected attributes on the outcome variable by analyzing counterfactuals. These criteria primarily differ based on how the counterfactuals are constructed. Further, IF approaches aim to relax some of the strong assumptions CFF require to make them more practical in real-life settings.

Class	Criterion	Key Idea	Ref.
Ctf.	Counterfactual Fairness (CF)	A decision is fair towards an individual if it is the same in (a) the actual world and (b) a counterfactual world where the individual belonged to a different demographic group	Sec. 6.2.1
	Path-Specific CF	Refinement of CF by distinguishing between fair and unfair pathways in causal DAG	Sec. 6.2.2
	Causal Explanation Formula	Decomposition of CF into direct, indirect, and spurious discrimination	Sec. 6.2.3
Int.	Proxy Fairness	Intervention on proxy variables of protected attributes instead of the latter directly	Sec. 6.3.1
	Justifiable Fairness	Opens up the possibility of not requiring a full causal model: requires only separation of variables into admissible and inadmissible	Sec. 6.3.2

Table 6.1: Criterion Overview of Causal Fairness.

Notation

- A** Protected (or “sensitive”) attributes of an individual
- X** Remaining attributes of an individual
- U** Exogenous attributes of an individual
- W** Attributes between **X** and **Y**
- Y** Outcome
- \hat{Y} Outcome predictor

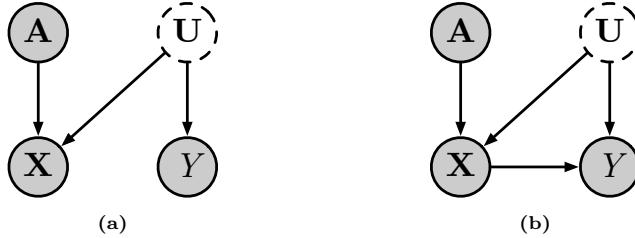


Figure 6.1: Two possible causal graphs considered by Counterfactual Fairness [175]: (a) shows the setup where only the latent confounder **U** causes the outcome; in contrast, (b) considers both **U** and **X** influencing **Y**.

6.1 Two more detailed examples

The aforementioned DNA example clarifies why one has to account for the causal structure of the data to ensure fair predictions. Now, we give two more detailed examples with concrete causal DAGs, as shown in Fig. 6.1. These examples are taken from Kusner et al. [175].

In both scenarios, using the seemingly fair attributes **X** to predict **Y** is unfair.

Fig. 6.1a illustrates a setting where only latent confounder **U** causes **Y**. Suppose, for example, a car insurance company wants to price insurance for car owners by predicting their accident rates **Y**, assuming there is an unobserved variable corresponding to aggressive driving **U** which (a) increases the likelihood that drivers will have an accident and (b) increases the likelihood of individuals preferring red cars, captured by **X**. Furthermore, people with specific demographic characteristics may prefer driving red cars. Using the red car feature **X** to predict the accident rate **Y** seems unfair, because these individuals are no more likely than anyone else to be aggressive or to get into accidents.

Fig. 6.1b illustrates the situation where both **U** and **X** influence the outcome **Y**. Consider a crime prediction scenario in which a municipality wishes to estimate crime rates by neighborhood to allocate police resources. The data set contains residents' neighborhood **X**, demographics **A**, and binary labels **Y** indicating a criminal arrest history. Due to historically segregated housing, the location **X** depends on **A**, and locations **X** do not have equally many police resources (unobserved). The number of arrests **Y** are higher in areas with more police resources **X**. **U** represents the totality of socioeconomic factors and policing practices that influence where an individual may live and how likely they are to be arrested. While **Y** depends on **X**, the predictor \hat{Y} could be influenced by values of **A** that are not explained by **U**.

For another, even more exhaustive example, including calculations of statistical and causal fairness criteria using concrete data, we direct the reader to Makhlouf et al. [173].

6.2 Counterfactual Fairness Criteria

6.2.1 Kusner et al. [175]'s Counterfactual Fairness

We start with the *counterfactual fairness* criterion by Kusner et al. [175]. It ensures that predictors of the outcome Y , which any individual with features $\{\mathbf{x} \cup \mathbf{a}\}$ receives in the real world, is the same as the one they would receive in a counterfactual world in which only the protected attribute of the individual $\mathbf{A} = \mathbf{a}'$ has changed, everything else remaining equal. In other words, the predictor is invariant to both counterfactuals. We formalize this criterion by recalling the definition of a counterfactual in Sec. 2.3.2.

Def.: 6.2.1: Counterfactual Fairness [175]

Given latent exogenous variables \mathbf{U} that are not caused by any of \mathbf{X} or \mathbf{A} , we say that the predictor \hat{Y} is counterfactually fair if

$$p(\hat{y}_a(\mathbf{u}) | \mathbf{x}, \mathbf{a}) = p(\hat{y}_{a'}(\mathbf{u}) | \mathbf{x}, \mathbf{a}), \quad \forall \mathbf{x}, \mathbf{a}, \mathbf{a}', y. \quad (6.1)$$

We interpret the final expression of Eq. (6.1) as the probability that \hat{Y} predicts y for a given individual for which we observe features \mathbf{x} and protected attribute \mathbf{a} , had the protected attribute been \mathbf{a}' instead of \mathbf{a} . An analogous interpretation of the first expression asserts that it equals $p(y | \mathbf{x}, \mathbf{a})$.

6.2.2 Path-Specific Counterfactual Fairness

The aforementioned counterfactual fairness criterion (Definition 6.2.1) considers the full effect of the sensitive attribute on the decision as problematic. However, this is not the case in certain scenarios, and we want to be more precise about their contribution to unfair predictions. Nabi and Shpitser [176] identify that discrimination can be formalized as the presence of an effect of a covariate on the outcome along specific causal pathways, but not all of them.

More concretely, Chiappa [177] substantiate this concern by using the famous Berkeley alleged gender bias case, commonly used as a textbook example of Simpson's Paradox. In this scenario, data on college admissions showed a bias for male applications overall. However, the university rejected female applicants more often than male applicants because they applied to more competitive departments with lower admission rates. Such an effect of gender through department choice is not unfair.

Motivated by this, Chiappa [177] propose *path-specific counterfactual fairness*, a more fine-grained fairness criterion that deals with sensitive attributes affecting the decision along both fair and unfair pathways. It states that a decision is fair toward an individual if it coincides with the one the model would have taken in a counterfactual world in which the sensitive attribute along the unfair pathways differed.

Def.: 6.2.2: Path-Specific Cf. Fairness [176, 177]

Define $\mathcal{P}_{\mathcal{G}_A}$ as the set of all directed paths from \mathbf{A} to Y in \mathcal{G} which correspond to all unfair chains of events where \mathbf{A} causes Y . Let $\mathbf{X}_{\mathcal{P}_{\mathcal{G}_A}^c} \subseteq \mathbf{X}$ be the subset of covariates not present in any path in $\mathcal{P}_{\mathcal{G}_A}$. Then, Predictor \hat{Y} is path-specifically counterfactually fair w.r.t. path set $\mathcal{P}_{\mathcal{G}_A}$ if $\forall \mathbf{x}, \mathbf{a}, \mathbf{a}', y :$

$$p(\hat{y}_{\mathbf{a}, \mathbf{x}_{\mathcal{P}_{\mathcal{G}_A}^c}}(\mathbf{U}) | \mathbf{x}, \mathbf{a}) = p(\hat{y}_{\mathbf{a}', \mathbf{x}_{\mathcal{P}_{\mathcal{G}_A}^c}}(\mathbf{U}) | \mathbf{x}, \mathbf{a}). \quad (6.2)$$

Similar to Chiappa [177], Wu et al. [178] define a path-specific counterfactual fairness criterion too. Their criterion enables us to reason about any population sub-group, not just individuals. Further, they address situations where identifiability may not hold by proposing a method that bounds the path-specific effects.

6.2.3 Causal Explanation Formula

Zhang and Bareinboim [179] propose the *causal explanation formula*, which allows practitioners to decompose the total observed disparity of decisions into three fine-grained measures.

First, the authors assume that there is a disadvantaged group \mathbf{a}_1 and an advantaged one, \mathbf{a}_0 . Further, \mathbf{W} denotes all observed intermediate variables between \mathbf{A} and Y . Next, they define different effects; if the effect is non-zero, the predictor is unfair. They distinguish between direct and indirect spurious effects; the latter considers the back-door paths between \mathbf{A} and Y , that is, paths with an arrow into \mathbf{A} .

Def.: 6.2.3: Counterfactual Effects [179]

We denote all observed mediator variables between \mathbf{A} and Y as \mathbf{W} . Then, the direct effect (DE), indirect effect (IE), and spurious effect (SE) are respectively defined as:

$$\text{DE}_{\mathbf{a}_0, \mathbf{a}_1}(y | \mathbf{a}) = p(y_{\mathbf{a}_1, \mathbf{W}_{\mathbf{a}_0}} | \mathbf{a}) - p(y_{\mathbf{a}_0} | \mathbf{a}), \quad (6.3)$$

$$\text{IE}_{\mathbf{a}_0, \mathbf{a}_1}(y | \mathbf{a}) = p(y_{\mathbf{a}_0, \mathbf{W}_{\mathbf{a}_1}} | \mathbf{a}) - p(y_{\mathbf{a}_0} | \mathbf{a}), \quad (6.4)$$

$$\text{SE}_{\mathbf{a}_0, \mathbf{a}_1}(y) = p(y_{\mathbf{a}_0} | \mathbf{a}_1) - p(y | \mathbf{a}_0). \quad (6.5)$$

Next, they aim to provide intuition on how the different effects relate to each other by decomposing the total variation (TV) into them. TV is the difference between the conditional distributions of Y when passively observing \mathbf{A} changing from \mathbf{a}_0 to \mathbf{a}_1 . Formally, the TV of event $\mathbf{A} = \mathbf{a}_1$ on $Y = y$ (with baseline \mathbf{a}_0) is defined as:

$$\text{TV}_{\mathbf{a}_0, \mathbf{a}_1}(y) = p(y | \mathbf{a}_1) - p(y | \mathbf{a}_0). \quad (6.6)$$

Def.: 6.2.4: Causal Explanation Formula [179]

The total variation (TV), spurious effect (SE), indirect effect (IE), and direct effect (DE) are related as

$$\text{TV}_{\mathbf{a}_0, \mathbf{a}_1}(y) = \text{SE}_{\mathbf{a}_0, \mathbf{a}_1}(y) + \text{IE}_{\mathbf{a}_0, \mathbf{a}_1}(y | \mathbf{a}_1) - \text{DE}_{\mathbf{a}_1, \mathbf{a}_0}(y | \mathbf{a}_1), \quad (6.7)$$

$$\text{TV}_{\mathbf{a}_0, \mathbf{a}_1}(y) = \text{DE}_{\mathbf{a}_0, \mathbf{a}_1}(y | \mathbf{a}_0) - \text{SE}_{\mathbf{a}_1, \mathbf{a}_0}(y) - \text{IE}_{\mathbf{a}_1, \mathbf{a}_0}(y | \mathbf{a}_0). \quad (6.8)$$

For example, the first formula shows that the total disparity experienced by the individuals who have naturally attained \mathbf{a}_1 equals the disparity experienced via spurious discrimination, plus the advantage is lost due to indirect discrimination minus the advantage it would have gained without direct discrimination.

6.3 Interventional Fairness

Counterfactual fairness involves modeling counterfactuals on an individual level, which is problematic. For example, [180, 181] point out that effects of race or gender are challenging to model *even at the group level*. Thus, interventions on such (often ill-defined) protected attributes are typically hard to envision. Kilbertus [182] illustrates this through the following thought experiment: imagine a pregnant woman's job application gets rejected. Could we imagine her life as a man in that world? Was she born male or perceived as male during the hiring process? Is she a pregnant man now? The notion of counterfactual fairness renders meaningless if we compare (fictitious) individuals who are vastly different from each other.

6.3.1 Proxy Fairness

To address these concerns, Kilbertus et al. [183] examines population-level interventional distributions, which they call *proxy discrimination*. The idea is to separate the protected attribute **A** from its potential proxies, such as names, visual features, languages spoken at home, etc. Hence, an intervention based on proxy variables presents a more manageable problem. In practice, we are frequently limited to imperfect measurements of **A**, so separating the root concept from a proxy is prudent [182]. A proxy **P** is a descendant of **A** in the assumed causal graph.

Def.: 6.3.1: Proxy Fairness [182]

Predictor \hat{Y} exhibits no proxy discrimination based on a proxy **P** if

$$p(y | \text{do}(\mathbf{p})) = p(y | \text{do}(\mathbf{p}')), \quad \forall \mathbf{p}, \mathbf{p}'. \quad (6.9)$$

6.3.2 Justifiable Fairness

All previous approaches require knowledge of the causal graph. To deal with settings where a causal graph is missing but a partial knowledge of admissible variables is given, Salimi et al. [184] formalize interventional fairness as a database repair problem. They present data pre-processing algorithms for providing fairness guarantees about classifiers trained on pre-processed training data.

First, they assume that the causal graph is given and define the *K-fair* criterion, which captures group-level fairness, similar to proxy fairness.

Def.: 6.3.2: K-Fair [184]

For a set of attributes $\mathbf{K} \subseteq \mathbf{X}$, we say that the predictor \hat{Y} is \mathbf{K} -fair w.r.t. protected attributes \mathbf{A} if $\forall \mathbf{k}, \mathbf{a}, \hat{y}$:

$$p(\hat{y} | \text{do}(\mathbf{a}, \mathbf{k})) = p(y | \text{do}(\mathbf{a}', \mathbf{k})). \quad (6.10)$$

A model is deemed *interventionally fair* if it is \mathbf{K} -fair for every set \mathbf{K} . This notion differs from proxy fairness in that it ensures that \mathbf{A} does not affect Y in any configuration obtained by fixing other variables to arbitrary values. As opposed to counterfactual fairness, it does not try to capture fairness at the individual level, so it employs level-2 interventions.

Next, they define the *justifiable fairness* criterion, which allows the user to distinguish only admissible and inadmissible variables. The former variables are a subset of the protected attributes \mathbf{A} through which it is still permissible to influence the outcome.

Def.: 6.3.3: Justifiable Fairness [184]

A predictor \hat{Y} is justifiably fair if it is \mathbf{K} -fair w.r.t. all supersets $\mathbf{K} \supseteq \mathbf{A}$.

6.4 Fairness under Distribution Shifts

Singh et al. [185] study the problem of learning fair prediction models under covariate shift, i.e., with test set covariates distributed differently from the training set. Given the ground-truth causal graph describing the data and anticipated changes, they propose an approach based on feature selection that exploits conditional independencies in the data to estimate accuracy and fairness metrics for the test set.

Schrouff et al. [186] evaluates how realistic the assumptions made by Singh et al. [185] as well as other works are [187, 188]. They categorize distribution shifts into four categories: *demographic shift*, *covariate shift*, *label shift*, and *compound shift*, independently of the causal graph considered. Their study examines two real-world applications in dermatology and electronic health records and shows that clinically plausible shifts directly affect aspects of the data distribution simultaneously. There-

fore, and as empirically demonstrated, compound shifts impact the transferability of fairness properties in these applications.

6.5 Open Problems

6.5.1 Alternatives to equality

Most fairness definitions in the literature emphasize equality, ensuring that each individual or group receives the same resources, attention, or outcome. In contrast, *equity* [189, 190] has received little attention, which means that all individuals and groups have access to the resources they need to thrive. An exciting future direction is to operationalize this definition and examine how it enhances or contradicts existing definitions of fairness [3].

Another alternative to equality might be considering a model’s *harm*: Richens et al. [191] propose a definition for a harm criterion that should be able to answer the following three questions precisely: “***Q1:*** *Did the actions of an agent cause harm, and if so, how much?* ***Q2:*** *How much harm can we expect an action to cause before taking it?* ***Q3:*** *How can we identify actions that balance the expected harms and benefits?*” They then propose a family of counterfactual objective functions that mitigate harm.

6.5.2 Fairness beyond predictions

In this chapter, we studied predictive fairness criteria that evolved around supervised models’ predictive output. However, there is growing interest in assessing the fairness of ML techniques beyond prediction, e.g., in the realm of recourse methods (Sec. 5.2.2).

Gupta et al. [192] put forward that models should not limit the ability to achieve recourse to only those with access to expensive resources. Put another way; the model must fairly distribute the rights of recourse across (demographically defined) groups. von Kügelgen et al. [193] investigate the fairness of causal algorithmic recourse actions and conclude that there is still much work left. For example, the question of whether it is appropriate to perform societal interventions on all individuals in a subgroup. Huan et al. [194] deal with *equality of efforts*: they seek to determine whether the efforts made to achieve the same outcome level are the same or different between those in the protected group and those in the unprotected group.

6.5.3 Partial identification

Causal quantities, especially counterfactuals, are often unidentifiable (Sec. 2.6). That means that we can not compute them from a statistical quantity. While previous causal fairness work has partly addressed these issues [175, 178, 195], we foresee more important work to be done in sensitivity analysis through realistic simulations

to thoroughly test these methods and alter aspects of them to understand how violations of assumptions impact fairness estimation.

6.5.4 Manipulability of social categories

From a social science perspective, it is heavily debated whether social categories admit interventions. Kohler-Hausmann [196] criticize that counterfactuals require us to reduce race to only the signs of the category, e.g., the skin color or phenotype of a race. Hu [197] argues that causal theories about social categories such as race involve “*ineliminable substantive moral and political considerations, a feature for which interventionism can not well account*”. Hu and Kohler-Hausmann [198] challenge the validity of specifying a social group such as gender as a variable in a DAG while still assuming the model’s modularity assumption. More broadly, Kasirzadeh and Smart [199] review various papers at this intersection between sociology and causal modeling, concluding that social categories often do not admit counterfactual manipulation.

6.5.5 Trade-offs between fairness criteria

Kleinberg et al. [200], Corbett-Davies et al. [201], Chouldechova [202] show that fairness criteria can be opposed to each other, and sometimes, one cannot have multiple criteria simultaneously satisfied. Hence, in practice, the trade-off between multiple admissible criteria has to be balanced. Gultchin et al. [203] shed some first light on how to balance certain measures by using a multi-objective framework. Similarly, Nilforoshan et al. [204] highlight limitations and potential adverse consequences of causal fairness criteria. They demonstrate cases in which these criteria lead to policies that would be disfavored by every stakeholder.

6.5.6 Unfair models despite fair data

In debates about model fairness, a common assumption is that unfairness arises through biases in data. However, Ashurst et al. [205] point out that models can produce unfair predictions even when the training labels are fair. The authors refer to this phenomenon as *introduced unfairness* and investigate the conditions under which it may arise. Taking a causal perspective, they show that the notion of introduced unfairness can be applied to causal definitions of fairness too. For example, they consider *path-specific introduced effects* as the difference in some path-specific effects on labels and predictions.

Minimal work exists investigating when fair training labels may not yield a fair model. Ashurst et al. [205] conclude that it is difficult to rule out unfair disparities, even when the criteria of causal fairness are met. Future research may stress-test the validity of the in this section introduced fairness criteria.

7

Causal Reinforcement Learning

Reinforcement Learning (RL) is a framework in which autonomous agents interact with their environments to learn optimal behaviors, improving over time through trial and error. Its central goal is learning how to map situations to actions while maximizing a numerical reward signal [206]. RL researchers typically formalize their problem setup by using *Markov decision processes* (MDPs), which include three ingredients: sensation (observation), action, and goal (reward).

In this section, we highlight methods benefiting RL problems by exploiting the causal paradigm (and not the other way around, see, e.g., [207]). We refer to this family of approaches as *Causal Reinforcement Learning*. We summarize the benefits of these approaches in Table 7.1.

Notation

t	discrete time step
T	final time step of an episode t
A_t	action at time t
S_t	state at time t
L_t	regret/loss at time t
R_t	reward at time t
\mathcal{R}	return
π	policy (decision-making rule)
$\pi(a s)$	probability of taking action a in state s
s, s'	true states
x, x'	observed states
$v_\pi(s)$	value of state s under policy π (expected return)
$q_\pi(s, a)$	value of taking action a in state s under policy π
τ	trajectory, i.e., $\tau = \{x_t, a_t, x_{t+1}\}_{t=1}^T$

7.1 Isn't RL already "Causal"?

The short answer is yes. For many years, researchers have argued that there exist links between certain flavors of RL and causal inference [7, 208, 209, 210, 211, 212, 213, 214]. For example, Bottou et al. [208] frame bandit problems and Markov decision processes (MDPs) as special cases of causal models. However, despite con-

Problem	Output	Benefits over non-causal RL	Ref.
Causal Bandits	$\hat{\pi} = \arg \min_{\pi \in \Pi} L_n(\pi)$	Optimal simple regret guarantees	Sec. 7.2
Model-Based RL	$\hat{\theta} = \arg \min_{\theta \in \Theta} \ell(\theta, (R_{t+1}, S_{t+1}))$	Deconfounding	Sec. 7.3
Multi-Environment RL	$\hat{\pi} = \arg \max_{\pi \in \Pi} \mathbb{E}_{c \sim p(c)} [\mathcal{R}(\pi, \mathcal{M}^c)]$	Interpretable task embeddings, systematic generalization	Sec. 7.4
Off-Policy Policy Evaluation	$\hat{v}_\pi(s) = \mathbb{E}_{x \sim d_0} \left[\sum_{t=0}^{T-1} \gamma^t r_t \mid x_0 = x \right]$	Deconfounding	Sec. 7.5
Imitation Learning	$\hat{\pi} = \arg \min_{\pi \in \Pi} \mathbb{E}_{x \sim d_{\pi^*}} [\ell(x, \pi, \pi^*(x))]$	Deconfounding	Sec. 7.6
Credit Assignment	$\mathcal{M}_{a_t \rightarrow r_{t+k}}$ or $\mathcal{M}_{a_t \rightarrow s_{t+1}}$ or $\mathcal{M}_{a_t^i \rightarrow a_t^j}$	Intrinsic reward, Data-efficiency	Sec. 7.7
Counterfactual Data Augmentation	$\tilde{\tau} = \{\tilde{x}_t, \tilde{a}_t, \tilde{x}_{t+1}\}_{t=1}^T$	Data-efficiency	Sec. 7.8
Agent Incentives	Incentive criteria and measures	Avoiding unintended harmful behavior	Sec. 7.9

Table 7.1: Problem Overview of Causal Reinforcement Learning.

ceptual similarities, the two fields have mainly focused on different goals: on the one hand, the RL community has focused on building algorithms to *maximize rewards*; on the other hand, the focus in the causality literature has been on the *identifiability and inferences* of or based on given causal structure [214].

We attribute one reason for different foci among both communities to the type of applications each tackles. The vast majority of literature on modern RL evaluates methods on synthetic data simulators, able to generate large amounts of data. For instance, the popular ALPHAZERO algorithm assumes access to a boardgame simulation that allows the agent to play many games without a constraint on the amount of data [215]. One of its significant innovations is a *tabula rasa* algorithm with less handcrafted knowledge and domain-specific data augmentations. Some may argue that ALPHAZERO proves Sutton's *bitter lesson* [216]. From a statistical point of view, it roughly states that given more compute and training data, general-purpose algorithms with low bias and high variance outperform methods with high bias and low variance.

In contrast, in the causal inference literature, we are typically given a limited-size observational dataset from an unknown policy and unknown environment and cannot interact with the environment in an online fashion. The reason behind that problem convention is that much of causal inference methodology originated from domains like medicine, econometrics, online advertisements, and the social sciences, in which conducting experiments is infeasible due to ethical or cost/time-consuming reasons. Nonetheless, causal inference is typically applied in contexts where decisions directly influence human individuals. For example, Bottou et al. [208] illustrates how causal inference can be used in the ad placement system associated with the Bing search engine.

In causal inference, instead of maximizing a reward function in expectation, a common target is to learn the heterogeneous treatment effect (HTE) (Sec. 11.2.1.1): it quantifies the expected effect of changing the observed treatment \mathbf{t} to a different treatment \mathbf{t}' for a certain subgroup characterized by covariates \mathbf{x} , denoted as

$$\tau(\mathbf{t}', \mathbf{t}, \mathbf{x}) \triangleq \mathbb{E}[Y | \mathbf{x}, \text{do}(\mathbf{t}')] - \mathbb{E}[Y | \mathbf{x}, \text{do}(\mathbf{t})]. \quad (7.1)$$

By assumption, we observe only one treatment and outcome pair per subgroup. By using an estimator $\hat{\tau}$, decision-makers may reason about what treatments work well for what subgroups. Due to high risks in its applied domains, the HTE research communities prioritize strong theoretical guarantees like convergence rates as a function of the dataset size or analytical confidence intervals. This preference has led to advances like doubly-robust plug-in estimators capable of utilizing machine learning methods [217, 218].

Somewhere in between the two is the more recent subfield of *offline reinforcement learning* (ORL) [219]. Here, the goal is to learn optimal policies from a data set containing trajectories generated from an unobserved policy. *Offline* refers to the fact that the algorithms have to exploit a batch dataset from an unknown environment without access to online exploration, which matches the common problem setup in causal inference. Despite these similarities, there are also differences between the two methodologies, which we list in Table 7.2. Nonetheless, we hope to see constructive cross-pollination between these two fields.

7.2 Causal Bandits

A bandit problem is a sequential game between a learner and an environment [220]. The game is played over $n \in \mathbb{N}$ rounds, where n is also called the *horizon*. In each round $t \in [n]$, the learner first chooses an action A_t from a given set \mathcal{A} (also called *arms*), and the environment then reveals a reward $R_t \in \mathbb{R}$.

When $|\mathcal{A}| = k \in \mathbb{N}$, we refer to the problem as *k*-armed bandits. When $k \geq 2$, but k itself is irrelevant, we simply call it *multi-armed bandits*.

Lattimore et al. [221] formalize *causal bandit problems*, a class of stochastic sequential decision problems in which rewards are given for repeated interventions on a fixed causal model. The motivation is to exploit the causal information for predict-

Approach	Input data	Goal	Typical Desiderata
Offline RL	Multi-Step Trajectories $\{(\mathbf{x}_t, \mathbf{a}_t, r_t)\}_{t=1}^T$	Policy $\hat{\pi}(\mathbf{X})$	Reward Maximization in Test Environments
HTE	One-Step Individuals $\{(\mathbf{x}_i, \mathbf{t}_i, y_i)\}_{i=1}^N$	HTE $\hat{\tau}(\mathbf{X}, \mathbf{T}, \mathbf{T}')$	Analytical Convergence Rates

Table 7.2: Differences between Offline RL and Heterogeneous Treatment Effect (HTE) estimation. Both methodologies share the motivation of extracting information for decision-making from observational data generated by unknown policies. The shown typical desiderata are simplistic, yet, we hope they clarify differences between the two schools of thought.

ing outcomes of interventions without explicitly performing them. Thereby, utilizing non-interventional (i.e. observational) data may improve the rate at which the policy learns high-reward actions. This framework generalizes classical bandits and contextual stochastic bandit problems: in the former, we have no additional observations besides a reward, and in the latter, we observe the context *before* an intervention is chosen. The causal bandit framework additionally allows us to use observations that occur *after* an intervention.

Lattimore et al. [221] describe the following scenario that motivates why using such extra observations can be helpful: Imagine a farmer who wants to maximize their crop yield. They know that crop yield is only affected by temperature, a particular soil nutrient, and moisture level, but their combination's precise effect is unknown. Each season, the farmer has enough time and money to intervene and control at least one of these variables: deploying shade or heat lamps will set the temperature to be low or high; the nutrient can be added or removed through a choice of fertilizer, and irrigation or rain-proof covers will keep the soil wet or dry. When not intervened upon, the temperature, soil, and moisture vary naturally from season to season due to weather conditions. These variations are all observed along with the final crop yield at the end of each season and can inform the farmer to conduct an experiment to identify the single, highest-yielding intervention in a limited number of seasons.

Def.: 7.2.1: Causal Bandit Problem [221]

Consider a causal model is given by a directed acyclic graph \mathcal{G} over a set of multivariate random variables $\mathcal{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_N\}$ and a joint distribution $p(\mathcal{X})$ that factorizes over \mathcal{G} . For each state $\mathbf{X} \in \mathcal{X}$, there is a reward variable Y that takes on values in $\{0, 1\}$. Further, given a set of allowed actions \mathcal{A} , we denote the expected reward for the action $a = \text{do}(\mathbf{x})$ by $\mathbf{U}_a := \mathbb{E}[Y | \text{do}(\mathbf{x})]$ and the optimal expected reward by $\mathbf{U}^* := \max_{a \in \mathcal{A}} \mathbf{U}_a$. The causal bandit game proceeds over T rounds. In round t , the learner intervenes by choosing $a_t = \text{do}(\mathbf{X}_t = \mathbf{x}_t) \in \mathcal{A}$ based on previous observations. In contrast to

the conventional bandit problem, before the agent takes the next action, it observes further samples for all non-intervened variables \mathbf{X}_t^c drawn from $p\{\mathbf{X}_t^c \mid do(\mathbf{X}_t = \mathbf{x}_t)\}$, including the reward $Y_t \in \{0, 1\}$.

Typically, the objective of the learner is to minimize the simple regret $R_T = \mathbf{U}^* - \mathbb{E}[\mathbf{u}_{\hat{a}_T^*}]$. After T observations the learner outputs an estimate of the optimal action $\hat{a}_T^* \in \mathcal{A}$ based on its prior observations. This regret objective is sometimes referred to as a *pure exploration* [222] or “best-arm identification” problem [223] and is most appropriate when, as in drug and policy testing, the learner has a fixed experimental budget after which its policy will be fixed indefinitely.

We call this particular objective the *parallel bandit* problem, in which we formalize each arm as a binary variable that is an independent cause of the reward variable. Lattimore et al. [221] propose and analyze an algorithm for achieving the optimal regret in this setting. The authors propose another, more general algorithm for general causal graphs but leave lower regret bounds for future work. Lastly, they empirically verify that both algorithms improve over the optimal successive elimination algorithm [224] through using causal knowledge of the parallel bandit problem under various conditions.

Similarly, Lee and Bareinboim [225] show that whenever the underlying causal model of a decision-making process is not taken into account, the standard strategies of simultaneously intervening on multiple variables can lead to sub-optimal policies, regardless of the number of interventions performed by the agent in the environment. Therefore, the authors suggest to reduce the search space for interventions that lead to optimal rewards and are not redundant. First, they formalize a *minimal intervention set* (MIS) and then aim at finding *possibly-optimal minimal intervention sets* (POMIS).

de Kroon et al. [226], Lu et al. [227] study causal bandit problems with unknown causal graph structures. de Kroon et al. [226] formulate a causal bandit algorithm that uses outputs of causal discovery algorithms. They utilize a separating set, defined as a set \mathbf{S} that renders a target variable Y independent of a context variable \mathbf{I} when conditioned upon, i.e., $\mathbf{I} \perp Y \mid \mathbf{S}$. The context variable encodes the intervention. Then, they separately model how interventions influence the separating set \mathbf{S} and the expected reward given \mathbf{S} . Their proposed estimator is unbiased and possesses lower variance than the sample mean.

Lu et al. [227] propose an algorithm for a class of causal graph types and prove that it achieves stronger worst-case regret guarantees than non-causal algorithms. Formally, they demonstrate that their goal of exploiting meaningful causal relations among variables cannot be achieved for general causal graphs, and in the worst case, there is no chance to do better than standard algorithms.

Similar to the causal bandit problem, Silva [228] study the related problem of dose-response learning, i.e., how an outcome variable Y (which can be a reward) varies under different levels of a control variable \mathbf{X} (which can be an arm). The difference

to causal bandits is that their goal is not to maximize Y but rather learn the relationship $f(\mathbf{x}) \equiv \mathbb{E}[Y | \text{do}(\mathbf{x})]$, $\mathbf{x} \in \mathcal{X}$, where $\mathcal{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_N\}$ is a pre-defined set of actions. Their method couples different Gaussian process priors that combine observational and interventional data. They also consider active learning schemes to choose arms informed by the GPs' uncertainties.

Similarly motivated to both the causal bandit [221], and the dose-response learning problem, Aglietti et al. [229] generalize Bayesian optimization to scenarios where causal information is available. Bayesian optimization is an efficient heuristic to optimize objective functions whose evaluations are costly and cannot be analytically described [230], e.g., the final performance of a neural network after training with a specific set of hyper-parameters. Aglietti et al. [229] argue that exploiting the causal graph noticeably increases reasoning capabilities about optimal decision-making strategies, decreasing the optimization cost and avoiding suboptimal solutions. By integrating real interventional data with estimated intervened effects computed using do-calculus, they propose an algorithm to balance two trade-offs: exploration vs. exploitation and observation vs. intervention.

Bareinboim et al. [209] take a causal perspective on the conventional multi-armed bandit problem and show that the existence of unobserved confounders render observational and interventional data distinct (recall our discussion in Sec. 2.2.1.1). Further, they show that formalizing this distinction implies that previous bandit algorithms try to maximize rewards based on the observational distribution, which is not always the best strategy to pursue. Therefore, they propose an agent loss function that incorporates both observational and interventional distributions, improving over previous ones.

Saengkyongam et al. [231] develop a causal framework for characterizing the environmental shift problem in *offline* contextual bandit problems. While the formerly discussed work has focused on exploiting causal knowledge for improving the finite sample performance or the regret bound in a single environment, their work focuses on modeling distributional shifts and the ability to generalize to new environments.

Lu et al. [232] propose *causal MDPs*, extending the idea behind causal bandits to MDPs. The motivation is similar: they use prior causal knowledge about the state transition and reward functions to obtain conditional independence relations among action, reward, and state variables and use them to develop efficient algorithms.

7.3 Model-Based RL

Model-based reinforcement learning (MBRL) learns a model of the environment dynamics (state transition and reward function) in addition to the policy, effectively combining learning and planning methods (the latter have *reversible* access to the MDP dynamics) [233]. Generative dynamics models are sometimes referred to as *world models* [234]. The enticing benefit of MBRL is the promise of improving sample efficiency compared to model-free methods by extracting valuable information from

the observed trajectories and enabling the ability to sample simulated experiences from the model instead of the actual environment.

Confounded partial models: In prior works, dynamics models are often *partial* in the sense that they are neither conditioned on, nor generate the full set of observed data. For example, the popular MuZero model [235] is a partial model, because it predicts the state observation y_T at $T > 1$ given actions a_t and updated hidden states h_t but using no observational data beyond the initial state s_0 at $t = 0$. In other words, it generates y_T directly without generating intermediate observations.

Rezende et al. [236] demonstrate that the aforementioned partial models can be causally incorrect: they are confounded by the observations $y_{<T}$ that they do not model, and can therefore lead to incorrect planning. The observations $y_{<T}$ are confounders because they are used by the policy to produce the actions $a_{<T}$, while the partial model is missing the $y_{<T}$ in its input. Therefore, the dynamics model is not robust against changes in the behavior policy.

To remedy confounding, the authors propose to use backdoor adjustment (Sec. 2.6) to make the action a_t conditionally independent of the agent state s_t . They refer to models that are conditioned on the backdoor as *Causal Partial Models* (CPM) and models that are not as *Non-Causal Partial Models* (NCPM), see Fig. 7.1. They list multiple possible choices for the backdoor and discuss their trade-offs. For future work, they suggest focusing on dynamics model robustness against other types of interventions in the environment besides policy changes.

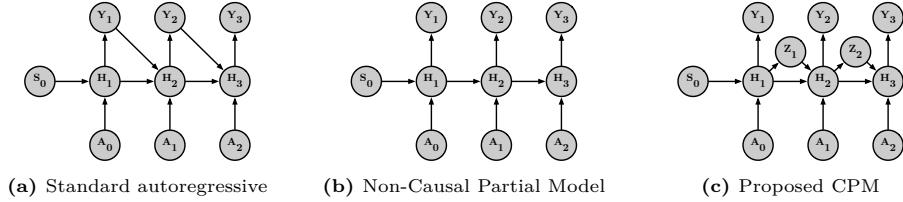


Figure 7.1: Causal Partial Models (CPM) motivation [236]: Comparison of causal graphs behind common dynamics models in Sec. 7.3

Causal World Models: Learned models of the environment dynamics are sometimes referred to as *world models*, particularly when the dynamics are observed from a sequence of high-dimensional raw pixel frames. Fundamentally, they are still used to estimate the observational conditional of the state transition function $p(s^{t+1} | s^t, a^t)$ (analogously, the reward function).

Li et al. [237] argue that such observational conditional (world) models can be biased in the real world where confounding factors exist. To make this issue apparent, they first contrast partially-observable MDPs (POMDPs) (Fig. 7.2a) used by conventional models with the causal POMDPs (Fig. 7.2b) that their *causal word model* (CWM) aims to learn from.

Given the causal POMDP, the authors define their quantity of interest as the interventional conditional $p(s^{t+1} | \text{do}(s^t = s_I), a^t)$. Then, they argue that in many

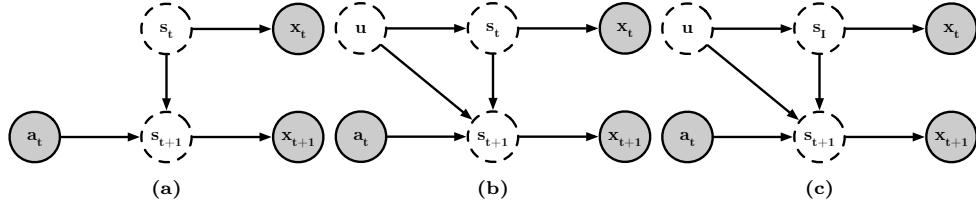


Figure 7.2: The graphical model of (a) POMDPs used by conventional world models [237], (b) causal POMDPs used by causal world models (CWMs), and (c) CWMs after intervention (do -operation).

real-world cases, the observational conditional and the interventional conditional are different because of the existence of confounding factors \mathbf{u} . The observational conditional can be written as

$$\begin{aligned} p(\mathbf{s}^{t+1} | \mathbf{s}^t, a^t) &= \int_{\mathbf{u}} p(\mathbf{s}^{t+1} | \mathbf{u}, \mathbf{s}^t, a^t) p(\mathbf{u} | \mathbf{s}^t, a^t) d\mathbf{u} \\ &= \int_{\mathbf{u}} p(\mathbf{s}^{t+1} | \mathbf{u}, \mathbf{s}^t, a^t) \frac{p(\mathbf{s}^t, a^t | \mathbf{u})}{p(\mathbf{s}^t, a^t)} p(\mathbf{u}) d\mathbf{u}, \end{aligned} \quad (7.2)$$

and the interventional conditional as

$$p(\mathbf{s}^{t+1} | \text{do}(\mathbf{s}^t), a^t) = \int_{\mathbf{u}} p(\mathbf{s}^{t+1} | \mathbf{u}, \mathbf{s}^t, a^t) p(\mathbf{u}) d\mathbf{u}. \quad (7.3)$$

The difference is highlighted in red.

Consequently, CWMs infer the interventional query “Given that we have observed $\mathbf{x}^{t:T} = \mathbf{x}^{t:T}$ in the real world, what is the probability that $\mathbf{x}^{t+1:T}$ would have been $\mathbf{x}_I^{t+1:T}$ if \mathbf{x}^t were \mathbf{x}_I^t in the dream world?” Technically, they want to intervene upon the abstract state variable as shown in Fig. 7.2c, where $s_I^t \in \mathcal{S}$ is the counterfactual value in the dream environment. This intervention is then rendered as an observable change applied to \mathbf{x}^t (such as, for instance, object displacement or removal) by the conditional observation distribution $\mathcal{U}(\mathbf{x}^t = \mathbf{x}_I^t | \text{do}(\mathbf{s}^t = s_I^t))$, where $\mathbf{x}_I^t \in \mathcal{O}$ represents the value of the counterfactual observation.

Task-Independent State Abstractions: Standard, non-causal MBRL dynamics models are *dense* in the sense that they predict the next step value of each variable based on the action and all variables in the current state. As such, they are sensitive to spurious associations.

For example, consider the example in Fig. 7.3a, where a robot faces two doors it can open and additionally observes a wall clock. Subfigure (a) shows a dense model. When door B is at angles unseen during training or the clock is at unseen times, this dense model’s prediction of door A can be inaccurate due to unnecessary dependence on the other variables. This issue has motivated state abstraction techniques [238, 239], which group many states into an abstract state by omitting some state variables, as illustrated in Subfigure (b). However, [240] argue that the generalization issues persist, as dense models are still used for the remaining variables, leaving unnecessary dependencies in the abstract states.

To this end, Wang et al. [240] introduce *Causal Dynamics Learning (CDL) for Task-Independent State Abstraction*, as illustrated in Subfigure (c). This approach learns

a causal model that explains which actions and state variables affect which variables from data. In the above example, its predictions about door A do not rely on door B and the clock and are therefore more robust to spurious associations than dense models.

More generally speaking, if there exist certain state variables which no other variables depend on (e.g., the clock), they can be omitted for planning. This motivates a novel form of state abstraction: Wang et al. [240] suggest partitioning the state variables into three groups: (i) those that the model can change with its actions (*controllable variables*, e.g., doors A and B), (ii) those that it cannot change but that still influence actions' results on those that it can (*action-relevant variables*, e.g., an obstacle that blocks door A's motion), and (iii) the remainder which can be omitted entirely (*action-irrelevant variables*, e.g., the clock).

As shown in Fig. 7.3b, their approach represents the transition dynamics through a causal graphical model, which is then split into subgraphs corresponding to the three previously described partitions. To learn the causal dynamics model, a key challenge is to determine whether a causal edge exists between two state variables, i.e., whether $s_t^i \rightarrow s_{t+1}^j$ holds. For that, they leverage Mastakouri et al. [241]'s conditional independence test, which relies on approximating conditional mutual information.

To expose the causal relationships thoroughly, one has to collect trajectories with wide coverage of the state space. For that exploratory phase, they use a reward function that is the prediction difference between the dense predictor and the causal predictor learned so far:

$$r_t = \tanh \left(\tau \cdot \sum_{j=1}^{d_S} \log \frac{\hat{p}(s_{t+1}^j | s_t, a_t)}{\hat{p}(s_{t+1}^j | \mathbf{pa}_{s^j})} \right), \quad (7.4)$$

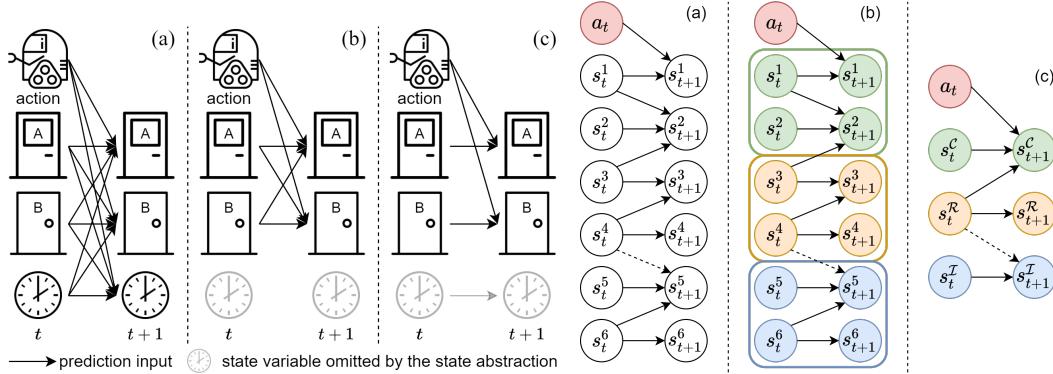
where τ is a scaling factor and \tanh bounds the reward. This reward encourages the exploratory agent to take transitions where the dense predictor is better than the causal predictor, which usually suggests the learned causal graph is inaccurate.

Lastly, to solve downstream tasks, CDL simultaneously learns a transition dynamics model (including reward function) and uses a planning algorithm depending on that model for action selection, as most MBRL algorithms do. In experiments, they verify that this improves generalization and sample efficiency on the learned dynamics models and the policies for downstream tasks.

7.4 Multi-Task RL

Multi-Task RL learning refers to settings where we expect an agent to solve multiple environments¹. Further, some works assume that the environments have never been encountered during training, but there exist some invariances across all environments (e.g., the same dynamics, but observation functions with different noise sources) or

¹We use task and environment interchangeably.



(a) Abstraction example. Imagine an environment with two doors that the robot can open and go through as well as a clock on the wall. (a): Standard MBRL dynamics model predict dynamics by unnecessarily using all variables. (b): State abstractions learn to omit the clock based on a pre-defined representation whether it exists does not affect s^5 to be ward but still use a dense model for the remaining variables. (c): CDL only keeps necessary dependencies (doors A and B depend on the action individually) and derives a state abstraction independent from any reward function.

(b) State variables can be split into three types. (a) Complete causal dynamics model. (b) Split: controllable (green), action-relevant (orange), and action-irrelevant (blue). The dashed arrow represents whether it exists does not affect s^5 to be action-irrelevant. (c) Causal graph can be split into three subgraphs, one for each type of state variable.

Figure 7.3: Causal Dynamics Learning (CDL) for Task-Independent State Abstraction [240].

there may exist an adaptation phase with limited exposure to the new task (also referred to as *meta-reinforcement-learning* [242]).

Causal Curiosity: Sontakke et al. [243] consider the setting in which the environment dynamics depend on hidden parameters [244] that differ across environments or over time. For example, if a body in an environment loses contact with the ground, the coefficient of friction between the body and the ground no longer affects the outcome of any action that is taken. Likewise, the outcome of an upward force applied by the agent to a body on the ground is unaffected by the friction coefficient.

However, in contrast to previous approaches that learn latent task representations, e.g., [91, 245], they aim to recover *disentangled* representations for the factors, i.e., independent causal mechanisms that affect the outcomes of behaviors in each environment. The motivation is that disentangled embeddings of the causal factors make the changing behaviors interpretable.

Block MDPs: Zhang et al. [246] consider the problem of learning state abstractions that generalize in block MDPs, families of environments in which the observations may change, but the latent states, dynamics, and reward function are the same. By leveraging ideas from IRM (Sec. 3.1.2.1), they propose to learn invariant state abstractions from stochastic observations across different interventions on variables in the observation space (e.g., the background color of 3D-rendered Physics simulation).

The block structure assumption states that each observation \mathbf{X} can uniquely determine its generating state \mathbf{S} . Two additional causal assumptions the authors make about the causal structure of the environment are 1) the environment state at time t can only affect the values of the state at time $t + 1$ and the reward at time t ,

and 2) each environment corresponds to an intervention on a single variable in the observation space. For example, in one of their experiments, they intervene on the background color of the environment and set it to a value sampled at random.

Model-Invariant State Abstractions: Tomar et al. [247] introduce *model-invariant state abstractions* for systematic generalization to unseen states in a single-task setting. These abstractions are built on two ideas: (1) *causal sparsity* in the transition dynamics over state variables and (2) *causal invariance* in the learned representations. (1) means that given a set of state variables, each variable only depends on a small subset of those variables in the previous timestep. (2) dictates that given a set of features, the learned representations comprise only those features that are consistently necessary for predicting the target variable of interest across different interventions. Therefore, it likely contains the true causal features and will generalize well to possible shifts in the data distribution.

Schema Networks: Kansky et al. [248] propose *Schema Networks*, a generative model for model-based reinforcement learning. Their approach relies on *Schemas*, which are local cause-effect relationships involving one or more object entities. The model’s knowledge about the class of environments is represented with schemas such that in a new environment, these cause-effect relationships are traversed to guide action selection. Thereby, experience from one scenario can be transferred to other similar scenarios that exhibit repeatable structure and sub-structure. For example, they demonstrate that Schema Networks are able to generalize to variations of the Atari Breakout [249] game with perturbed positions of objects, while model-free baselines fail to do so.

Systematic Generalization: Systematic generalization aims at learning universal (causal) relations of the environment dynamics from interactions with a few environments, such that we can deal with unseen states in a single-task setting, or approximately solve unseen other environments without further interactions in a multi-task setup.

Mutti et al. [250] define the following systematic generalization problem. First, they define a *universe*: a large, potentially infinite, set \mathbb{U} of environments modeled as discrete MDPs without rewards,

$$\mathbb{U} := \{\mathcal{M}_i = ((\mathcal{S}, d_S, n), (\mathcal{A}, d_A, n), P_i, \mu)\}_{i=1}^{\infty}. \quad (7.5)$$

The agent’s goal is to acquire sufficient knowledge to approximately solve any task that can be specified over the universe \mathbb{U} by drawing a finite amount of interactions.

A task is defined as any pairing of an MDP $\mathcal{M} \in \mathbb{U}$ and a reward function r . Solving it refers to providing a slightly sub-optimal policy via planning, i.e., without taking additional interactions.

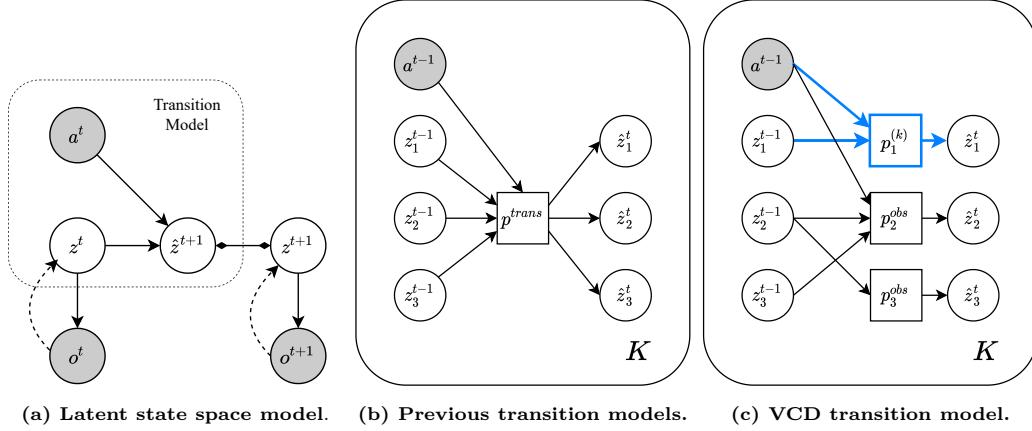


Figure 7.4: Variational Causal Dynamics (VCD) [251]. The shown DAGs illustrate one timestep and K environments. The VCD model combines causal discovery with learning transition dynamics, aiming to generalize across K environments by identifying sparse changes in the underlying causal graph. Each latent state variable z_i is represented by a separate conditional distribution p_i that is conditioned on a subset of the previous states and the action at each timestep. In blue, we highlight the intervened mechanism specific to environment k . In black, we denote mechanisms shared across environments.

Def.: 7.4.1: Systematic Generalization [250]

For any latent MDP $\mathcal{M} \in \mathbb{U}$ and any given reward function $r : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$, the systematic generalization problem requires the agent to provide a policy π , such that $V_{\mathcal{M}, r}^* - V_{\mathcal{M}, r}^\pi \leq \epsilon$ up to any desired sub-optimality $\epsilon > 0$.

Since the set \mathbb{U} is infinite, the authors posit the presence of common causal structure underlying the transition dynamics of the universe. This assumption makes the problem feasible and allows them to yield a provably efficient algorithm that achieves systematic generalization with polynomial sample complexity. They verify their algorithm's effectiveness on a synthetic universe, where each environment is a person, and the MDP represents how a series of actions the person can take influences their weight and academic performance.

Variational Causal Dynamics: Lei et al. [251] learn a causally factorized latent state-space dynamics model. They sparsity regularization from causal discovery to capture sparse causal structures describing the environment dynamics. The motivation behind their model is that it can adapt to novel environments by intervening on learned latent independent mechanisms without affecting the others.

In contrast to the other approaches discussed in this section, they parameterize the belief over the causal adjacency matrix \mathbf{G} as a random binary matrix. Each entry of that matrix follows a Bernoulli distribution with learned scalar variables. To jointly learn \mathbf{G} as well as the latent intervention masks for each environment, they utilize variational inference and derive an ELBO objective. Fig. 7.4 summarizes their approach.

7.5 Off-Policy Policy Evaluation

The purpose of policy evaluation is to measure the expected return $\mathbb{E}_{\pi}[G]$ of a target policy π , where G denote some return, e.g., the discounted return $\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$ at time step t with discount rate $\gamma \in [0, 1]$ and reward R . This evaluation uses sample trajectories $\mathcal{D} = \{\hat{\mathbf{h}}_T^i\}_{i=1,\dots,N}$ consisting of logged episodes $\hat{\mathbf{h}}_T^i = (\hat{x}_1^i, \hat{a}_1^i, \dots, \hat{a}_{T-1}^i, \hat{x}_T^i)$ generated either by the same policy π (on-policy) or by another policy \mathbf{U} (off-policy). When using the latter, we refer to the evaluation as *off-policy policy evaluation* (OPPE) and the policy used to generated the behavior is called the *behavior policy*.

In principle, OPPE is an attractive problem with many potential use cases in which online learning is not feasible, e.g., due to cost or ethical constraints on experimentation. From a technical perspective, it can guide policy learning methods to reuse off-policy experiences and find good policies more data-efficiently. However, it is generally difficult because there is a distributional mismatch between the trajectories generated by the target policy and the behavior policy. This mismatch often causes policy evaluation methods to have high variance and slow convergence [206].

Interestingly, Bannon et al. [212] emphasize that the causal inference task *counterfactual inference* (CFI) and OPPE are two different approaches to similar (under certain conditions identical) problems.

Causal Perspective: 7.5.1: Off-Policy Policy Evaluation [212]

Observations, interventions and the *query variable* in CFI directly correspond to *off-policy episodes*, the *target policy*, and the *expected return*, respectively. Similarly, Parbhoo et al. [252] argue that formulating the task of OPPE from a generalized and causal perspective opens the possibility for *counterfactual* or *retrospective* off-policy evaluation at the level of individual units (e.g. patient-level) of the population.

Observing this correspondence begs the question if the techniques developed in both fields are complementary. In the following, we discuss methods that answer in the affirmative.

Counterfactual policy evaluation: Buesing et al. [253] present *Counterfactually-Guided Policy Search* (CF-GPS), which uses counterfactual inference (Sec. 2.3.2) for off-policy evaluations (CF-PE) in the setting of *model-based (off-policy) policy evaluation* (MB-PE). In this setting, we want to evaluate the policy on synthetic data sampled from a model \mathcal{M} , i.e., we can estimate the expected return $\mathbb{E}_{\pi}[G]$ by sampling the *scenarios* $\mathbf{U} \sim p_u$ (all aspects of the environment that cannot be influenced by the agent, e.g., the initial state distribution), and then simulating a trajectory τ from the functions f_i and computing its return.

Assuming that we use an SCM \mathcal{M}_{SCM} instead of a statistical model \mathcal{M}_{ST} , Buesing et al. [253] show that CF-PE should be less biased than MB-PE. First, they

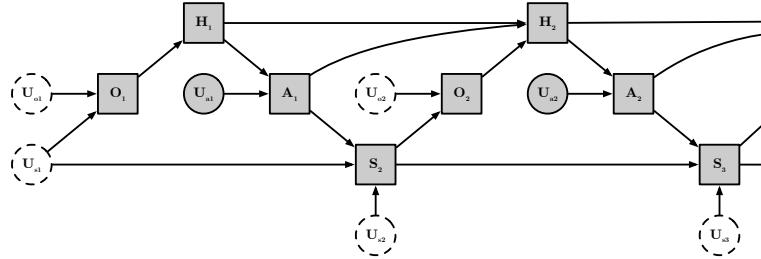


Figure 7.5: Representing a POMDP as an SCM [253]: We denote initial state $\mathbf{U}_{s_1} = \mathbf{S}_1$, states \mathbf{S}_t and histories \mathbf{H}_t . The mechanism that generates the actions A_t is the policy π . Scenarios \mathbf{U} summarize immutable aspects, some of which are observed (grey), some not (white).

show that one can represent any given partially-observable Markov decision process (POMDP) under a policy π as SCM \mathcal{M} over trajectories \mathcal{T} (for details on POMDPs, we refer the reader to [206]).

Causal Perspective: 7.5.2: POMDPs [253]

We can represent any given Partially-Observable MDP (POMDP) by an SCM \mathcal{M} over trajectories \mathcal{T} in the following way. By exploiting ideas related to the reparameterization trick commonly used in variational inference (e.g., VAEs) [104], we can express all conditional distributions, e.g. the state transitions $p(\mathbf{S}_{t+1} | \mathbf{S}_t, A_t)$, as deterministic functions with independent noise variables \mathbf{U} , such as $\mathbf{S}_{t+1} = f_{st}(\mathbf{S}_t, A_t, \mathbf{U}_{st})$. The DAG of \mathcal{M} is shown in Fig. 7.5.

Based on this representation, they then prove that counterfactual inference (Sec. 2.3.2) yields an unbiased estimator of $p^{\text{do}(I)}(\mathbf{x})$, the POMDP's distribution over trajectories after intervention I took place. In contrast, any bias in a statistical model \mathcal{M}_{ST} propagates from p^π to the estimate $\mathbb{E}_{p^\pi}[G]$.

Algorithm 2 summarizes the CF-GPS procedure. Given data $\mathcal{D} \sim p^u$ and assuming no model mismatch, i.e. $p^u = p^u$, we can regard the task of off-policy evaluation of π as a counterfactual query with data \hat{h}_T^i , intervention $I(\mathbf{U} \rightarrow \pi)$ and query variable G (reward). Then, the difference to MB-PE is that instead of sampling from the prior $\mathbf{u}^i \sim p(\mathbf{u})$, we sample scenarios from the posterior $\mathbf{u}^i \sim p^u(\mathbf{u} | \hat{h}_T^i)$ where we inferred scenarios U in hindsight from given off-policy data \hat{x}_o . Then, we evaluate the agent on these specific scenarios.

Sampling from the posteriors $N^{-1} \sum_{i=1}^N p^u(\mathbf{u} | \hat{h}_T^i)$ has access to strictly more information than the prior $p(\mathbf{u})$ by taking into account additional data \hat{h}_T^i (Sec. 2.3.2). This semi-nonparametric distribution can help to de-bias the model by effectively winnowing out parts of the domain of \mathbf{U} which do not correspond to any real data.

After demonstrating that CF-PE outperforms MB-PE in partially observed grid-world settings, they conclude that one may expect CF-PE to outperform MB-PE when the transition and reward kernels f_{st} are accurate models of the environment dynamics, but the marginal distribution over the noise sources P_U is difficult to model.

Algorithm 2 Counterfactual Policy Evaluation [253]

```

// Counterfactual inference (CFI)
1: procedure CFI(data  $\hat{\mathbf{x}}_o$ , SCM  $\mathcal{M}$ , intervention  $I$ , query  $\mathbf{X}_q$ )
2:    $\hat{\mathbf{u}} \sim p(\mathbf{u}|\hat{\mathbf{x}}_o)$                                       $\triangleright$  Sample noise variables from posterior
3:    $p(\mathbf{u}) \leftarrow \delta(\mathbf{u} - \hat{\mathbf{u}})$                           $\triangleright$  Replace noise distribution in  $p$  with  $\hat{\mathbf{u}}$ 
4:    $f_i \leftarrow f_i^I$                                           $\triangleright$  Perform intervention  $I$ 
5:   return  $\mathbf{x}_q \sim p^{\text{do}(I)}(\mathbf{x}_q|\hat{\mathbf{u}})$             $\triangleright$  Simulate from the resulting model  $\mathcal{M}_{\hat{\mathbf{x}}_o}^I$ 
6: end procedure

// Counterfactual Policy Evaluation (CF-PE)
7: procedure CF-PE(SCM  $\mathcal{M}$ , policy  $\pi$ , replay buffer  $\mathcal{D}$ , number of samples  $N$ )
8:   for  $i \in \{1, \dots, N\}$  do
9:      $\hat{\mathbf{h}}_T^i \sim D$                                           $\triangleright$  Sample from the replay buffer
10:     $g_i = \text{CFI}(\hat{\mathbf{h}}_T^i, \mathcal{M}, I(\mathbf{U} \rightarrow \pi), G)$            $\triangleright$  Counterfactual evaluation of return
11:   end for
12:   return  $\frac{1}{N} \sum_{i=1}^N g_i$                                  $\triangleright$  Return averaged counterfactual return
13: end procedure

```

Following a similar methodology, [254] use counterfactually-generated trajectories to highlight episodes where the target and behavior policy returns differ substantially. They interpret this as a helpful procedure for off-policy “debugging” in high-risk settings, such as healthcare. In contrast to [253], who use counterfactuals to approximate draws from the interventional distribution, they treat the counterfactual distribution as the primary object of interest and demonstrate their method’s utility in a sepsis management simulation.

7.5.1 Unobserved Confounding

In OPPE, the caveat of not using the current policy for evaluation is that we almost inevitably face unobserved confounders (Sec. 2.5) that causally affect the behavior policy. Kallus and Zhou [255] illustrate the problem with the following clinical example: Suppose we want to compare the efficacy of different drugs. During normal clinical practice, we observe the outcomes (rewards) of those prescribed the drug. A drug may appear less clinically effective if those prescribed it were more unhealthy to begin with and, therefore, would have had less successful outcomes regardless. Conversely, if the drug was given only to patients who would benefit most, it could be mistakenly regarded as beneficial to everyone. While these issues can potentially be addressed by controlling for more factors that may have affected treatment decisions, they can never be entirely eliminated. Healthcare databases are often incomplete about medical histories, patient severity notes, etc., so they are especially susceptible to unobserved confounding.

To this end, Kallus and Zhou [255] study the setup of *confounding-robust policy improvement* (including OPPE) to account for possible unobserved confounding (UC). They develop a method for minimizing the worst-case estimated regret of a candidate policy against a baseline policy over a set of propensity weights that control the extent of UC. As a result of their theoretical analysis, they obtain generalization

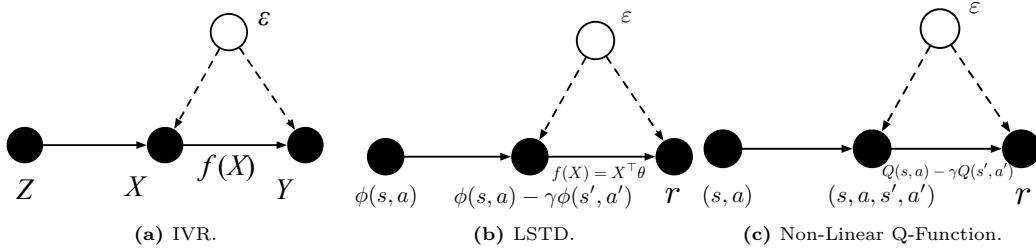


Figure 7.6: Causal DAGs revealing the relationship between Instrumental Variable Regression (IVR), Least-Squared Temporal Difference (LSTD) and non-linear Q-functions [265].

guarantees that ensure that their policy will be safe when put into practice. Further, it will yield the best possible uniform control on the range of all possible population regrets consistent with UC's extent.

Infinite-Horizon: Kallus and Zhou [256] and Bennett et al. [257] consider the above setup extended to the infinite-horizon setting, as e.g., commonly considered in continuous control of physical systems [258] or quantitative trading [259]. Similarly, Namkoong et al. [260] study the case when UC occurs only at a single time step, as when an expert makes an initial decision based on unrecorded information and then follows a set of protocols based on well-recorded observations.

POMDPs: Tennenholz et al. [261] consider OPPE in partially-observable MDP (POMDP) environments, where the unobserved variables may have confounding effects, motivating them to propose the *decoupled POMDP* model, which is a class of POMDPs for which observed and unobserved variables are distinctly partitioned. Bennett and Kallus [262] utilize the framework of proximal causal inference to reveal POMDP settings where identification of the target policy value is possible. Further, they construct semi-parametrically efficient estimators for these settings.

Combining offline and online data: Gasse et al. [263], Wang et al. [264] focus on using offline data to warm start online RL. Gasse et al. [263] suggests learning a latent-based transition model that explains both the interventional and observational regimes and then inferring the standard POMDP transition model. Wang et al. [264] propose *confounded MDPs*, which naturally captures both the offline and online setting as well as their mismatch due to confounding. They then construct deconfounding algorithms in the episodic setting with linear function approximation.

Relationship to instrumental variable regression: Many OPPE rely on estimates of the state-action value (Q-) function by minimizing the mean squared Bellman error. Chen et al. [265] reveal that this strategy leads to a regression problem with an unobserved confounder between the inputs and the output noise: Fig. 7.6 displays that the causal relationships between instrumental variable regression, the least-squared temporal difference (linear Q-function) and non-linear Q-functions are equivalent.

7.6 Imitation Learning

Imitation learning (IL) aims to learn control policies directly from examples of demonstrations provided by human experts [266]. The motivation behind it is to remove the need for extensive interaction with the environment during policy learning and/or designing reward functions specific to the task. Cameras and sensors of today collect and transmit vast amounts of data quickly, and processors with significant computing power are getting quicker in mapping the sensory inputs to actions. Hence, developing real-time perception and reaction models assisted through IL opens up many potential applications, such as humanoid robots, self-driving vehicles, or human-computer interaction systems.

One challenge in IL is confounding. In the following, we consider confounded settings in which it is difficult for the imitator to recover the expert's performance. At their core, these issues stem from similar roots as the ones we have seen in the subsection on OPPE (Sec. 7.5), where one aims at learning policies based on trajectories that were not controlled by the agent itself but an external entity. However, as the goal of IL is different from OPPE, different remedies have been developed.

By and large, *causal* imitation learning (CIL) aims at solving confounding issues. The following works aim at deconfounding observed expert trajectories such that imitator policies achieve similar performances or define necessary and testable conditions for such.

Causal Confusion: de Haan et al. [267] study the setup of *causal confusion*: in which the inputs to the expert policy are fully observed; however, the mechanisms of the expert policy are latent, i.e., it is not known which of the observed input variables are actual causes of the expert's actions and which are not (i.e., "nuisance variables"). Here, the expert and imitator observe the same contexts, but the causal diagram is not available to the imitator.

The authors show that in this setting, the phenomenon of *causal misidentification* (CM) can occur: if there is a shift between inputs causing the expert's actions and the imitator's actions (e.g., when the latter relies on nuisance variables), then access to more data can yield worse performance. In other words, CM happens when cloned policies fail by misidentifying the true causes of expert actions, and training them on more observations can exacerbate the performance.

Consider Fig. 7.7 as an example: we aim to train a neural network to drive a car. The model input for scenario A is an image of the dashboard and windshield. The input to the model (with identical architecture) is the same image, but the dashboard is masked out. Although both cloned policies achieve low training loss, model B performs well when tested on the road, while model A does not. The dashboard has an indicator light that comes on immediately when the brake is applied, and model A wrongly learns to apply the brake only when the light is on. Even though the brake light is an effect of braking, model A could achieve low training error by misidentifying it as a cause.

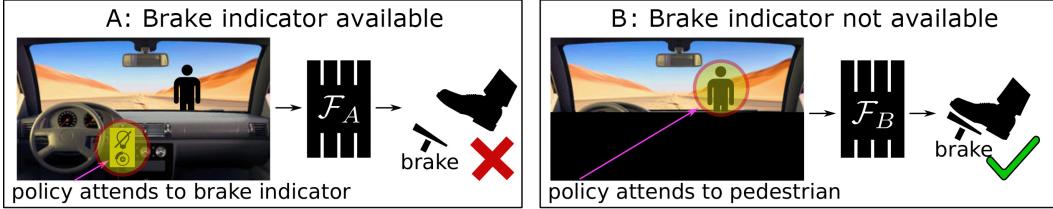


Figure 7.7: Causal misidentification: more data can yield worse imitation learning performance [267]. There exist a spurious association between braking and the car’s brake indicator. A naive behavior-cloned model (\mathcal{F}_A) relying on the braking indicator will apply the brake only when the brake light is on. This model works well in the left Scenario A; however, if the braking indicator is unavailable (Scenario B, right), the imitator will fail. A better model (\mathcal{F}_B) decides whether to brake by attending to the pedestrian.

To remedy these issues, the authors suggest finding the true causal model of the expert’s actions. A two-stage pipeline can automate this: in stage 1, we jointly learn policies corresponding to various causal graphs; in stage 2, we perform targeted interventions to efficiently search the hypothesis set for the correct causal model.

Tien et al. [268] provide a systematic study of causal confusion in the context of learning reward functions from human inputs that take the form of pairwise preferences or rankings. Specifically, they demonstrate across three different robot learning tasks that causal confusion over the true reward function occurs, even with large numbers of pairwise preferences over trajectories. The authors investigate different factors affecting causal confusion: type of training data, reward model capacity, and the preference training data generation mechanism. One of their conclusions is that while human demonstrations seeking to be pedagogic lead to better sample efficiency, all the preference data collection methods were prone to causal confusion.

Self-Delusion: Ortega et al. [269] demonstrate that the common perception that popular sequence models “do not understand the cause and effect of their actions” is the result of conditioning on actions rather than treating them as causal interventions. The reason is that the model update differs depending on whether the collected data originated from within the model (i.e., from actions) or outside of it (i.e., from a third person whose policy we do not know), and mixing them up results in incorrect inferences.

To give intuition, the authors give a minimal example, as illustrated in Fig. 7.8. Consider a prize-or-frog problem, where there are two boxes (1&2), one containing a prize and the other a frog, respectively. The objective is to open the box containing the prize. For simplicity, let us assume that the two configurations are equiprobable. The true DGP has a joint distribution $p(\theta, A, O)$ over the box configuration θ , the chosen box A (1 or 2), and the observed content O (± 1 reward).

Suppose we learn a probabilistic model from data generated by an expert who opens the correct box when told where the prize is. Crucially, unlike the expert, we would not observe θ . The source of the delusion can be seen by comparing the two posterior beliefs $p(\theta | a, o)$ and $p(\theta | \text{do}(a), o)$ over the task parameter, that is, where the past action is treated as a condition and an intervention, respectively (differences are

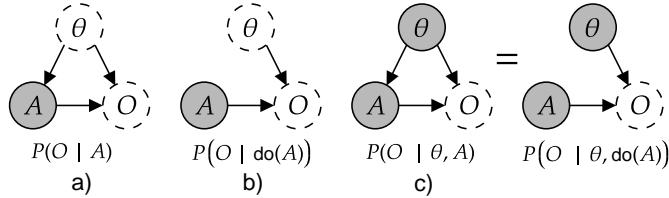


Figure 7.8: The Self-Delusion problem illustrated through a prize-or-frog game [269]: Panel (a) illustrates the self-delusion problem: conditioning on the self-generated action leads to wrong inferences about O , since A and O are confounded by θ . Panel (b) shows that treating the self-generated action as a causal intervention circumvents the self-delusion because no information can flow backward from A into θ . Panel (c) corresponds to the fully observable case. When θ is observed, conditioning or intervening on the self-generated action A leads to the same prediction over O .

highlighted):

$$p(\theta | a, o) = \frac{p(\theta) p(a | \theta) p(o | \theta, a)}{\sum_{\theta'} p(\theta') p(a | \theta') p(o | \theta', a)} \quad (7.6)$$

$$p(\theta | \text{do}(a), o) = \frac{p(\theta) p(o | \theta, a)}{\sum_{\theta'} p(\theta') p(o | \theta', a)}. \quad (7.7)$$

Here, we can see that the intervention causes the dismissal of the evidence $p(a | \theta)$ produced by the self-generated action $A = a$.

Unobserved confounding: Zhang et al. [270] tackle the setting in which a causal graph is provided, but *unobserved* confounders (UC) affect both actions and outcomes of the expert demonstrations. This means that the expert's input observation may differ from the ones available to the imitator. For instance, self-driving cars rely solely on cameras or lidar, ignoring the auditory dimension. However, most human demonstrators can exploit this data, especially in dangerous situations (car horns, screeching tires) [271].

To formalize and study this problem in single-stage settings, they introduce *partially observable SCMs*, a particular type of SCMs that explicitly allows one to model the unobserved nature of some endogenous variables. Then, they show that the problem of imitability is orthogonal to identifiability and introduce a data-dependent, graphical criterion for determining whether imitating an expert's performance is feasible. Based on this criterion, they provide an algorithm that checks whether there are instruments in the data, permitting it to satisfy the graphical imitability criterion. Finally, for the case of the criterion being met, they propose a procedure for estimating an imitating policy.

Sequential data: Kumor et al. [271] extend Zhang et al. [270]'s results to sequential settings, where the imitator must make multiple decisions per episode. To this end, they introduce a generalized backdoor criterion that allows one to learn imitating policies across a sequence of states and actions. They prove that their criterion is *necessary* for imitability.

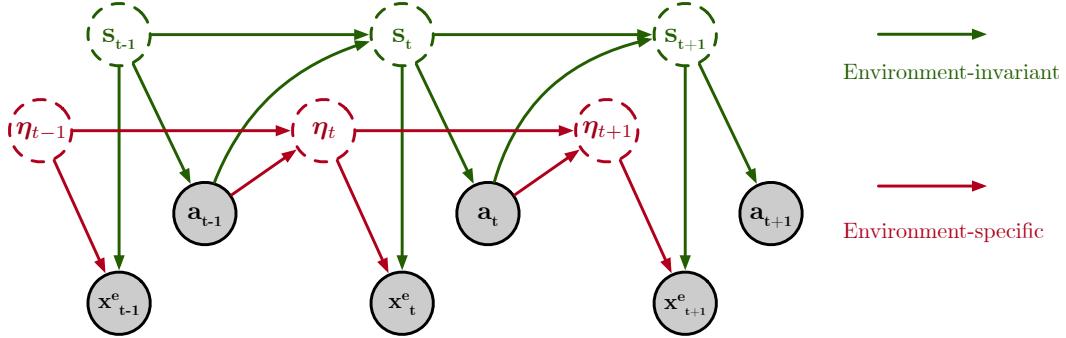


Figure 7.9: Problem setup of Invariant Causal Imitation Learning [272]. The observations are decomposable into environment-invariant features s_t and environment-specific η_t that encapsulate spurious associations with the actions. The goal is to recover s_t such that the learned policy $\pi(\cdot | s_t)$ generalizes well to new environments.

Multiple Environments: Bica et al. [272] consider the imitation learning setting with multiple environments. Their goal is to learn a policy that matches the expert behavior in all possible environments from a class of environments that share a certain structure for the observations and the transition dynamics, similarly, as we have discussed in Sec. 7.4. Fig. 7.9 visualizes the causal graph of their setup.

To incentivize the learned imitation policy to stay within the distribution of the expert’s observations, the authors propose to minimize the energy of the next observation obtained by following π given the current observation. This way, they assign a low loss to the imitation policy for staying within high-density areas of the expert’s occupancy measure and a high loss for straying from it.

Temporally Correlated Noise: Swamy et al. [273] hypothesize that in many imitation learning settings, recordings of the expert may be corrupted by an unobserved confounder that is temporally correlated, i.e., *temporally correlated noise* that is not independently distributed across timesteps. Table 7.3 illustrates settings where such noise can occur. The first row, for example, considers a quadcopter pilot who might have been flying under the persistent wind. An imitation learner will likely reproduce these deviations and perform poorly in test-time environments with different weather conditions. This setup is related to Sec. 7.6 but makes stronger assumptions on the unobserved confounder (additive noise).

To this end, Swamy et al. [273] leverage instrumental variable regression (IVR) to deconfound the data (Sec. 11.2.1.2). Put simply, IVR leverages an *instrument*, a source of random variation independent of the confounder, to deconfound inputs to a model via conditioning on the instrument.

The key idea behind their approach is to frame past states as instruments to break the spurious association between states and actions caused by the unobserved confounder. The motivation behind this is that historical transitions are unaffected by future confounding [274].

They formalize the TCN setup by an SCM and the target estimand as the interventional effect of the expert policy $\mathbb{E}[\pi_E(s) | s] = \mathbb{E}[a | \text{do}(s)] \neq \mathbb{E}[a | s]$ based on

Setting	State	Expert Action	Observed Action	Confounder
Quadcopter Flying [275]	Position	Intended Heading	Actual Heading	Persistent Wind
Product Pricing [276]	Demand	Profit Margin	Price	Raw Materials Cost
ICU Treatment [277]	Symptoms	Intent to Treat	Patient Treated	Comorbidity
Shared Autonomy [273]	User State	Intended Action	Executed Action	Assistance

Table 7.3: Examples of Temporally Correlated Noise in Causal Imitation Learning [273]: In these settings, spurious associations (Sec. 2.5) between states and actions can lead to inconsistent estimates of the expert’s policy.

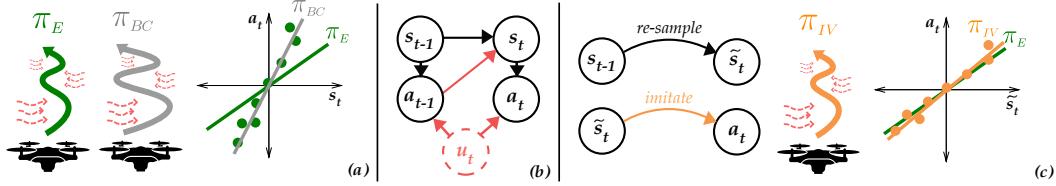


Figure 7.10: Causal Imitation Learning under Temporally Correlated Noise (TCN) [273]: (a) Unobserved confounders (e.g. wind) induce spurious associations between states and expert actions. Non-causal imitation learning approaches like behavioral cloning can learn policies that rely on this noise, leading to poor test-time performance. (b) Causal DAG illustrating how the TCN confounder u_t affects both the input (s_t) and output (a_t). (c) Using instrumental variables Sec. 11.2.1.2, We can re-sample state transitions from a past state, producing fresh samples (\tilde{s}_t). We can then regress from these sampled states to observed expert actions to recover the expert’s policy as the noise on inputs and outputs is no longer correlated.

IVR. To estimate that estimand, they propose two approaches inspired by generative modeling and game-theoretic IVR approaches. The former utilizes access to a simulator, while the latter can be run entirely offline. Further, they derive performance bounds, which assume that the same distribution of TCN affects the learner at test time.

7.7 Credit Assignment

One key issue in RL is *credit assignment* (CA), the task of understanding the causal relationship between actions and rewards and determining to what extent an outcome is due to external, uncontrollable factors [278, 279]. That means that a correct value function allows us to disentangle the relative aspects of “skill” and “luck” in an agent’s performance. However, unfortunately, because of partial observability, scale, long time horizons, or increasing number of actions, each action taken by an agent may have only a vanishing effect on the outcome, making it increasingly difficult to learn proper value functions.

In the following, we look at methods that assign credits by measuring the causal influence of an action on the observed reward (Sec. 2.7).

7.7.1 Causal influence detection

Seitzer et al. [280] postulate that an agent can only influence its environment in certain situations. If, for instance, a robotic arm is placed in front of an object of interest on a table, the object can only be moved once contact is made between the robot and object. Some situations have an immediate causal effect, while others do not.

The authors argue that *causal action influence* (CAI), which quantifies the causal influence of states, can help guide the learning algorithm to seek out states with stronger (predicted) influence, even enabling it to discover helpful behavior in the absence of task-specific rewards. To derive CAI, they introduce a causal model, which allows them to quantify whether the agent is in control given its state by conditional mutual information (CMI), which they learn through neural network models. Based on their experimental results, they conclude that CAI improves sample efficiency and performance by (i) better state exploration through an exploration bonus, (ii) causal action exploration, and (iii) prioritizing experiences with causal influence during training.

7.7.2 Counterfactual credit assignment

Mesnard et al. [279] propose *counterfactual credit assignment* (CCA), a framework that uses the notion of counterfactuals to deal with the credit assignment issue. They propose to implicitly perform counterfactual policy evaluation (Sec. 7.5) by conditioning value functions on future event embeddings, which learn to extract relevant information from a trajectory. They then use the estimated counterfactual returns to form unbiased and lower variance estimates of the policy gradient by building future-conditional critics.

While conventional state-action functions estimate the return for all actions, they do so by averaging all possible futures. In contrast, CCA estimates the return for different actions while keeping many external factors constant between the return and the counterfactual estimate. This makes CCA potentially finer-grained and improves data efficiency in environments with complex credit assignment structures.

The critical ingredient for the FC-PG algorithm is to learn the hindsight statistics Φ , which embed a trajectory while not precluding any action a , which was possible for π from having potentially produced Φ . For example, $\Phi_t = A_t$ does *not* satisfy this condition, because it precludes any action $a \neq A_t$ from having produced Φ_t . The authors discuss several options to learn Φ and mainly focus on using a hindsight network φ which extracts Φ directly from the observed state-action-reward triplets, i.e., $\Phi_t = \varphi(\mathbf{X}, A, R)$.

By establishing a connection to counterfactually-guided policy search [253], introduced in Sec. 7.5, the authors explain that the CCA estimator can be understood as sidestepping modeling the environment by discarding information from the whole

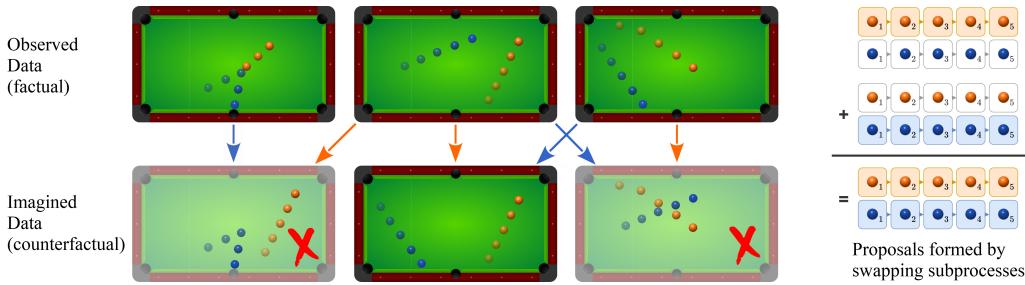


Figure 7.11: Pool example motivating Counterfactual Data Augmentation (CoDA) [286]. Knowing the local causal structure allows us to mix and match factored subprocesses to form counterfactual samples based on three factual samples. We reject the first proposal because one of its factual sources (the blue ball) is not localized. Due to the swapped proposal not being locally factored in, the third proposal is rejected. Accepted proposal two may be used as additional training data for reinforcement learning agents.

trajectory τ which would require a model, and left with still useful information Φ_t with which counterfactuals can be computed in a model-free way.

7.7.3 Multi-Agent Reinforcement Learning

Multi-agent systems consist of autonomous, distributed agents that interact with each other in a shared environment [281, 282]. Each agent strives to accomplish an assigned goal, and the interplay between agents can vary depending on the task, such that agents cooperate or act competitively to beat the competition.

Here, the credit assignment problem arises in cooperative settings when the collection of the agents’ actions generates only a global, shared reward, making it difficult for each agent to deduce its contribution to the team’s success [283, 284].

Additionally to the shared rewards given to all agents, [285] proposes that agents be rewarded for influencing the actions of other agents. Agents receive this intrinsic reward in addition to their immediate reward, reflecting the effect one agent has on another. They refer to it as “causal influence”, similarly as Seitzer et al. [280] in the single-agent setup discussed in Sec. 7.7.1. Using counterfactual reasoning, an agent simulates possible actions and determines the effect they would have had on another agent’s behavior at every time step. Actions that significantly affect the other agent’s behavior are considered highly influential and rewarded. Generally, this added reward helps agents learn interpretable communication protocols and obtain higher collective rewards than baselines, which sometimes fail to learn completely.

7.8 Counterfactual Data Augmentation

The following techniques apply the idea of counterfactual data augmentation (CFDA) (Chapter 4) to RL trajectories.

Local Causal Models [286]: Consider a game of billiards, as illustrated in Fig. 7.11: each ball can be seen as its own physical process. Before the opening break, due to

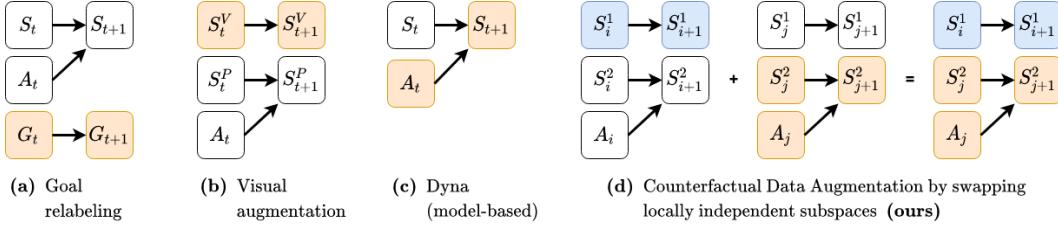


Figure 7.12: Existing RL data augmentation techniques can be interpreted as specific instances of CoDA [286]: Orange nodes are relabeled, exogenous noise variables omitted for clarity. (a) Goal relabeling [287], including HER [288], augments transitions with counterfactual goals. (b) Visual feature augmentation [289, 290] changes visual features S_t^V (e.g., lighting, camera positions, etc.) that do not impact the physical state S_{t+1}^P . (c) Dyna [291], including MBPO [292], augments observed states with new actions and re-samples the next state using a learned dynamics model. (d) Given two transitions that share local causal structures, CoDA swaps connected components to form new transitions.

their initial placement, each ball has a non-zero chance of colliding with the others. Hence, to predict the expected outcome of the opening break, we require a transition dynamics model that considers all balls. However, besides the initial timestep, most interactions between balls remain sparse. In other words, only a small subset of all balls are involved at most timesteps.

Pitis et al. [286] take advantage of the fact that during the time between their interactions (*locally*), the subprocesses are causally independent: they propose a counterfactual data augmentation technique that is compatible with any agent architecture and does not require a forward dynamics model. By inferring whether or not local interactions occur, they swap factorized subspaces of observed trajectory pairs whenever two trajectories have the same local factorization.

The key idea behind CoDA is to leverage the principle of independent mechanisms (Def. 2.3.4). Assume that we can decompose state and action space into multiple subspaces, e.g., represented through subgraphs $\mathcal{G}_i, \mathcal{G}_j \subset \mathcal{G}$, where \mathcal{G} is the causal DAG of the global transition dynamics. Then, the causal mechanisms represented by $\mathcal{G}_i, \mathcal{G}_j$ are independent when \mathcal{G}_i and \mathcal{G}_j are disconnected in \mathcal{G} . In other words, when the global dynamics governed by \mathcal{G} are divisible into two (or more) connected components, we can reason about each subgraph as an independent causal mechanism. Existing RL data augmentation techniques rely on similar global independence relationships, as illustrated in Fig. 7.12.

Personalized Policies [293]: Lu et al. [293] propose a CFDA method to tackle environment heterogeneity by *personalized policies*, e.g., for healthcare settings, where patients may exhibit different responses to identical treatments. Similar to CF-GPS (Sec. 7.5), they formalize the transition dynamics process as an SCM. Then, similar to methods in Sec. 7.3, they learn that SCM including its structural functions and exogenous noise variables (Sec. 2.3), they use *Bidirectional Conditional GANs* [294].

The authors propose two algorithms, differing in whether we assume latent environment heterogeneity in the data or not. If so, they explicitly include an environment variable in the causal system to consider variability across observations.

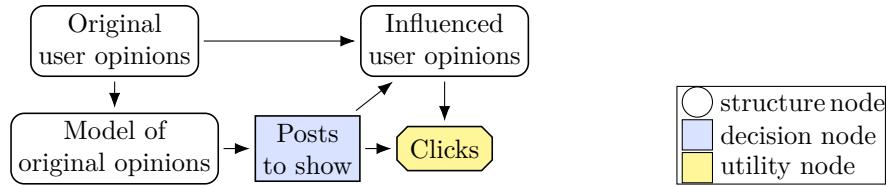


Figure 7.13: Example of a Causal Influence Diagram (CID) in the context of a content recommendation system [301]. The goal of a content recommendation system is to choose posts that will maximize the user's click rate. However, the system's designers prefer the system not to manipulate the user's opinions to obtain more clicks.

7.9 Agent Incentives

The development of artificial general intelligence (AGI), systems that equal or exceed human intelligence in a wide variety of cognitive tasks, raises serious safety concerns [295, 296]. The key concern underlying AGI safety research is that these systems may produce autonomous agents much more intelligent than humans, which consequently pursue goals that conflict with our own. One way to take precautions against potential catastrophes caused by AGIs is to *align* their *incentives* with ours, also referred to as *AI alignment* [297, 298].

Many agent incentive concepts rely on causal relationships among variables influencing the agent’s utility. For example, Miller et al. [299] postulate that classifiers designed to incentivize improvement *necessarily* require to perform causal inference. In other words, the authors show that any type of agent that strategically adapts performs causal modeling “in disguise”.

In the following, we follow Everitt et al. [300] and summarize recent progress on formalizing agent incentives causally.

7.9.1 Causal Influence Diagrams

One formalism for studying agent incentives is a *causal influence diagram* (CID) [301, 302, 303], a causal DAG with decision (\square) and utility (\diamond) variables. After defining a CID, one can analyze incentives through graphical criteria, e.g., the *value of information* (VoI) [304], which measures which variables a decision-maker would benefit from knowing before making a decision.

Fig. 7.13 shows an example: Imagine a content recommendation engine recommending a series of posts (decision node) to a user. For the algorithm designers to maximize the number of clicks (utility node), the content should be tailored to the user's interest (structure node). They do not, however, want the algorithm to use provocative content to manipulate the user's opinion (structure node). Evans and Kasirzadeh [305] refers to this problem as *user tampering*. Using a simulation study, the researcher demonstrates that a Q-learning algorithm learns how to exploit its opportunities to polarize users with its early recommendations to have more consistent success with later recommendations that cater to that polarization.

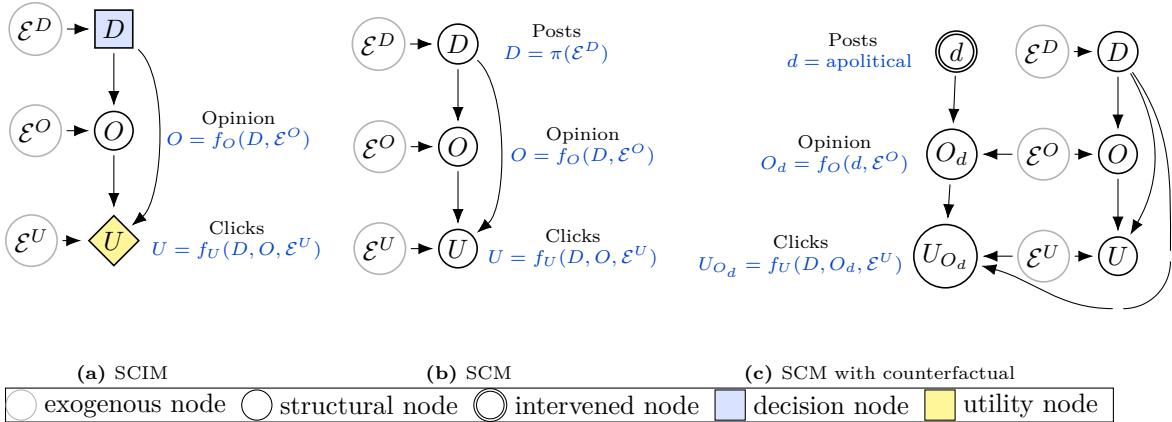


Figure 7.14: An example of a Structural Causal Influence Model (SCIM) in a content recommendation scenario [301]. The system aims to select posts that maximize click-through rates without manipulating user opinions. (a) Displaying political or apolitical posts D will affect the user's opinion O . D and O influence the user's clicks U . (b) Given a policy, the SCIM becomes an SCM. (c) We can use this SCM to estimate counterfactuals. For example, the counterfactual U_{O_d} represents the number of clicks if the user has the opinions they would arrive at after viewing apolitical content.

More generally, we define a CID as follows.

Def.: 7.9.1: Causal Influence Diagram [301, 302, 303]

A *causal influence diagram* (CID) is a directed acyclic graph \mathcal{G} with vertices \mathcal{V} partitioned into $\mathcal{V} = \{\mathcal{X}, \mathcal{D}, \mathcal{U}\}$ with *structure nodes* \mathcal{X} , *decision nodes* \mathcal{D} and *utility nodes* \mathcal{U} . Utility nodes have no children. Edges into decisions indicate what information is available at the time of the decision, so we call them *information links*.

Next, to define incentive concepts, Everitt et al. [301] use a *structural causal influence model* (SCIM) [306], which is similar to an SCM (Def. 2.3.1), but not equivalent. There are three key differences: (i) its structural functions are only defined for non-decision endogenous variables, (ii) the vertices are partitioned as described in Def. 7.9.1, and (iii) it holds that the utility variable domains are a subset of the real numbers, i.e., $\mathcal{U} \subset \mathbb{R}$.

Returning to the content recommendation example, Fig. 7.14 illustrates its SCIM. By specifying a policy π , which comprises a structural function for decision nodes \mathcal{D} , we turn a SCIM into an SCM, allowing us to use the standard definitions of causal interventions and counterfactuals.

Being equipped with CIDs and SCIMs, we can now analyze agent incentives and other safety-related concerns.

7.9.2 Incentive Concepts

We summarize five incentive concepts informally and direct readers interested in more details to Everitt et al. [301].

Materiality [307]: : Which observations are *material* for optimal performance? Nodes can be identified as immaterial based on the graphical structure, e.g., if they are *d-separated* (i.e., conditionally independent) from the utility nodes.

Value of Information [304]: Generalization of materiality to unobserved nodes to assess which variables a decision-maker would benefit from knowing *before* making a decision.

Response Incentive [301]: An alternative generalization of generalizing a material observation from the perspective of it being one that influences optimal decisions. Under this interpretation, variables that have a response incentive are the ones that influence the optimal agent's decisions.

Value of Control [308]: Can the agent benefit from setting a variable's value? More specifically, can the attainable utility be increased by letting the agent decide the structural function for the variable?

Instrumental Control Incentive [301]: What is the agent interested in and able to control? Would that choice matter if the agent chooses decision $D \in \mathcal{D}$ to influence $X \in \mathcal{X}$ independently of how D influences other aspects of the environment? Everitt et al. [301] call it an *instrumental* control incentive because the control of X is a means to achieve utility.

7.9.3 User Interventions

Sometimes it is desirable to constrain or modify the policies of agents when they are deployed in practice. For example, action constraints have been used to prevent damage when training robot policies, and transformations to the policy can ensure that an agent stays within some safe region of state space [309, 310]. Whenever it is difficult to specify unsafe states formally, a human overseer might interrupt the agent [300].

Langlois and Everitt [311] address this issue by introducing *modified-action MDPs* (MAMDPs), where the agent's policy outputs the agent's decision rather than the actions. The actions are assumed to be not fully under the agent's control but can also be influenced by the action modification. Since the agent might learn the interruption scheme from interruptions during training, the corresponding CID includes an information link from the action modification to the policy.

The MAMDP framework, including its CIDs, enables the authors to analyze the causal assumptions that prototypical RL algorithms make about the environment

(a) Rocks and Diamonds Gridworld.

(b) Rocks and diamonds with a modifiable RF.

Figure 7.15: Gridworld Example of Reward Tampering [312]. The agent can avoid getting the implemented reward function updated and even change it.

and to understand how they adapt to interruption via a graphical criterion for path-specific response incentives.

They find that black-box optimization algorithms such as evolutionary strategies take all causal relationships of the environment into account and try to both obscure its policy and disable its off-switch. In contrast, with the action chosen by the agent, Q-learning and SARSA assume that the next action will be taken optimally, with no action modification. In other words, these two algorithms ignore causal effects from the action modification to the actual action taken. Another version of SARSA, based on the modified action, ignores the effect of action modification on the current action but considers the effect on subsequent actions and tries to disable the off-switch.

7.9.4 Reward Tampering

By default, an agent aims to maximize its observed reward, given by a (programmed) reward function. However, sometimes, it may be incentivized to tamper with this function, which we call *reward tampering* [312]. Everitt et al. [300] give three examples:

1. **Wireheading:** Rewriting the source code of its implemented reward function.
2. **Feedback tampering:** Influencing users that train a learned reward model.
3. **Input-tampering:** Manipulating the reward function's inputs.

Fig. 7.15 showcases an example of an MDP with a modifiable implemented reward function. Imagine a gridworld where the agent can push rocks and diamonds by walking towards them from an adjacent cell, as illustrated in Fig. 7.15a. At time t , the reward function rewards diamonds and punishes rocks for each item pushed to the goal area:

$$R_t = \# \text{ diamonds in goal area} - \# \text{ rocks in goal area}. \quad (7.8)$$

Now, let us distinguish between *intended rewards* that encourage completion of the intended task and *observed rewards*, which are the rewards received by the agent. Generally speaking, we parameterize the returning reward function as $R_t =$

$R(S_t; \Theta^R)$, where Θ_t^R will denote the parameter for an implemented reward function at time t , and Θ_*^R the parameter of an intended reward function.

To model the possibility of the agent influencing the intended function, we include a user and two reward parameters Θ_{rock}^R and $\Theta_{\text{diamond}}^R$. The new, modifiable reward function is

$$R_t = \Theta_{\text{diamond},t}^R \cdot (\#\text{diamonds in goal area}) + \Theta_{\text{rock},t}^R \cdot (\#\text{rocks in goal area}), \quad (7.9)$$

where the reward parameters toggle between -1 and $+1$ if the agent stands on top of them. Further, the reward parameters are set to their intended value of diamond-gathering ($\Theta_{\text{rock}}^R = -1$ and $\Theta_{\text{diamond}}^R = 1$) when the agent visits the user tile.

The intended agent's task is to gather diamonds. However, in its initial implementation, Θ_1^R rewards rocks instead of diamonds, which is corrected when the agent passes the user. According to the intended reward function, the agent should not walk around the user or visit the reward parameter tile. Unfortunately, by breaking either of these conditions, the agent can observe more rewards.

Everitt et al. [312] claim that a standard RL agent may have an instrumental goal to tamper with its implemented reward function. By modeling the above problems with CIDs and analyzing their previously proposed solutions, the authors find that the latter avoid undesirable incentives by removing causal links in a way that avoids instrumental control incentives. However, there are still differences between the proposed methods, e.g., whether they use previously observed or latent variables to formulate an objective avoiding reward tampering.

7.9.5 Delicate States

A *delicate* state is as part of the environment which is hard to define a reward for (*subtle*) and vulnerable to deliberate manipulation towards bad outcomes (*manipulable*). Farquhar et al. [313] try to remove any incentive for the agent to control the delicate part of the state-space (as having a control incentive over it is dangerous).

To remove it, they declare a *stable* state condition, which is robust against unmotivated action and renders side-effects unlikely to be bad. They formalize these conditions through CIDs' graphical criteria for the presence of incentives to control certain parts of the state.

7.10 Open Problems

Lack of unified evaluation environments: For example, in causal model-based RL, all methods share the same goal of learning the environment transition dynamics (e.g., to improve sample efficiency for policy search). However, there is no overlap in their evaluation environments. Rezende et al. [236] evaluate on AVOIDFUZZYBEAR (proposed in the paper), MINIPACMAN [314], and visual 3D [315] environments. Wang et al. [240] use a Chemical and manipulation environment. Li et al. [237]

conduct their experiments on the CoPHY benchmark [316] (Sec. 9.2). Similar inconsistent evaluation procedures occur in causal multi-task RL.

Unification of Formalisms: Due to the parallels between causality and reinforcement learning and the lack of maturity at the intersection of both, we fear redundancy among papers, i.e., an analogy of *reinventing the wheel* [317] taking place. An increasing number of formalisms in the literature may lead to decision paralysis for practitioners and, similarly as we have argued in the previous paragraph, slows down progress by increasing barriers to comparing approaches. We suggest that future work compares the applicability of different formalisms more rigorously and possibly unifies them.

For example, consider a set of MDPs with shared dynamics (e.g., the set of all physically feasible pendulums), where each single MDP is specified by additional dynamics parameters (e.g., the length and mass of a pendulum, see [91] for more concreteness). Let us assume that these parameters are unobserved.

In the literature, we notice a myriad of MDP formalisms proposed for that type of setup: MDP with Unobserved Confounders [318], Confounded MDPs [319], Causal POMDPs 1 [237], Causal POMDPs 2 [243] (they are not the same formulation), Causal MDPs [232], and discrete MDPs (a *universe*, Sec. 7.4) [250]. Even in the non-causal RL literature, we find multiple formalisms with similar motivations, e.g., Contextual MDPs [320], Hidden Parameter MDPs [244], and Bayes Adaptive MDPs [321]. We understand that these formalisms may not be technically identical but hypothesize that unification is possible and can facilitate progress.

Deconfounding Offline RL: Whenever we aim to learn policies from an offline dataset without access to observed policy or the environment, this dataset likely includes confounding biases. As we have seen throughout this section, a significant motivation for using causality for RL is to deconfound observed data, i.e., transform it so that confounding biases are removed. We believe that deconfounding observational data is under-explored in pure offline RL settings, and only a few works have addressed this setup [256].

Counterfactual Decision-Making: Bareinboim [213] expose that inference of counterfactuals is neglected in the RL literature, possibly due to the difficulty of learning correct SCMs. However, counterfactual reasoning in agents may bring additional benefits to agents only reasoning based on observational distributions, such as less bias during policy search (Sec. 7.5), assigning credit to single agents in a shared-reward cooperative multi-agent system [284] or considering the intended actions of humans in a human-in-the-loop system [322].

Bottou et al. [208] illustrate how counterfactual reasoning can assist large-scale real-world machine learning systems, such as the ad placement system associated with the Bing search engine. They propose a causal model for ad placement and apply it to optimizing ad auction tuning parameters. This work is very detailed and discusses multiple fundamental causal inference techniques, connections to RL, and different

design choices. We hope to see more carefully-designed counterfactual agents for real-world systems that may take inspiration from Bottou et al. [208].

8

Modality-specific Applications

In the previous chapters, we familiarized ourselves with different methodologies that applied two causal primitives, interventions and counterfactuals, throughout 5 different problem domains; ranging from how to learn invariances in data up to dealing with confounding in sequential decision making settings.

In this chapter, we review methods designed for particular data modalities, namely: image (computer vision), text (natural language processing), and graph (graph representation learning) data. Some of these works are off-the-shelf applications of the core methodology introduced in the previous chapters; some are not but are too modality-specific to be presented in the earlier chapters.

We observe a few common themes across all three areas, which we use to divide the methods within each modality. For convenience, we recap and briefly summarize them here:

- **Causal Supervised Learning** (Chapter 3): Methods disposing of spurious associations by extracting (invariant) feature maps containing only the causal parents of the predictor variable Y .
- **Counterfactual Data Augmentation** (Chapter 4): Methods augmenting training data with counterfactual modifications to a subset of the causal factors, leaving the rest of the factors untouched.
- **Counterfactual Explanations** (Sec. 5.2.1): Methods explaining model predictions by computing a (minimally) altered features instantiation of an individual that would cause the underlying model to classify it in a different class.
- **Causal Fairness** (Chapter 6): Methods ensuring that model predictions are fair w.r.t. causal relationships involving protected attributes (e.g., gender, race, etc.).
- **Miscellaneous**: Everything else that does not fit into the above categories.

8.1 Causal Computer Vision

8.1.1 Causal Supervised Learning

Long-Tail Classification: Stochastic gradient descent (SGD) methods are central to neural network optimization [323]. Modern SGD variants (e.g., Adam, SGD-M, etc.) commonly used for image classification often involve some form of *momentum*

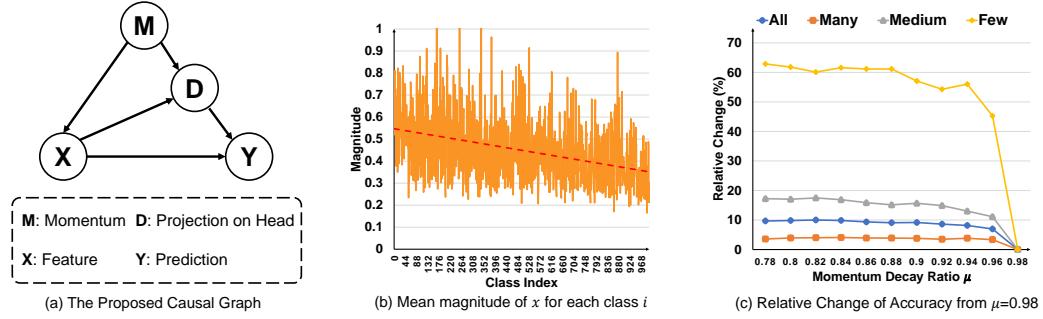


Figure 8.1: SGD Momentum as Confounder [324]. (a) Causal DAG explaining the causal effect of momentum. (b) Sorted mean magnitudes of feature vectors for each class i after training with momentum $\beta = 0.9$. (c) The relative accuracy change from $\text{beta} = 0.98$ as a function of β indicates that the few-shot tail is more sensitive to it.

(or *acceleration*), which accumulates historical gradients to speed up convergence, akin to a heavy ball rolling down the loss function landscape.

Tang et al. [324] interpret momentum \mathbf{M} as a confounder in long-tailed classification settings. If a dataset is balanced, each class contributes equally to the momentum. Long-tailed datasets, however, will be dominated by *head* samples (i.e., samples belonging to the majority class). Long-tailed datasets are very common in practice, e.g., in NLP due to Zipf’s law [325] or in image segmentation, where increasing the images of tail class instances like “remote controller” requires more head instances like “sofa” or “TV” simultaneously.

Let us consider input features \mathbf{X} , momentum \mathbf{M} , classification logits Y , and \mathbf{D} , which denotes \mathbf{X} ’s projection on the head feature direction that eventually deviates \mathbf{X} . In a long-tailed dataset, few head classes possess most training samples, which have less variance than the data-poor but the class-rich tail. Therefore, the moving averaged momentum will point to a stable head direction. Specifically, the authors demonstrate that one can decompose any random feature vector \mathbf{X} into $\mathbf{X} = \ddot{\mathbf{X}} + \mathbf{D}$, where \mathbf{D} is a function of the exponential moving average features and the number of training iterations (for brevity, not shown here).

The authors show that the following causal relationships emerge: \mathbf{M} confounds the input features \mathbf{X} ($\mathbf{M} \rightarrow \mathbf{X}$) and the classification logits Y (via $\mathbf{M} \rightarrow \mathbf{D} \rightarrow Y$), as illustrated in Fig. 8.1(a). The backdoor path $\mathbf{X} \leftarrow \mathbf{M} \rightarrow \mathbf{D} \rightarrow Y$ causes a spurious association between \mathbf{X} and Y . The authors call this association a “bad” bias. Fig. 8.1(b,c) demonstrate that momentum contributes to this bias empirically. Further, the mediation path $\mathbf{X} \rightarrow \mathbf{D} \rightarrow Y$ is considered a “good” bias, because it respects the inter-relationships of the semantic concepts in classification.

To remedy the “bad” momentum bias, Tang et al. [324] apply backdoor adjustment to account for the backdoor confounding path, while keeping the “good” mediation path. Further, for the final prediction logits, they compute the intervention $p(y | \mathbf{x})$ to yield the direct causal effect of $\mathbf{X} \rightarrow Y$. Their method improves over baselines in long-tailed image classification and instance segmentation benchmarks.

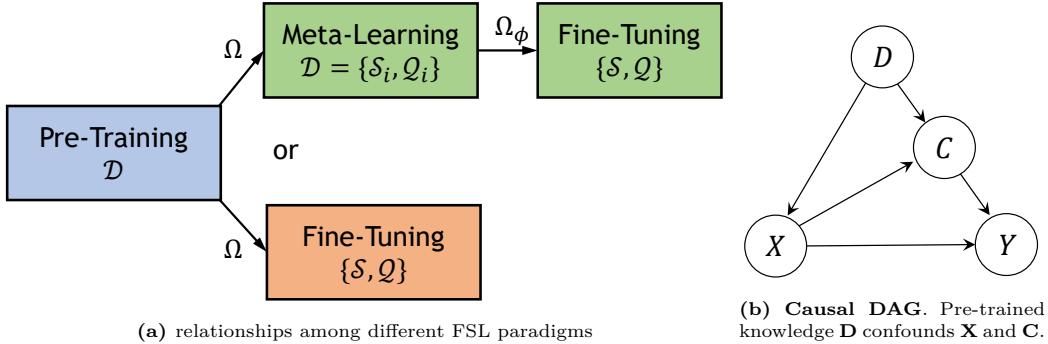


Figure 8.2: Confounding (Sec. 2.5) in Few-Shot-Learning [328]: Pre-trained weights, obtained through conventional pre-training or meta-learning, can be interpreted as a confounder that may introduce transfer deficiencies (Sec. 8.1.1).

Few-Shot Learning: Few-shot learning (FSL) methods assume that deployed models in production will likely encounter novel tasks for which only a few samples with labels are available; yet, these data-poor tasks may have some structural similarity to other data-rich tasks [326, 327]. Murphy [327] give the following example: consider the task of classifying endangered bird species, which are by definition rare. However, birds bear many structural similarities across species (wings, feathers, etc.). Hence, training a model on a large dataset of non-endangered species and then transferring that knowledge to the small datasets of endangered birds may result in better performance than training on the small datasets alone.

A common strategy to deal with such settings is *transfer learning*. A simple transfer learning strategy is *fine-tuning*, which consists of two phases: First, we perform a *pre-training* phase, in which we train a model on a large source dataset. Second, we freeze some pre-trained parameters and continue training the rest on the few-shot target (training) dataset of interest.

Another, more advanced transfer learning strategy is *meta-learning*, which aims to learn a meta-model that quickly adapts to different few-shot datasets. We will not discuss how these methods work but direct interested readers to [326].

Yue et al. [328] study both FSL strategies from a causal perspective and find that pre-trained knowledge is a confounder that can limit their performance. By adjusting for this confounder, the authors develop three algorithms that achieve new state-of-the-art results on several FSL benchmarks.

For illustration, recall the exemplary problem in Fig. 3.1, where we are interested in predicting the cow label based on cow-specific features and not the background features (e.g., the mountains). Using pre-trained knowledge **D** (e.g., a large dataset \mathcal{D} or a third-party pre-trained model with parameters θ), $p(y | \mathbf{x})$ may fail to generalize well, because it induces spurious associations: the pre-trained weights generate features ($\mathbf{D} \rightarrow \mathbf{X}$), and semantics ($\mathbf{D} \rightarrow \mathbf{C} \rightarrow Y$) that may overly rely on mountain specifics.

The authors argue that to make FSL more robust, we need to pursue the true causality between **X** and **Y**, i.e., the causal intervention $p(y | \text{do}(\mathbf{x}))$. To illustrate

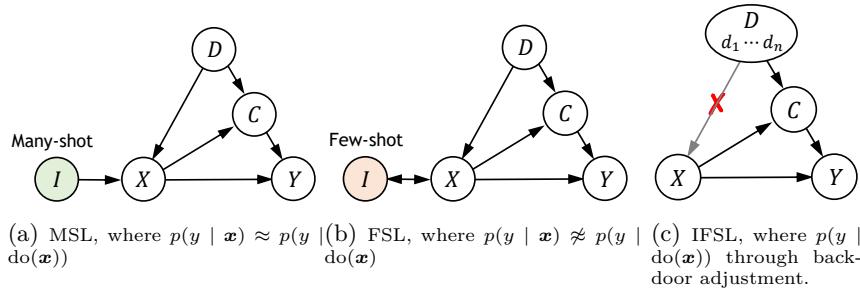


Figure 8.3: **Many-shot (MSL) vs. few-shot (FSL) vs. interventional few-shot learning (IFSL)** [328]. MSL and IFSL are robust against confounding of \mathbf{D} , while FSL is not.

why this is, they contrast FSL with many-shot learning (MSL), which refers to fine-tuning with a larger target dataset. Naturally, we would expect MSL to work better simply because we use more data; yet, the authors argue that this would not answer why MSL converges to the true causal effects as the number of samples approach infinity. Causal Perspective 8.1.1 explains this in more detail.

Causal Perspective: 8.1.1: Many-Shot vs. Few-Shot Learning [328]

Yue et al. [328] postulate that $p(y | \text{do}(\mathbf{x})) \approx p(y | \mathbf{x})$ in MSL while $p(y | \text{do}(\mathbf{x})) \not\approx p(y | \mathbf{x})$ in FSL. To show why this holds, let us introduce the sample ID I and assume that $\mathbf{x} \sim p(\mathbf{x} | i)$. Naturally, we may assume that we can use $p(y | i)$ to estimate $p(y | \mathbf{x})$, so we can incorporate I into the estimation of $p(y | \mathbf{x})$, by writing

$$p(y | \mathbf{x}_i) := \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|i)} [Y | \mathbf{x}, i] = p(y | i). \quad (8.1)$$

Now, we compare how I enters the causal graphs for MSL and FSL, illustrated in Figs. 8.3a and 8.3b. For MSL, we find that $I \rightarrow \mathbf{X}$, and not $\mathbf{X} \rightarrow I$, because tracing \mathbf{X} 's ID out of many samples is like “finding a needle in a haystack”. This makes I an instrumental variable (Sec. 11.2.1.2), effectively meaning that I and \mathbf{D} are independent and $p(y | \mathbf{x}) = p(y | i) \approx p(y | \text{do}(\mathbf{x}))$ (details in [328]’s Appendix). However, $\mathbf{X} \rightarrow I$ persists in FSL, because it is easier to guess the correspondance, e.g., in the 1-shot extreme case that has a trivial 1:1 mapping for $\mathbf{X} \leftrightarrow I$.

Next, the authors propose *Interventional FSL*, where the idea is to use backdoor adjustment to estimate $p(y | \text{do}(\mathbf{x}))$ without the need for many-shot samples, as illustrated in Fig. 8.3c. This adjustment requires observing and stratifying the confounding variable, which is non-trivial when \mathbf{D} is a third-party delivered pre-trained network. The authors suggest three implementations for this: (i) feature-wise adjustment, (ii) class-wise adjustment, and (iii) combined adjustment. They show that these implementations improve the baselines across all query hardnesses.

Attention Models: To make visual attention-based models less susceptible to capture spurious correlations and more robust in OOD settings, Wang et al. [319]

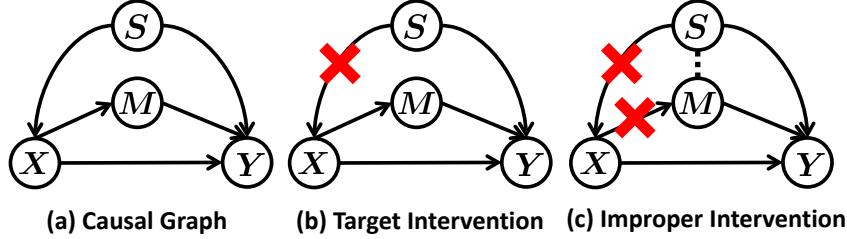


Figure 8.4: Causal DAG for Causal Attention Module [319], consisting of image \mathbf{X} , label Y , mediator \mathbf{M} and unstable context confounder \mathbf{S} . The goal is to learn the causal association $\mathbf{X} \rightarrow Y$, while disposing associations along the pathways of $\mathbf{X} \leftarrow \mathbf{S} \rightarrow Y$.

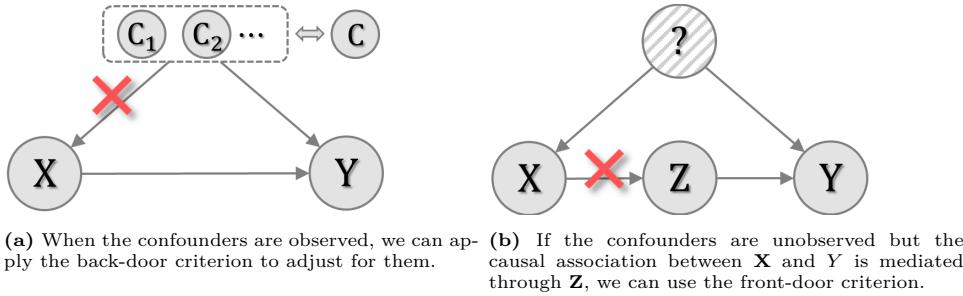


Figure 8.5: Confounder Identification-free Causal Feature Learning [329].

propose a *causal attention module* that annotates confounders in an unsupervised way. Fig. 8.4 shows their causal graph: $\mathbf{X} \rightarrow Y$ is the desired causal effect from image \mathbf{X} to label Y . Further, we assume an invariant mediator variable \mathbf{M} that contains discriminative object parts, e.g., a bird’s wing. \mathbf{S} is the style confounder, e.g., the non-discriminative background of an image (sky when the bird flies, ground when its wings are down). The target intervention is to disentangle the background \mathbf{S} and the mediator \mathbf{M} , as shown in Fig. 8.4b). However, since a perfect intervention is typically not easy to obtain when working with visual datasets, the authors propose to perform an improper intervention, illustrated in Fig. 8.4c).

To identify the content and style variables, they train two separate attention mechanisms $f_a(\cdot)$, $f_{\bar{a}}(\cdot)$ in an adversarial fashion, respectively. Similar to training Generative Adversarial Networks [105], each iteration of the training pipeline corresponds to solving a bilevel optimization problem consisting of a minimization and maximization step. The minimization step optimizes the invariant feature extractor $f_a(\mathbf{x}) = \mathbf{c}$, while the maximization step updates the style confounder extractor $f_{\bar{a}}(\mathbf{x}) = \mathbf{s}$.

Unobserved Confounders: Most previous approaches adopt the backdoor criterion to mitigate the effect of confounders. However, the backdoor criterion requires the explicit identification of confounders. Since it can be challenging to identify confounders in many real-world scenarios, Li et al. [329] explore a confounder identification-free method using the front-door criterion. Fig. 8.5 illustrates the difference between using the backdoor and front-door criterion.

The front-door criterion does not require identifying the confounders if we have access to an intermediate variable \mathbf{Z} , such that $\mathbf{X} \rightarrow \mathbf{Z} \rightarrow Y$. The authors propose a strategy to simulate interventions on \mathbf{Z} , relying on meta-gradients. They also connect this strategy to gradient-based meta-learning methods and explain why methods like MAML [330] work from a causal perspective. Overall, their method improves the cross-domain performance of vision models.

Motion Forecasting: Liu et al. [331] study motion forecasting, the task of forecasting the location of a tracked object from a video. Fig. 8.6a illustrates their causal DAG: they treat each video as a domain $e \in \mathcal{E}$, and include style \mathbf{S} , content \mathbf{C} , as well as *domain invariant variables* \mathbf{Z} , in contrast to \mathbf{S} and \mathbf{C} which have dependence on the domain $e \in \mathcal{E}$. For example, \mathbf{Z} may capture laws of physics.

Fig. 8.6b illustrates their model architecture: an invariant encoder $\phi(\cdot)$ models domain invariant relationships, the style encoder models $\psi(\cdot)$ models domain-specific relationships, and the style modulator $f(\cdot)$ extracts useful domain-specific information while ignoring spurious information. A contrastive loss $\mathcal{L}_{\text{style}}$ is employed to encourage learning useful representations of domain specific information.

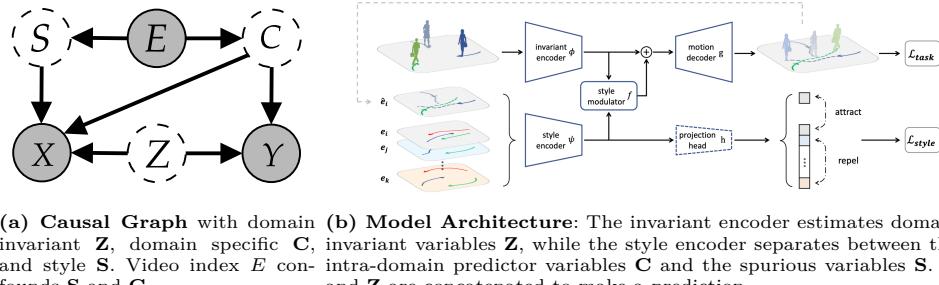


Figure 8.6: Causal Motion Forecasting [331]: An approach for forecasting the location of an object over time that generalizes across environments.

Weakly-Supervised Semantic Segmentation: Zhang et al. [332] present a framework called *Context Adjustment* (CONTA) to deconfound pixel-level pseudo-masks in weakly-supervised semantic segmentation (WSSS). Specifically, the goal is to obtain non-spurious pixel-level pseudo-masks \mathbf{M} of image segments Y . The authors identify that three types of style features \mathbf{S}^1 lead to inaccurate pseudo-masks:

- Object ambiguity:** Objects often co-occur with each other under certain contexts. For example, if most horse images contain people riding horses, there exists a spurious association between these two distinct objects. A model relying on it will learn that most horses are with people, outputting pseudo-masks that do not draw boundaries between a person and a horse.
- Incomplete background:** Backgrounds are often composed of multiple (unlabeled) semantic objects. Hence, there is a co-occurrence of foreground, and back-

¹The authors refer to this variable as “context” in their work, but it is semantically equivalent to our concept of “style” (Def. 3.1.1) that we use consistently across methods.

ground objects, e.g., some parts of the background “floor” can be misclassified as the foreground “sofa”.

3. **Incomplete foreground:** Semantic parts of the foreground object may co-vary with different contexts, e.g., the window of a car may include reflections of its surroundings. Hence, a model may learn to segment less discriminative parts to represent the foreground, e.g., the “wheel” of a “car” (instead of the whole car).

Therefore, they propose an SCM capturing the causal relationships between images, styles, and class labels, as illustrated in Fig. 8.7. Their deconfounding method, named *context adjustment*, removes the confounding bias in semantic segmentation by simulating an intervention $p(y | \text{do}(\mathbf{x}))$ using an approximation of the unobserved \mathbf{S} .

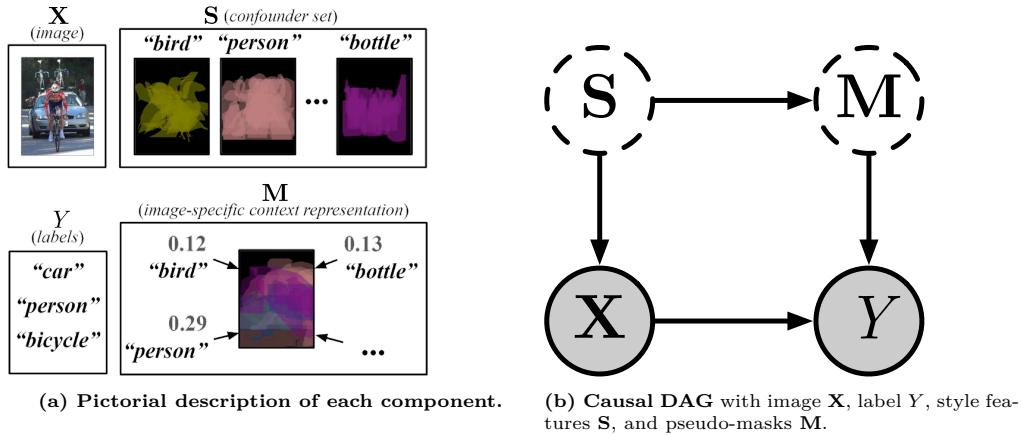


Figure 8.7: Context Adjustment (CONTA) for Weakly-Supervised Semantic Segmentation [332]. The goal is to deconfound pixel-level pseudo-masks.

Visual Grounding: Visual grounding is the task of mapping a free-form natural language query (phrase or sentence) onto its corresponding region of the image, e.g., “the dog next to the person who is driving a car” [333]. By investigating failure cases of existing grounding methods, Huang et al. [333] reveal spurious associations between certain subjects and their locations in common grounding datasets. For example, pictures corresponding to queries including “sheep” are often located in the central area; “corner of” tend to be of smaller size; “standing” tend to display standing people in the center because they are commonly the focus of the photographer; however, if encounter images where most people stand aside, this spurious association does not hold anymore.

Fig. 8.8 shows the causal graph Huang et al. [333] propose. The causal estimand of interest is the interventional distribution $p(l | \text{do}(r), \mathbf{x}) \neq p(l | r, \mathbf{x})$, where \mathbf{X} denote the image, R the language query, and L the object location. The inequality is due to the unobserved confounder \mathbf{G} . To deal with this unobserved confounder, the authors exploit the previously proposed *Deconfounder* algorithm by Wang and Blei [334]. This algorithm allows one to learn a generative model of the substitute confounder $\hat{\mathbf{G}}$, which Huang et al. [333] use to perform the backdoor adjustment. We note that

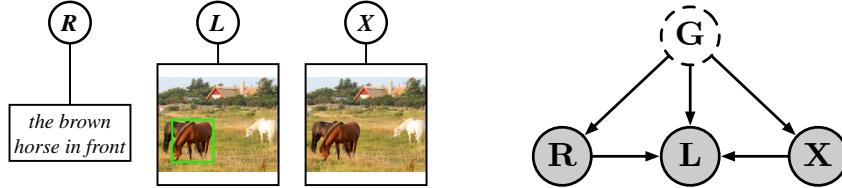


Figure 8.8: Deconfounded Visual Grounding [333]: \mathbf{G} : unobserved confounder, \mathbf{X} : pixel-level image, \mathbf{R} : language query, \mathbf{L} : the location for the query.

this algorithm has caused some controversy, and Ogburn et al. [335], D’Amour [336] questioned its legitimacy.

Video moment retrieval: The task of identifying the start and end of a segment in a video that corresponds to a textual query is known as *video moment retrieval* (VMR). Otani et al. [337] argue that VMR models often exploit spurious temporal location biases in datasets rather than learning the cross-modal matching. In other words, the temporal location of moments is a hidden confounder that spuriously correlates user queries and moment locations, making the model ignore the actual video content.

Nan et al. [338] propose to remove such spurious correlations from the video grounding model by considering interventions on the text and video input. They call this method *Interventional Video Grounding (IVG)*, as it uses backdoor adjustment to model $p(y | \text{do}(\mathbf{x}))$. They then propose a method to approximate the unobserved confounder as one cannot sample it directly from the dataset. They learn representations of the video and text inputs and enforce alignment between these representations of similar concepts with a contrastive approach.

Similarly and concurrently, Yang et al. [339] construct an SCM consisting of four variables: \mathbf{Q} (query), \mathbf{V} (video moment), Y (prediction), and \mathbf{L} (moment location). Then, they replace the non-causal query $p(y | \mathbf{q}, \mathbf{v})$ with $p(y | \text{do}(\mathbf{q}, \mathbf{v}))$. Thereby, the query is forced to fairly interact with all possible locations of the target based on the intervention.

8.1.2 Counterfactual Explanations

Goyal et al. [340] produce counterfactual explanations Sec. 5.2.1) to scrutinize the prediction of image classification models. Given a *query* image \mathbf{X} with class label y , the goal of the counterfactual visual explanation is to identify how to change the image such that an image classification model would predict a different class y' .

Their methods work as follows: First, one has to select a *distractor* image \mathbf{X}' that the model predicts as the class y' . Then, they identify spatial regions in the images \mathbf{X} and \mathbf{X}' , replacing the spatial region in \mathbf{X} with the area in \mathbf{X}' would lead to predicting the class y' . Beyond providing explanations for model prediction, the authors find that counterfactual visual explanations can help human users distinguish different classes better.

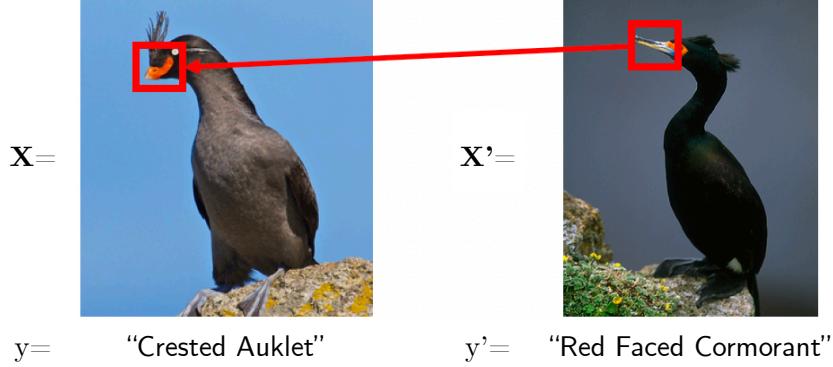


Figure 8.9: Counterfactual visual explanations (CVE) [340]: Consider two images of two similar looking yet different birds: \mathbf{X}, \mathbf{X}' is the query and distractor image with label y, y' , respectively. CVE identify regions in both images such that if the highlighted region in \mathbf{X} looked like the highlighted region in \mathbf{X}' , the resulting image \mathbf{X}^* would be classified more confidently as y' .

Hendricks et al. [341] propose to generate counterfactual explanations by inspecting which evidence an input data point misses but might contribute to a different classification decision if present in the image. The motivation is to generate post-hoc natural language explanations about what attributes might change classification decisions if present in an image, e.g., “This is not a Scarlet Tanager because it does *not* have black wings.”

8.1.3 Causal Generative Modeling

8.1.3.1 Counterfactual Data Augmentation

Zero-Shot Learning: Yue et al. [342] suggest to generate counterfactual samples for out-of-distribution generalization of classifiers, by means of a generative model which disentangles class attributes Y and sample-specific attributes Z of \mathbf{X} . Then, they generate counterfactual samples from a counterfactual distribution where Z has been intervened upon while keeping Y constant, $\tilde{x} \sim p(x | y, \text{do}(\tilde{z}))$.

To ensure that \tilde{x} lies in the true distribution of the seen or unseen samples, they enforce counterfactual faithfulness by applying the contrapositive of the Consistency Rule [343]: if x is dissimilar to \tilde{x} , the ground-truth attribute of x cannot be y . We can implement this criterion by training a binary classifier that distinguishes between seen and unseen data \mathbf{X} .

Cross-Domain Pose Estimator: Zhang et al. [344] propose using causal representation learning to improve cross-domain 3D pose estimation tasks. Specifically, they train a counterfactual feature generator that takes domains and contents as input. They change domains to simulate interventions and steer the model to produce counterfactual features. This strategy facilitates the model to learn transferable features across domains.

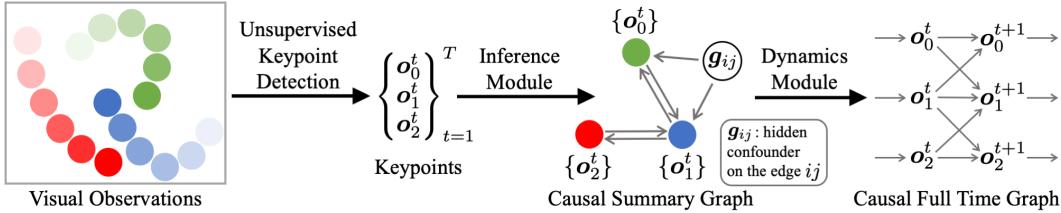


Figure 8.10: Visual Causal Discovery Network [345]: it extracts unsupervised keypoints as the state representation, infers a causal graph, and then learns a dynamics module conditioned on both.

8.1.3.2 Counterfactual Trajectory Generation

Our goal is to generate a counterfactual sequence $\mathbf{x}_{1:T}^{\text{CF}}$ under a new initial condition \mathbf{x}_0^{CF} , given an observed initial condition \mathbf{x}_0 and a sequence of T frames $\mathbf{x}_{1:T}$.

Li et al. [345] propose the *Visual Causal Discovery Network* (V-CDN), which is akin to tackling all three levels of causal representation learning (Def. 2.4.1). It consists of three modules: (i) a perception network employing an unsupervised keypoint detection algorithm [346] to extract useful features from a video; (ii) a structural inference module using a GNN to learn a causal graph; and (iii) a dynamics prediction model conditioned on the causal graph and the current state features to make predictions into the future. The model tackles a task that purely associational object learners fail on: modeling interactions between objects beyond collisions only.

Following up on Li et al. [345], Janny et al. [347] complement learning latent representations based on the unsupervised discovery of keypoints with additional information. Specifically, this addresses two problems: First, one must encode the shape, geometry, and relationships between multiple moving objects through the relative positions of points because each object is only discriminated through its 2D positions. Second, a 2D keypoint space may not be the optimal representation for modeling the dynamics of a physical system because of the additional imaging process that may confound the data.

The authors deploy their proposed architecture on video data from a 3D simulator of stacked blocks. An encoder learns a representation of the blocks in the form of 2D keypoints and information coefficients which encode for shape and appearance. Latent confounders are extracted from this representation, and a dynamic model forecasts a trajectory from a representation of an observation. A decoder subsequently maps the trajectory prediction into video data. The authors also introduce a novel benchmark for counterfactual trajectory prediction, which is discussed in Sec. 9.2.

8.2 Causal Natural Language Processing

Despite these challenges, text has been proven to be useful in causal inference applications, because text can serve as confounder [348], outcome [349], and treat-

ment [350, 351] in causal inference. Causal inference has been applied in a variety of natural language processing (NLP) tasks such as natural language explanations [352, 353], fairness [354], text classification [355], and machine translation [356]. Next we start introducing some representative work.

8.2.1 Causal Supervised Learning

8.2.1.1 Visual Question Answering

Imagine working at a social network company, and your team's task is to build models detecting hate speech. Colleagues of yours have already developed great models for text data only; however, a remaining challenge is detecting it in memes, where the model needs to understand the combined meaning of the words and pictures.



Figure 8.11: Multimodal “mean” memes [357]. *Left:* Mean memes, *Middle:* Benign Image Confounders, *Right:* Benign Text Confounders.

This problem falls into the category of *Visual Question Answering* (VQA), i.e., the task of answering previously unseen questions framed in natural language about a previously unseen image (e.g., “Does this image include hate speech?”). The classical approach for VQA is to learn a model from a training set made up of image \mathbf{v} , question \mathbf{q} and answer a triplets $\mathcal{D} = \{\langle \mathbf{q}_i, \mathbf{v}_i, a_i \rangle\}_{i=1}^n$. The model infers an embedding of the questions $\mathbf{e}^q = f_q(\mathbf{q})$, an embedding of the image $\mathbf{e}^v = f_v(\mathbf{v})$ and a fusion function of the two $\mathbf{z} = h(\mathbf{e}^q, \mathbf{e}^v)$ into what is known as the *joint space*.

One challenge for such models is to deal with spurious associations between the image and text modalities. For instance, Fig. 8.11(*Left*) illustrates common mean memes; memes with text like “love the way you smell today” might be spuriously associated with images of unpleasant smells. These spurious associations make it harder for models to truly capture multimodal understanding, as shown by Kiela et al. [357]. The authors develop a benchmark with *benign confounders*, which are minimum replacement images/texts that flip the labels for a given multimodal meme from hateful to non-hateful. Benign Image and Text confounders are shown in the

Middle and *Right* of Fig. 8.11, respectively. They find that when evaluated on this benchmark, state-of-the-art methods perform poorly compared to humans.

In the following, we look at two causal VQA methods that propose to break spurious associations between text and image.

Counterfactual Vision and Language Learning: Abbasnejad et al. [358] suggest to train VQA models on both observational and generated counterfactual samples to improve generalization. The motivation behind this procedure is to force the model to use both input modalities instead of relying on correlations from one modality only. Typical VQA models possess feature extractors $e^q = f_q(\mathbf{q})$ and $e^v = f_v(\mathbf{v})$ for question and visual input \mathbf{q} and \mathbf{v} , respectively. In contrast, Abbasnejad et al. [358] construct an SCM in which the feature extractors depend on exogenous variables: f_v is replaced by $\tilde{f}_v(\mathbf{v}, \mathbf{u}^v)$ and f_q by $\tilde{f}_q(\mathbf{q}, \mathbf{u}^q)$ where \mathbf{u}^v and \mathbf{u}^q are exogenous variables for image (vision module) and question (language module), respectively. During training, they intervene on either \mathbf{q} or \mathbf{v} , denoted by $\tilde{\mathbf{q}}$ and $\tilde{\mathbf{v}}$, respectively, and yield the corresponding embeddings \tilde{e}^q and \tilde{e}^v . They find their approach effective on both unimodal vision and language tasks as well as multi-modal vision-and-language tasks.

Counterfactual VQA: Niu et al. [62] formulate language confounding as the direct causal effect of questions on answers. To deconfound such spurious associations, they propose to subtract the direct language effect from the total causal effect. They refer to their approach as *Counterfactual VQA* (CFVQA).

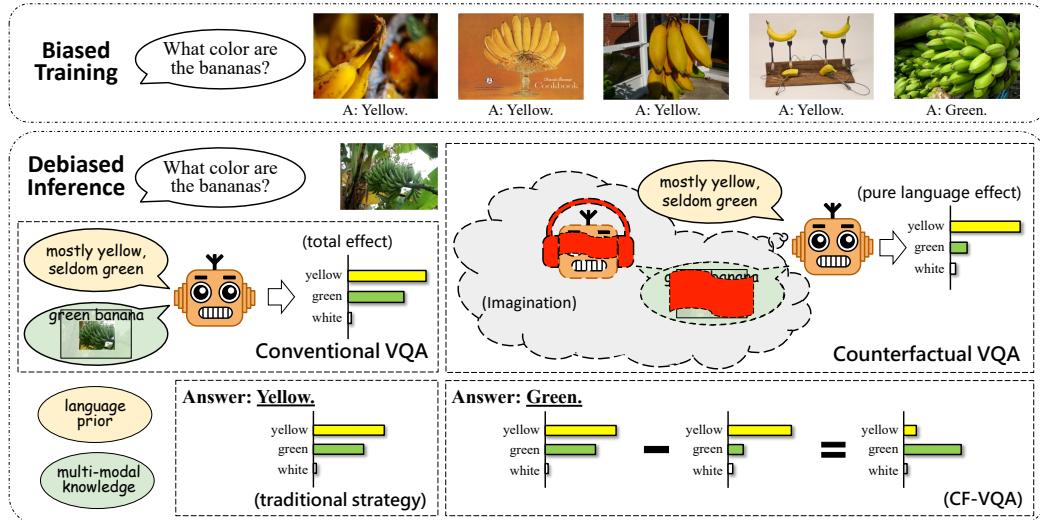
Fig. 8.12a illustrates how conventional and counterfactual VQA differ in their workings. Conventional VQA asks “What will answer A be, if machine hears question Q , sees image V , and extracts the multimodal knowledge K ?”. Doing so cannot disentangle the single-modal linguistic effect and the multimodal reasoning effect.

To isolate the linguistic effect, Niu et al. [62] consider the following counterfactual question: “What would have happened if the model had not performed multimodal reasoning?”. This corresponds to the counterfactual query in which the model the machine considers Q , but the multimodal knowledge K is intervened under the no-treatment condition, i.e., as if V and Q had not been accessible. Since the response of K to Q is now blocked, the model can only rely on the single-modal effect, effectively isolating the language bias. Fig. 8.12c depicts the corresponding causal DAGs.

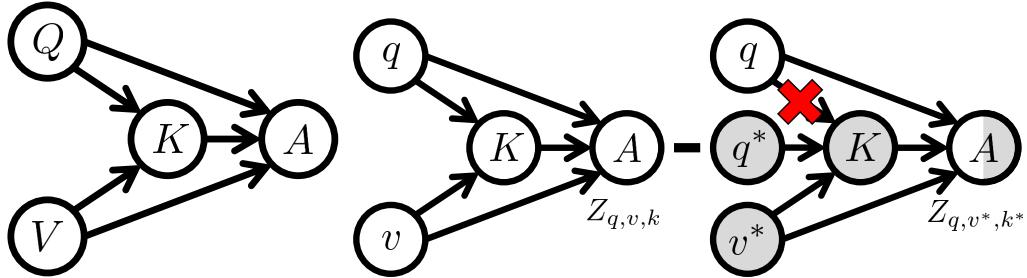
By inferring the counterfactual query “What would A be, if the model reads Q , but had not extracted K or seen V ?”, CFVQA extracts the language bias. Then, to reduce it in the final query, it is subtracted from the total effect of $\mathbf{V} = \mathbf{v}$ and $\mathbf{Q} = \mathbf{q}$ on $A = a$, also referred to as the *total direct effect* (TDE). This TDE estimand is different than conventional VQA’s posterior distribution $p(a | \mathbf{v}, \mathbf{q})$.

8.2.1.2 Certified Robustness Against Natural Language Attacks

Alzantot et al. [359] expose that sentiment analysis models can be fooled by synonym substitution attacks, as illustrated by their adversarial examples in Table 8.1. This



(a) Example of spurious associations between questions and answers. The query “What color are the bananas?” is most often associated with the answer “yellow”.



(b) Causal DAG for VQA. Q: question. V: image. K: multi-modal knowledge. A: answer.
(c) Comparison between conventional VQA (left) and counterfactual VQA (right). Unshaded nodes possess the value $V=v$ and $Q=q$. Shaded nodes are at the value $V=v^*$ and $Q=q^*$, where v^*, q^* denote the non-treatment condition where v and q are not given.

Figure 8.12: Counterfactual VQA [62]. Conventional VQA relies on spurious associations. CFVQA isolates the language effect by imaging the scenario where the model reads the question but the multi-modal knowledge (including image) is removed. By subtracting the pure language effect from the total effect, CFVQA deconfounds the answer

<p style="text-align: center;">Original Text Prediction = Negative. (Confidence = 78.0%)</p> <p><i>This movie had terrible acting, terrible plot, and terrible choice of actors. (Leslie Nielsen ...come on!!!) the one part I considered slightly funny was the battling FBI/CIA agents, but because the audience was mainly kids they didn't understand that theme.</i></p>
<p style="text-align: center;">Adversarial Text Prediction = Positive. (Confidence = 59.8%)</p> <p><i>This movie had horrible acting, horrible plot, and horrible choice of actors. (Leslie Nielsen ...come on!!!) the one part I regarded slightly funny was the battling FBI/CIA agents, but because the audience was mainly youngsters they didn't understand that theme.</i></p>

Table 8.1: Natural Language Adversarial Examples for the sentiment analysis task [359]. We highlight modified words in green and red for the original and adversarial texts, respectively.

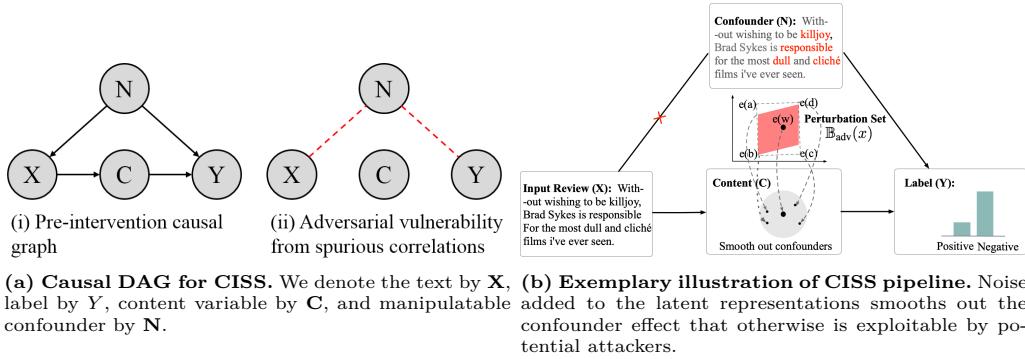


Figure 8.13: Causal Intervention by Semantic Smoothing (CISS) [364]: A framework towards robustness against natural language attacks by learning $p(y | \text{do}(\mathbf{x}))$.

has motivated a myriad of works making NLP models more robust against such attacks [360, 361, 362, 363].

Zhao et al. [364] take a causal perspective on the natural language attack problem and frame the source of adversarial vulnerability as the spurious association induced by confounders. Fig. 8.13a illustrates their Causal DAG. For instance, when considering a movie review (\mathbf{X}) from the IMBD dataset [365], a professional reviewer likely uses jargon (\mathbf{N}), has high standards, and therefore, is likelier to assign a lower score on average. This spurious association can be exploited, e.g., by adding more jargon words to a positive movie review.

To defend models against such attacks, Zhao et al. [364] show that a Gaussian-based randomized classifier models the interventional distribution $p(y | \text{do}(\mathbf{x}))$ and is therefore robust against l_2 -bounded attacks. However, textual input spaces are not continuous, and text substitutions do not follow Gaussian distributions. To circumvent these issues, they propose smoothening out the latent semantic space of the learned content variable \mathbf{C} instead, as illustrated in Fig. 8.13b. They refer to this framework as *Causal Intervention by Semantic Smoothing* (CISS).

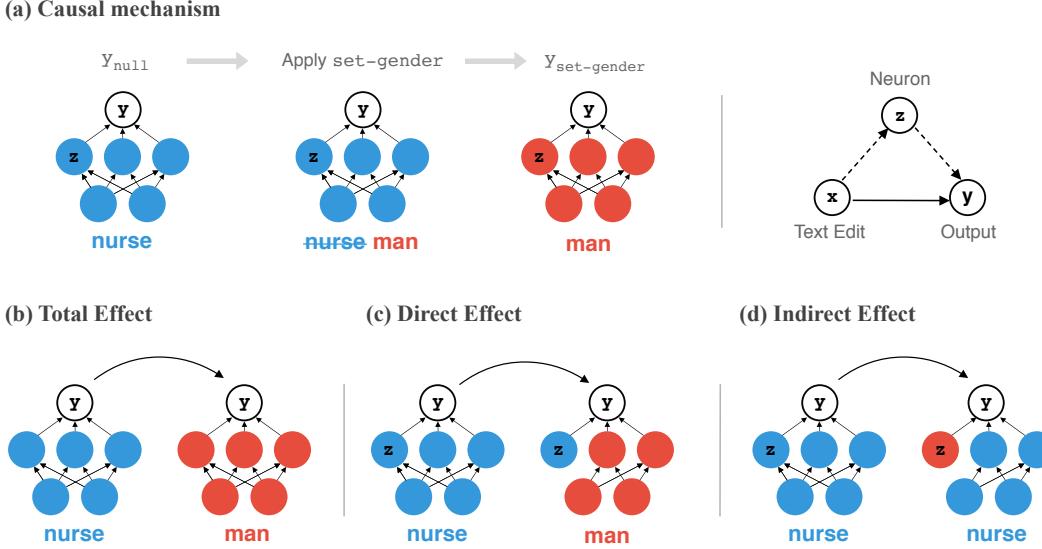


Figure 8.14: Investigating Gender Bias using Mediation Analysis [368]. Given a prompt u such as “The nurse said that”, we ask a language model to generate a continuation. A biased model may assign a higher likelihood to *she* than to *he*. To understand the role of model components on this biased prediction, we perform the do-operation $x = \text{set-gender}$, which changes u from *nurse* to *man* in this example. By inferring direct and indirect effects, we can analyze the causal role of specific mediators (neurons) between x and y .

8.2.2 Counterfactual Explanations

Causal Model Explanation Through Counterfactual Language Models: Understanding the effect of a concept in the input on a model is crucial for explanation and model dissemination. However, this usually requires generating counterfactual sequences by dropping/replacing the concept of interest, which is challenging for existing text generation models. Feder et al. [366] propose a framework named CausalLM to generate counterfactual representations instead of counterfactual sequences. To this end, Feder et al. [366] fine-tune deep contextualized embedding models with auxiliary adversarial tasks to encourage the model to “forget” the concept of interest. The representation of the input sequence and the counterfactual representation forgetting the concept of interest are fed into a classifier to measure the effect of the concept on the classifier’s prediction.

Investigating Gender Bias Using Causal Mediation Analysis: Many text corpora contain spurious associations due to gender stereotypes propagated or amplified by NLP systems. For example, the sentence “he is an engineer” is more likely to appear in a corpus than “she is an engineer” due to the current gender disparity in engineering [367].

Vig et al. [368] propose a method to interpret which parts of a model are biased. Their method uses causal mediation analysis to locate which parts of a neural model are causally implicated, as illustrated in Fig. 8.14. This approach is arguably better than other analysis tools like probing, as probing can only measure whether the in-

formation is encoded in hidden representations rather than whether the information is used by the model.

Counterfactual Fairness in Text Classification through Robustness: Text classifiers are sensitive to specific contents of the input. For example, a toxicity model predicts that the toxicity of "Some people are gay" is 98%, while the toxicity of "Some people are straight" is 2%. Garg et al. [369] study counterfactual fairness in text classification by asking the counterfactual question "How would the output of a classifier change if some sensitive tokens were different?". They define a metric named *counterfactual token fairness* to measure the differences between the outputs before and after substituting tokens associated with identity groups. They further evaluate three methods to promote fairness: replacing all sensitive tokens with a special token, counterfactual data augmentation, and a regularization called *counterfactual logit pairing*.

8.2.3 Counterfactual Data Augmentation

Counterfactual Generator: Zeng et al. [370] propose a new data augmentation algorithm for entity recognition with counterfactual inference. Each input sequence is divided into two parts, including entity and context. An entity in the input sequence is replaced with another entity of the same entity type. The augmented example is kept if a discriminator correctly recognizes the replaced entity. Zeng et al. [370] find that this data augmentation method can improve the generalization ability in low-resource settings. Besides, training on augmented examples can partially eliminate spurious correlations between contexts and output labels.

Counterfactual Data Augmentation for Neural Machine Translation: Liu et al. [356] design a counterfactual data augmentation method for neural machine translation. The method interprets language models and phrasal alignment causally. It creates augmented parallel translation corpora by answering the question "Given that Y_j is aligned to X_i , what would Y_j have looked like, had $X_i = \hat{x}_i$ and all other phrases $\mathcal{X}_{-i}, \mathcal{Y}_{-j}$ had been held constant?". They use a masked language model and a translation language model to replace the source and target phrases, respectively. Compared to previous work, this method takes context and alignment into account for data augmentation.

Mitigating Gender Stereotypes in Languages with Rich Morphology:

In Sec. 8.2.2, we learned that many text corpora include gender biases. Typically, such biases exist in many different languages around the world. Most NLP research has focused on mitigating gender stereotypes in English [371, 372]. However, these approaches often create ungrammatical sentences in morphologically rich languages like Spanish. To this end, Zmigrod et al. [373] present a counterfactual data augmentation approach for mitigating gender stereotypes associated with nouns representing people in such languages.

Their unsupervised approach uses dependency trees, lemmata, part-of-speech tags, and morpho-syntactic tags from Universal Dependencies corpora [374]. It consists of four steps: (1) Analysis of the sentence (including parsing, etc.), (2) Intervention on a gendered word, (3) Inference of the new morpho-syntactic tags, and (4) Reinflection of the lemmata to their new forms. Using four different languages, they demonstrate that their approach reduces gender stereotyping by an average of 2.5 without sacrificing grammaticality.

8.2.4 Miscellaneous

Causal Direction of Data Collection: The supervised learning problem aims to predict a label Y based on features \mathbf{X} . From a causal perspective, we can further distinguish this problem into two cases: *causal* and *anti-causal* learning, determined by the data collection process being $\mathbf{X} \rightarrow Y$ or $Y \rightarrow \mathbf{X}$, respectively. In words, if, during the data collection process, \mathbf{X} is generated first, and then Y is collected based on \mathbf{X} (e.g., through annotation), we say that \mathbf{X} causes Y (causal learning). Vice versa, if Y is generated first, and then \mathbf{X} is collected based on Y , we say that Y causes \mathbf{X} (anti-causal learning).

Previous works show that this simple distinction has important implications for scenarios such as covariate shift, transfer learning, semi-supervised learning [375], and adversarial examples [376].

Category	NLP Tasks
Causal	Summarization, parsing, tagging, data-to-text generation, information extraction
Anticausal	Author attribute / review sentiment classification
Mixed	Machine translation, question answering, question generation, text style transfer, intent classification

Table 8.2: Classification of typical NLP tasks into causal, anticausal, and mixed learning tasks [377]. Causal learning means that the model takes the cause as input and predicts the effect; anticausal refers to settings where the model takes the effect as input and predicts the cause. Some tasks do not have a clear causal interpretation of the data collection process, or a mixture of both types of data is typically used.

Jin et al. [377] study the causal direction of the data collection process for common NLP datasets, as summarized in Table 8.2. For example, they observe that language sentence pairs are mixed up in machine translation regardless of their original source-to-target direction (e.g., whether a sentence originated in English and was translated into Spanish or vice versa). Splitting the data into subsets reveals that they exhibit different properties, such as performance differences in self-supervised or domain adaptation settings. In light of these findings, the authors make various recommendations for future research, e.g., annotating the causal direction when collecting new NLP data and incorporating it into the model.

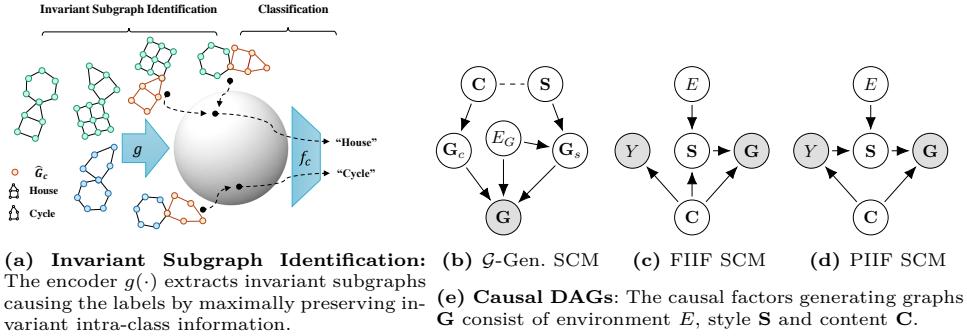


Figure 8.15: Graph Out-Of-Distribution (GOOD) Generalization Framework [382]: A Style- and Content Decomposition (Def. 3.1.1) perspective on graph distribution shifts.

8.3 Causal Graph Representation Learning

In numerous applied fields, graphs represent systems of relations and interactions. Examples include social networks [378], molecular graphs [55], protein associations [379], and programming code syntax trees [380]. Graph Neural Networks (GNNs) form an effective framework for learning representations over graph-structured data.

8.3.1 Causal Supervised Learning

Discovering Invariant Rationales for GNNs: Intrinsic interpretability of graph neural networks aims to discover a small subset of an input graph that contributes most to the model prediction. However, most work on intrinsic interpretability is prone to discover data biases and spurious correlation, failing to capture causal patterns. Besides, these methods suffer from performance degradation on out-of-distribution data as they are sensitive to spurious correlation. Wu et al. [381] propose a new method by discovering invariant rationale. A rationale generator is first applied to split each input graph into causal and non-causal subgraphs. Next, they encode causal and non-causal subgraphs into hidden representations. Then causal interventions are conducted on the non-causal representations to create perturbations. Finally, they utilize two classifiers for the final prediction, trained through an invariant risk loss function.

Invariance Principle Meets Out-of-Distribution Generalization on Graphs: Chen et al. [382] propose the Graph Out-Of-Distribution (GOOD) framework, akin to the Style- and Content Decomposition (Def. 3.1.1), to learn invariant graph features. To this end, a featurizer is designed to distinguish invariant subgraphs from the other parts of the graph that domain shifts can easily perturb. The authors demonstrate that this approach improves out-of-domain generalization.

Assuming the true underlying causal DAG in Fig. 8.15b, Chen et al. [382] categorize the interactions between content \mathbf{C} and style \mathbf{S} into three possible DAGs, depending on whether \mathbf{C} is fully informative about Y , i.e., $(\mathbf{S}, E) \perp\!\!\!\perp Y \mid \mathbf{C}$: Fully Informative Invariant Features (FIIF, Fig. 8.15c), Partially Informative Invariant Features (PIIF,

Fig. 8.15d). The main difference is that \mathbf{S} is directly controlled by \mathbf{C} in FIIF, and indirectly controlled by \mathbf{C} through Y in PIIF, which can exhibit different behaviors in the observed distribution shifts.

To extract the invariant subgraph containing the latent content features \mathbf{C} from the complete observed graph \mathbf{G} , the authors explicitly align the two causal mechanisms during the graph generation, i.e., $C \rightarrow G$ and $(\mathbf{G}_s, E_G, \mathbf{G}_c) \rightarrow G$. Fig. 8.15a illustrates the alignment procedure: they decompose a GNN into two sub-components: i) a featurizer GNN $g : \mathcal{G} \rightarrow \mathcal{G}_c$ aiming to identify the desired \mathbf{G}_c ; and ii) a classifier GNN $f_c : \mathcal{G}_c \rightarrow \mathcal{Y}$ that predicts Y based on the estimated \mathbf{G}_c , where \mathcal{G}_c refers to the space of subgraphs of \mathbf{G} .

Deconfounded Training for Graph Neural Networks: Sui et al. [383] aims to discriminate each input graph into critical and trivial parts. Since the trivial part is a confounder between the critical part and the label, this opens a backdoor and leads to spurious correlations. Therefore, Sui et al. [383] propose deconfounded training to mitigate the confounding effect. They use attention modules to split each graph into a critical and trivial subgraphs. Then they eliminate the confounding effect of trivial subgraphs by performing the backdoor adjustment.

8.3.2 Counterfactual Data Augmentation

Learning from Counterfactual Links for Link Prediction: A common GNN task is to predict edges (or *links*) between node pairs. For example, links can refer to citations between papers [384], relations in knowledge graphs [380], or molecule interactions [55].

Zhao et al. [385] point out that the causal relationship between the graph structure and link existence has been largely ignored. The authors give the following example of a social network. Imagine Alice and Adam live in the same neighborhood, and they are close friends. Associating neighborhood belonging with friendship could be too strong to identify the essential components of friendship, like mutual interests or family ties. Such spurious associations may also explain why they live in the same neighborhood.

Consequently, Zhao et al. [385] propose to examine the counterfactual query “would Alice and Adam still be close friends if they were not living in the same neighborhood?”. Generally speaking, any counterfactual link prediction queries follow the structure of “would the link exist or not if the graph structure was different from that observed?”. Hence, if we train link prediction models on such corresponding counterfactuals, they rely less on spurious associations, as described above.

Zhao et al. [385] generate counterfactual links to augment the training data for link prediction tasks. Their approach estimates the causal relationship between the observed graph structure (considered as the intervention) and link existence (the outcome). Fig. 8.16 summarizes their approach.

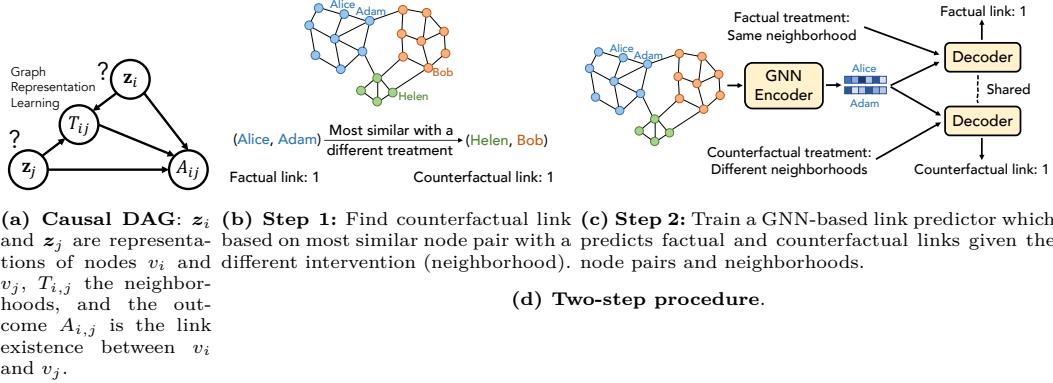


Figure 8.16: Counterfactual Link Prediction Framework [385].

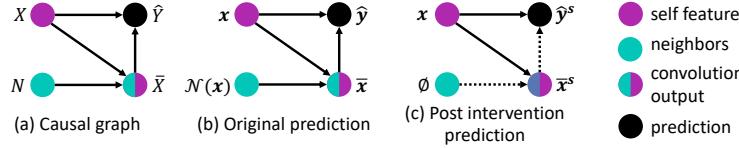


Figure 8.17: Causal Graph Convolutional Network Inference (CGI) [387]. (a) Causal graph of CGI; (b) Original prediction; (c) causal intervention $do(N = \emptyset)$ where the dashed arrow indicates that the effect from the predecessor is blocked.

8.3.3 Miscellaneous

Tackling Over-Smoothing by Causal Effect Uncertainty: Real-world graphs usually exhibit locally varying structures, i.e., uneven distributions of properties such as homophily and degree. However, a commonly-faced issue faced by GNNs is *over-smoothing*, which means that the representations of the graph nodes of different classes become indistinguishable, often causing local structure discrepancies [386].

Feng et al. [387] aim to resolve such node-specific structure discrepancies *at inference time* by taking into account the uncertainty of the causal effect of the local structure on the prediction. The causal effect uncertainty indicates whether a test node should be trusted when predicting its label based on its local structure. For example, if the local structure exhibits different properties from those observed, the causal effect uncertainty will be high, suggesting that the predicted class may be incorrect.

Fig. 8.17 illustrates this approach: Using the causal graph (a), we can compare the original (b) and post intervention prediction, which estimates what the prediction would be if the target node has no neighbors (c). By computing the difference between the two, we assess the causal effect of the neighborhood on the label prediction. Depending on this causal effect and a decision threshold, we include the node or discard it from the training data.

Relating Graph Neural Networks to Structural Causal Models: Graph neural networks are a powerful tool for learning with relational data. Zečević et al. [388] provide a theoretical analysis to establish several connections between graph neural networks and SCMs. They design a new neural-causal model class by formalizing

interventions for graph neural networks. The authors also provide theoretical results and proofs on the feasibility, expressivity, and identifiability of this new neural-causal model class for causal inference.

9

Causal Benchmarks

This section gives an overview of benchmarks designed for CAUSALML tasks and includes interventional or counterfactual ground-truth data. We discuss the current limitations of benchmarks more critically in Sec. 10.3.1.

We believe it is worth mentioning that for some tasks, one can (and should) evaluate CAUSALML methods on “conventional” benchmarks developed without causality-specific design choices. For example, in Chapter 3, we discussed methods that can be used for out-of-distribution (OOD) tasks. One may evaluate these approaches on standard OOD benchmarks, such as some causal invariance learning methods (Sec. 3.1.1.2).

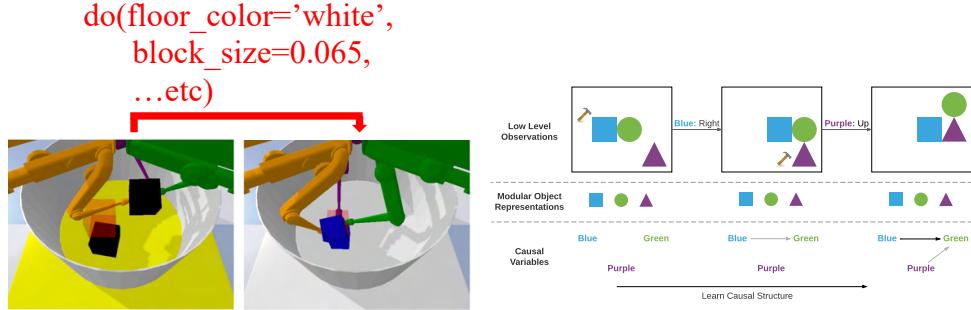
Similarly, conventional RL benchmarks rely on a simulation engine to generate trajectory data. Famous examples within the RL community are *MuJoCo* [389], a physics engine offering a suitable playground for continuous control tasks, or the Arcade Learning Environment, allowing agents to play Atari 2600 games [390]. One can use these existing benchmarks for evaluating CausalRL methods simply by interpreting (PO)MDPs as SCMs (Causal Perspective 7.5.2).

Nevertheless, not all conventional ML benchmarks are suitable for evaluating the techniques discussed in this work. For example, to probe visual reasoning models on causal comprehension (e.g., “What effect will pushing this object have?”), we require question and answer pairs that encompass causal relationships.

9.1 Reinforcement Learning

Several authors designed RL simulators with high-level causal variables to facilitate the symbiosis of Causality and RL. While conventional benchmarks such as MuJoCo can be tweaked to allow an agent to intervene on environment variables (such as masses or lengths of the physical system [91]), CausalRL benchmarks such as *CausalWorld* offer a well-defined API and underlying causal graph to simplify and extend environment interventions.

CausalWorld: *CausalWorld* [10] is a robotic manipulation simulator that provides a combinatorial family of tasks with common causal structures and underlying factors (e.g., robot and object masses, colors, and sizes). The user or the agent can intervene on a subset of causal variables that determine the environment dynamics, allowing them to control how similar tasks are.



(a) CausalWorld [10]: A robotic manipulation benchmark for causal structure and transfer learning. Tasks consist of constructing 3D shapes from a given set of blocks, inspired by how children learn to build complex structures.

(b) Visual Causal Discovery [391]: A benchmark for learning representations of high-level causal variables from raw pixel data. Environments have objects that interact according to the underlying causal graph.

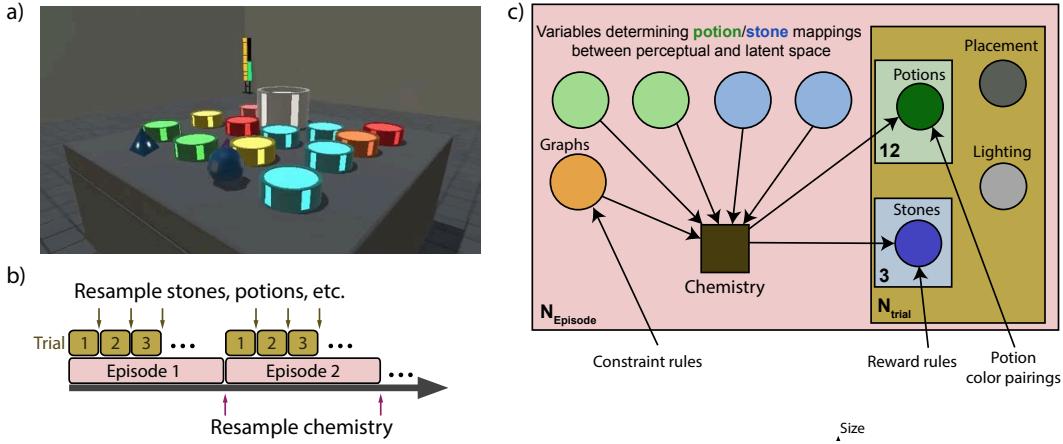


Figure 9.2: Alchemy [392]: A 3d video game implemented with the Unity game engine, revealing an accessible ground-truth task parameterization. It involves a causal structure (c) that is periodically resampled (b), offering a testbed for RL, structure learning, online inference, and hypothesis testing. It provides two observation spaces; one is symbolic and the other is image-based.

Visual Causal Discovery: Ke et al. [391] note that RL agents often only observe low-level variables like pixels in images, and have to induce high-level causal variables (Sec. 4.3.1). To evaluate the ability of model-based RL methods to identify these causal variables and structures, they design a suite of physics and chemistry environments whose underlying causal graphs are parameterizable.

Alchemy: Wang et al. [392] propose *Alchemy*, a 3D video game with a latent causal structure that is re-sampled procedurally in each episode. This environment provides a task distribution whose parameterization is accessible to the researcher yet, yields challenging tasks to solve, as demonstrated through experiments in which non-trivial deep RL methods fail. Based on probing experiments and analyses, they conclude that agents have to identify relevant parts of the latent structure to solve the tasks.

CausalCity: McDuff et al. [393] develop CAUSALCITY, a high-fidelity simulator for causal reasoning in the safety-critical context of driving. The goal is to navigate vehicles that have “agency”, high-level configurations controlling their sequence of actions (e.g., turn left at the next intersection), deciding their low-level behaviors

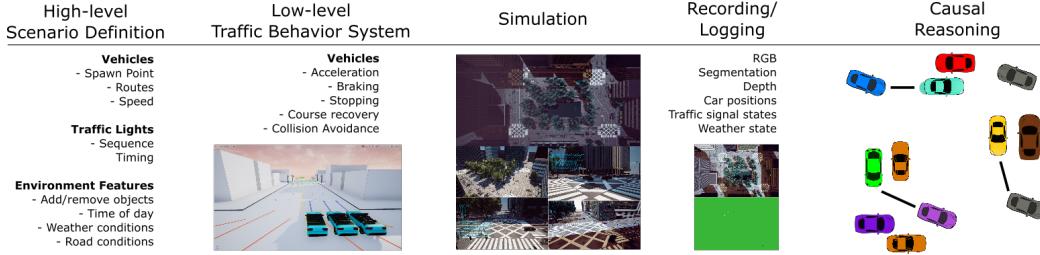


Figure 9.3: CausalCity [393]: A multi-agent driving simulator for complex temporal causal events, such as driving and vehicle navigation. It introduces *agency* abstractions, such that one can simulate complex scenarios by only defining high-level variables.

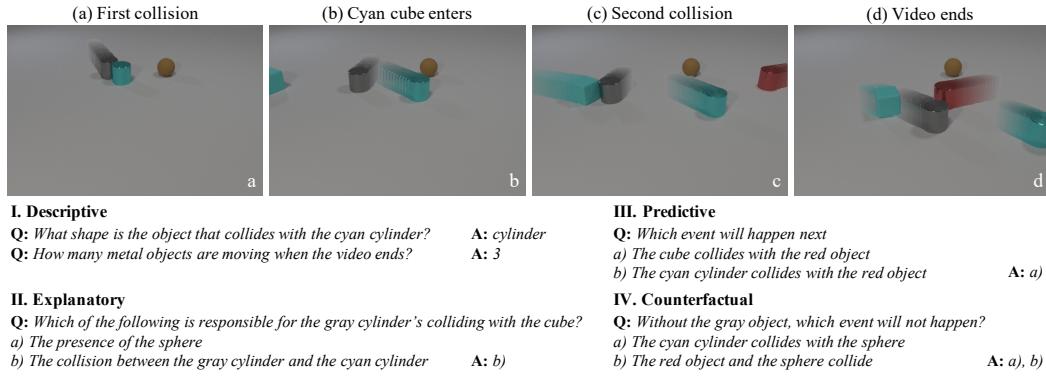


Figure 9.4: Sample video, questions, and answers from the CLEVRER dataset [394], designed to evaluate whether visual reasoning models understand the following classes of questions: (i) descriptive, (ii) explanatory, (iii) predictive, and (iv) counterfactual. All tasks, except (i), are considered causal. We include captions for the readers to understand the frames better, but they are not part of the dataset.

(e.g., their speed). The environment is designed to simulate scenarios with complex causal relationships, including different types of confounders (e.g., weather conditions)

9.2 Computer Vision

Clevrer: Yi et al. [394] introduce the *CoLlision Events for Video REpresentation and Reasoning* (CLEVRER) dataset. This video dataset allows us to evaluate models on four reasoning tasks: (i) descriptive, (ii) explanatory, (iii) predictive, and (iv) counterfactual. The authors interpret (ii)-(iv) as causal tasks and show that various state-of-the-art models for visual reasoning perform poorly on these. They conclude that future methods must learn the underlying causal relations between the objects seen in the images.

CoPhy: Baradel et al. [316] develop the CoPHY benchmark, a synthetic 3D environment for causal physical reasoning. It consists of a number of physical dynamics scenarios, such as *tower of blocks falling*, *spheres bouncing against walls* or *objects colliding*.

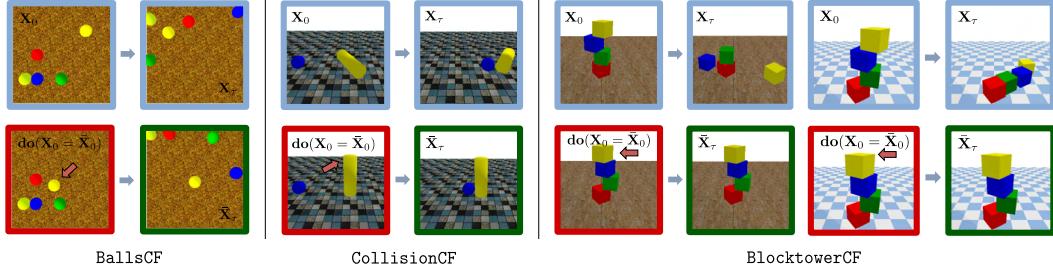


Figure 9.5: Counterfactual Physics benchmark (CoPhy) [316]. Given an observed frame \mathbf{X}_0 and a sequence of future frames $\mathbf{X}_{1:\tau}$, it generates ground-truth counterfactual trajectories corresponding to if we had intervened upon \mathbf{X}_0 and set it to $\hat{\mathbf{X}}_0$ (e.g., changing the initial positions of objects in the scene).

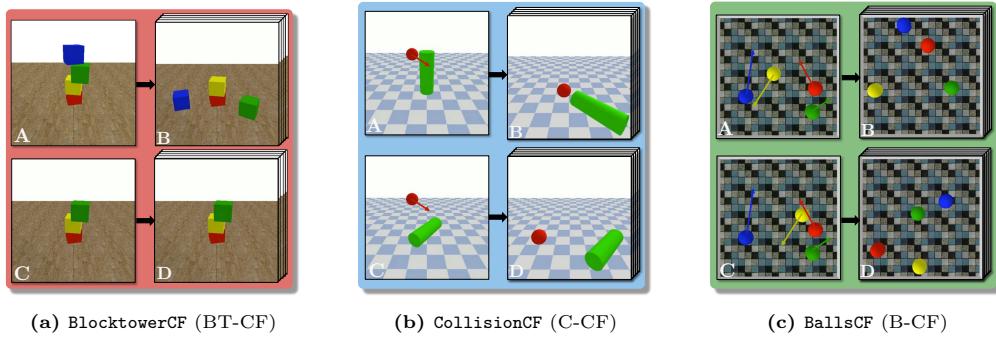


Figure 9.6: Filtered-CoPhy benchmark suite [347]. It contains three challenging scenarios involving 2D or 3D rigid body dynamics with complex interactions, including collision and resting contact. Similar to CoPhy, we can intervene on the initial conditions \mathbf{X}_0 and set them to $\hat{\mathbf{X}}$. Arrows indicate the initial motion.

lidings, as illustrated in Fig. 9.5. It allows us to intervene on the initial conditions of a trajectory and simulate its ground-truth counterfactual.

Filtered-CoPhy: Janny et al. [347] propose a benchmark to test for counterfactual trajectory prediction, where given an observed initial condition \mathbf{x}_0 and a sequence of T frames $\mathbf{x}_{1:T}$, one aims to predict a counterfactual sequence $\mathbf{x}_{1:T}^{\text{CF}}$ under a new initial condition \mathbf{x}_0^{CF} . There are three types of video data available to test: *BlocktowerCF*, *BallsCF* and *CollisionCF*, and confounders that affect the prediction are designed to be obtainable with a sufficiently powerful representation.

Causal3DIdent: Kügelgen et al. [41] introduce an image dataset of 3D objects with causal dependencies, that can be used to study the effectiveness of data augmentations for style and content decomposition techniques (Sec. 3.1), which aim at isolating invariant content and discarding varying style.

9.3 Natural Language Processing

Constructing causal benchmarks involving text is challenging as the vocabulary size of a text corpus is usually large (e.g., the vocabulary of BERT [395] has around 30K

tokens) and ground-truth data for multiple interventions is rare in natural language processing [349].

Kaushik et al. [396] provide two counterfactual NLP datasets for sentiment analysis and natural language inference, respectively. Given initial documents and labels, the authors recruited humans to revise the documents to conform to a counterfactual target label while ensuring internal consistency and avoiding gratuitous changes to facts unrelated to the target label.

Feder et al. [366] propose four NLP datasets designed for causal explanations, three of which include ground-truth counterfactual examples for a given concept. These datasets allow researchers to evaluate counterfactual representations that dropped a given concept of interest, as further explained in Sec. 8.2.2.

Frohberg and Binder [397] introduce CRASS, a counterfactual reasoning assessment benchmark for language models that is part of the BIG-bench suite [398]. For example, they showcase the query “*A woman sees a fire. What would have happened if the woman had fed the fire?*” with three possible answers (a) “*The fire would have become larger*”, (b) “*The fire would have become smaller*”, and (c) “*That is not possible*”. “*A women sees a fire*” is the base premise and “*What would have happened if the woman had fed the fire?*” is a questionized counterfactual conditional (QCC). The possible answers, i.e., a correct consequence and a set of potential effects as distractors, define a so-called *premise-counterfactual tuple* (PCT). More formally, A QCC has the form “*What would have happened if \mathbf{A}^{CF} ?*”, where \mathbf{A}^{CF} is some altered version of the base premise \mathbf{A}^O , effectively generating the counterfactual.

Yang et al. [399] propose a fine-grained causal reasoning dataset, including the three tasks of *causality detection*, *fine-grained causality extraction* and *Causal QA* tasks. To motivate the need for finer granularity, they give the following example: “the spread of COVID-19 has *led* to the boom in online shopping, [cause], but it also has *deterring* [prevent] people from going shopping centres”. The authors argue that previous datasets only considered the above-annotated cause relation, but not more fine-grained causal events like *enable* or *prevent*.

Resuming this example, consider a different passage “COVID-19 has accelerated change in online shopping, and given Amazon’s ... it will result in economic returns for years to come and offering more competitive prices compared to an offline business that brings pressures for the offline business recruitment.”. Previous datasets allow models to extract facts such as “COVID-19 causes an increase in online shopping”, yet, they cannot detect the subsequence for Amazon to “offer more competitive prices”, and the negative influence on offline business recruitment. Both of these can be useful, e.g., if we ask the what-if question “What if COVID stops?”, whose correct answer in the above context should include “there will be more offline business recruitments.” Their experiments reveal a significant gap between model and human ceiling performance (74.1% vs. 90.53% accuracy), providing evidence that statistical models still struggle to solve causal reasoning problems.

10

The Good, the Bad and the Ugly

In this section, we want to provide our perspective on what benefits CAUSALML may buy us (the *good*), what challenges the field is currently struggling with (the *bad*), and what price one has to pay for CAUSALML (the *ugly*). In other words, this section provides an informative discussion of the advantages and disadvantages of *current* CAUSALML methods.

10.1 The Good

We discussed many methods making explicit use of the various causal formalisms, such as SCMs (Sec. 2.3), interventions (Sec. 2.3.1) or counterfactuals (Sec. 2.3.2). Causality presents a framework (SCMs) to formally express assumptions about the data generating process that we would like to encode into our models and a mathematical tool (the do-operator) that can enforce properties thereof during the model training.

In the following, we briefly summarize some of the benefits causal formalisms can bring us across the discussed problem fields. Many of these benefits cannot be recovered with purely statistical reasoning.

With Causal Supervised Learning (Chapter 3), we can improve predictive models: Invariant feature learning approaches (Sec. 3.1) attempt to identify a set of content variables \mathbf{C} that represented the causal parents of Y , and learn a predictor $p(y | \mathbf{c})$ that is invariant to interventions on the style variables \mathbf{S} . In contrast, invariant mechanism learning approaches (Sec. 3.2) model the change of distribution caused by interventions on a set of independent unobserved confounders \mathbf{U} , and they learn separate mechanisms to model for each. To summarize, these methods build the foundation to learn domain-robust, reusable features or mechanisms [11].

Causal Generative Modeling (Chapter 4) offers a principled framework for the task of controllable generation (i.e., generation with control over certain attributes). The structural assignment learning (Sec. 4.1) approaches allow practitioners to add domain knowledge of an underlying causal graph for the DGP. These generate counterfactual samples that consider causal dependencies between the attributes of interest and the other generative variables. The causal disentanglement (Sec. 4.2) approaches go one step further and perform causal graph discovery as well as structural assignment learning, which allows for controllable generation by exploiting weaker forms of available causal domain knowledge.

Through Causal Explanations (Chapter 5), we gain interpretability of model predictions. Without access to the underlying causal dependencies, we can use feature attribution methods (Sec. 5.1) to identify which variables under intervention are most relevant to changes in the model output. With access to the underlying causal graph, we can generate contrastive explanations (Sec. 5.2), which suggest counterfactual model outputs dependent on actionable alternative inputs.

By using Causal Fairness (Chapter 6) criteria, we obtain counterfactual, or interventional, quantities that allow us to evaluate the fairness of prediction models dependent on sensitive attributes of interest. With access to underlying causal dependencies in the DGP, we can enforce case-specific fairness criteria that deconfound potential sources of selection bias.

In *Causal Reinforcement Learning* (Chapter 7), we reviewed publications that use interventions for formalizing actions (Sec. 7.2), changes in the environment (Sec. 7.4), deconfounding of observed trajectory data (Secs. 7.5 and 7.6), the effects of actions on the reward (Sec. 7.7), changing structure in the state space, exposing opportunities to generate counterfactual data augmentations, allowing us to recycle already observed data.

10.2 The Bad

10.2.1 Lack of Open-Source Ecosystem

Most open-source machine learning software packages or model hubs focus on observational models. Automatic differentiation frameworks like PyTorch [400], TENSORFLOW [401] and JAX [402] (including FLAX [403] and HAIKU [404]) paired with model libraries like TRANSFORMERS [405], TIMM [406], or PYG [407] facilitate rapid prototyping of model pipelines. Importing datasets, state-of-the-art (pre-trained) models, and launching a training loop can be done in a few lines of code. Model hubs like [Tensorflow Hub](#), [Huggingface Models](#) or [Pytorch Hub](#) provide thousands of pre-trained models across multiple modalities ready for fine-tuning and deployment.

Unfortunately, there is a much smaller open source ecosystem for learning or performing inference with SCMs. At the time of this writing, we are unaware of software libraries providing APIs for convenient identification tests, manipulating SCMs, or importing causal benchmarks (Chapter 9) as well as pre-trained SCMs to facilitate progress on novel research ideas.

We believe that the ingredients to either (a) build novel CAUSALML-focused platforms or (b) incorporate CAUSALML techniques and models into existing ones are ready. For example, causality researchers often prove causal estimand identification formally by hand, although, e.g., Xia et al. [408] develop an algorithm that verifies identifiability automatically for differentiable SCMs. Providing convenient APIs for such algorithms would be immensely helpful.

For estimating causal estimands using machine learning techniques, there exist some libraries. For example, Scutari [409] and Kalisch et al. [410] propose R packages for causal discovery; Sharma and Kiciman [411], Chen et al. [412], Bach et al. [413] propose Python packages for treatment effect estimation (see more details on these in Sec. 11.2). James Fox et al. [414] introduce a Python library for (multi-agent) causal influence diagrams, which are commonly used for decision-making problems under uncertainty.

10.2.2 Lack of Comparisons to Non-Causal Methods

Several CAUSALML papers lack experimental comparisons to non-causal approaches that solve similar, if not identical, problems. While the methodology may differ, e.g., depending on whether causal estimands are involved, some of these methods claim to improve performance on non-causal metrics, such as accuracy in prediction problems or sample-efficiency in RL setups. This trend of not comparing against non-causal methods evaluated on the same metrics harms the measure of progress and practitioners who have to choose between a growing number of methods.

One area in which we have identified indications of this issue is invariance learning (Sec. 3.1). Some of these methods are motivated by improving a model’s generalization to out-of-distribution OOD data; however, they do not compare their method against typical domain generalization methods, e.g., as discussed in Gulrajani and Lopez-Paz [415]. For example, Mouli and Ribeiro [52] tackle OOD tasks by learning counterfactually-invariant representations with asymmetry learning (discussed in Sec. 3.1.1.3); yet, in their experiments, they do not compare against any non-causal OOD method (see e.g., Gulrajani and Lopez-Paz [415], Wang et al. [416]).

Another area is the causal model-based RL literature (Sec. 7.3). Sontakke et al. [243] (Sec. 7.4) learn disentangled latent task embeddings, arguing these are more interpretable than previous (non-causal) latent task embedding methods, which define their setup using the formalisms of Hi-Param MDPs [244] and BAMDPs [321, 417]. In the experiments, they compare their agent in terms of sample efficiency against baselines not having access to any task embeddings, excluding Hi-Param MDP and BAMPD approaches (e.g., [245, 417]). Zhang et al. [246], presented in Sec. 7.4, argue that their method makes stronger assumptions on how tasks relate to each other compared to common multi-task RL methods. However, their experiments do not compare against any multi-task RL method (e.g. [245, 330, 418]). Mutti et al. [250] aim at learning an agent that can systematically generalize across a universe, i.e., an infinite set of possible MDP tasks (Sec. 7.4). Their experiments evaluate how well their method approximates the optimal value function but do not compare against non-causal models.

10.3 The Ugly

10.3.1 Difficulties of Obtaining Ground-Truth Evaluation Data

One of the biggest open problems in CAUSALML is the lack of public benchmark resources to train and evaluate causal models. Cheng et al. [419] find that the reason for this lack of benchmarks is the difficulty of observing interventions in the real world because the necessary experimental conditions in the form of randomized control trials (RCTs) are often expensive, unethical, or time-consuming. In other words, collecting interventional data involves actively interacting with an environment (i.e., *actions*), which, outside of simulators, is much harder¹ than, e.g., crawling text from the internet and creating passively-observed datasets (i.e., *perception*). Evaluating estimated counterfactuals is even worse: by definition, we cannot observe them, rendering the availability of ground-truth real-world counterfactuals impossible [420].

The *pessimistic* view is that yielding “enough” ground-truth data for CAUSALML to get deployed in real-world industrial practice is unlikely soon. Specifying how much data is “enough” is task-dependent; however, in other fields that require active interactions with real-world environments, too (e.g., RL), progress has been much slower than in fields thriving on passively-collected data, such as NLP. For example, in robotics, some of the best-funded ML research labs shut down their robotics initiatives due to “not enough training data” [421], focusing more on generative image and language models trained on crawled internet data.

Moreover, a curse of the need for simulated data is its manipulability and the consequent lack of consistency across papers. Authors can easily create novel or modify existing simulations tailor-made for an empirical verification for their particular setup/method, which may not generalize to other setups. Such inconsistencies hinder progress because it is harder to compare methods. Real-world fixed datasets such as IMAGENET [25] are less prone to such problems since papers altering them raise more suspicion.

The *optimistic* view is that the lack of benchmarks is simply due to the field being in its infancy, and both more available RCT datasets and simulators will alleviate the progress of CAUSALML methods. About a decade ago, similar problems existed in the field of (deep) RL, which shares a few parallels with CAUSALML, as debated in Chapter 7. In the meantime, RL simulators paved the way to beating world-champions in board games [422, 423], achieving super-human performance in Atari [424] and grandmaster level in StarCraft II [425], treating sepsis in intensive care [426], or navigating stratospheric balloons [427].

10.3.2 Untestable Assumptions Are Inevitable

By making assumptions about the data-generating process in our SCM, we can reason about interventions and counterfactuals. However, making such assumptions can

¹Real-world experiments in forms of randomized control trials are often expensive, time-consuming, or even impossible due to ethical concerns.

also result in bias amplification [428] and harming external validity [429] compared to purely statistical models. Using an analogy of Ockham’s Razor [430], one may argue that more assumptions lead to wrong models more easily.

For example, Pearl [428] illustrates bias amplification in a setting of hidden confounding (Sec. 11.2.1.2). They show that while adjusting for covariates acting like instrumental variables (i.e., variables that are more strongly associated with the treatment assignment than with the outcome), one may reduce confounding bias, but at the same time, residual bias carried by unmeasured confounders, can build up at a faster rate. Put simply, by making the causal model more complex through adding more covariates that should aid backdoor adjustment, the model residual bias of the causal effect increases in harmful ways. A “simpler” model that excludes covariates that are predictive of the treatments can work better ².

²Chapter 18.5 of [30] discusses such issues in more detail.

11

Related Work

11.1 Other Surveys

Schölkopf [431] discusses links between machine learning and graphical causal inference while introducing causality concepts along the way. Schölkopf et al. [11] reviews fundamental concepts of causal inference and relates them to open machine learning problems with a particular focus on representation learning. They highlight two issues in current deep learning systems: robustness to distribution shifts and learning reusable and modular mechanisms.

Feder et al. [432] examines the intersection of NLP and causality and argues that causal formalisms can make NLP methods more robust and understandable. They list three problems from the NLP literature that motivate this claim: purely associational models may (i) latch onto spurious associations, failing to generalize in OOD settings (e.g., [433]); (ii) exhibit unacceptable performance differences across groups of users (e.g., [372]); (iii) be too inscrutable to incorporate into high-stakes decisions (e.g., [434]).

Cheng et al. [435] discuss how causality may address ethical challenges in socially responsible artificial intelligence. They focus on seven causal inference tools, some of which we have discussed in Chapter 2 (e.g., the do-operator or counterfactual analysis), and some of which we have left out (e.g., mediation analysis or propensity scores).

Liu et al. [436] highlight how causal reasoning can support visual representation learning. Specifically, the authors categorize existing causality-aware visual representation learning work into (i) causal visual comprehension, (ii) causal visual robustness, (iii) causal visual question answering, and (iv) causal datasets. For future work, they enumerate the following potential directions: (i) more reasonable causal relation modeling, (ii) more precise approximations of intervention distributions, (iii) more proper counterfactual synthesizing processes, and (iv) large-scale benchmarks and evaluation pipelines.

Sanchez et al. [437] explore how causal inference can be incorporated into clinical decision support systems using modern machine learning techniques. As a running example throughout their paper, they use Alzheimer’s disease (AD) to illustrate how CAUSALML can benefit clinical scenarios. The authors observe that important challenges in healthcare applications are processing high-dimensional and unstructured data, generalization to out-of-distribution samples, and temporal relationships. All

these challenges may be addressed through CAUSALML, which the authors divide into (i) causal representation learning (Sec. 2.4), (ii) causal discovery (Sec. 11.2.2), and (iii) causal reasoning (estimating interventional distributions, Sec. 2.2.1).

Vlontzos et al. [438] discuss how causality can assist in creating robust and adaptable medical image analysis algorithms. Specifically, the authors point out that many healthcare machine learning approaches fail to translate into clinical practice due to the inability to adapt and be robust to real-world conditions. One cause of such existing models' inability is their lack of ability to distinguish between correlations and causation, which may result in deadly mistakes. For example, [439] identified several approaches that claim to have been able to diagnose COVID-19 from chest X-rays but ultimately failed to do so as they instead relied on spurious features like hospital identifiers and the patient's ethnicity.

11.2 Machine Learning for Causal Inference

This survey uses causality theory to solve common machine learning problems and provide new perspectives on such. An astute reader may wonder about research in the other direction, using machine learning to answer causal questions. We observe numerous studies using modern representation learning techniques to estimate and answer causal queries, such as causal effects. For completeness, we briefly list recent advances in two common causal inference tasks: *causal effect estimation* and *causal discovery*.

11.2.1 Causal Effect Estimation

Estimating causal effects from observational data is a fundamental problem in many fields that face challenges in running randomized control trials. We want to answer causal questions in many scientific or commercial setups; hence, it is fallible to argue just from observed associations. Supervised learning methods face two challenges in such settings: (i) missing interventions, i.e., the fact that we only observe one treatment for each individual means models must extrapolate to new treatments without access to ground truth, and (ii) confounding variables affecting both treatment assignment and the outcome, such that extrapolation from observation to intervention requires additional causal assumptions. The literature on *treatment effect estimation* deals with constructing models that address these issues.

11.2.1.1 Observed confounders

Generally speaking, we can identify causal effects if we observe all confounders. However, depending on the structure of the treatment effect (e.g., smoothness or sparsity [218]), different estimators behave differently. For example, Chernozhukov et al. [217] and Künzel et al. [440] show that simple regression models trained on all training data points regardless of their treatment can easily lead to biased estimates due to imbalance in the treatment assignment; Shalit et al. [441] and Shi

et al. [442] make similar arguments for neural network models. Hence, the bulk of treatment effect estimation works focuses on model regularization to identify the causal associations in the data.

One line of work for causal effect estimation that utilizes modern ML techniques are *meta-learners* (or *plug-in learners*) [440, 443]: they decompose effect estimation into multiple sub-problems (so-called *nuisance components*), each solvable using any modern machine learning technique [18, 217, 440, 444].

We highlight a few of these techniques that utilize neural networks. For binary treatments, Curth and van der Schaar [445] implement multiple meta-learning strategies with neural networks, concluding that theoretically-optimal estimators may not perform best in finite-simple regimes. For scalar-continuous treatments, Nie et al. [446] propose using splines to preserve continuity over the treatment domain. For higher-dimensional treatments, Kaddour et al. [18] present a meta-learning strategy that learns propensity features. For arbitrary treatments, Zhang et al. [447] utilize Transformer networks [83] to construct a flexible architecture.

For causal estimand identification, Malek and Chiappa [448] use the bandit framework to learn the arm that will produce the best estimator for identifying a causal estimand of interest. They assume an online setting in which the practitioner controls the data collection procedure and aims to find the estimand formula with the lowest asymptotic variance in as few samples as possible. Xia et al. [408] develop a necessary and sufficient algorithm that verifies identifiability automatically for differentiable SCMs. Their approach relies on enforcing L_1 -consistency with the data distribution.

11.2.1.2 Unobserved confounders

There are primarily three strategies to deal with unobserved confounding (Sec. 2.5): (1) estimating bounds based on additional assumptions, also referred to as *partial identification* [449, 450], (2) sensitivity analysis of how strong the confounder’s effect has to be to make the true estimand substantially different from our estimate [451, 452], and (3) utilizing other observed variables, such as *instrumental* [453] or *proxy* variables [454]. Specifically for the third category, some recent work explores ML techniques, such as NNs or kernel methods, discussed below.

Proxy Variables: *Proxy variables* contain relevant auxiliary information on the confounder; ideally, enough to completely recover the confounder [454, 455]. For example, we may be interested in estimating the impact of flight ticket prices on sales [456]. A possible confounder is the people’s desire to fly, e.g., driven by holiday seasons that affect both the number of ticket sales and the prices customers are willing to pay. A suitable proxy variable for this desire could be the number of views of the ticket reservation page. While simply adjusting on such a proxy variable alone would produce a biased effect estimate, there are recent methods aiming at producing less biased estimates [457, 458, 459, 460, 461].

Instrumental Variable Regression: Another class of methods rely on an *instrumental variable* I (often just called *instrument*) [462, 463, 464, 465, 466]. Three conditions have to hold for an instrument to be *valid*: (i) it is independent of the hidden confounder, $I \perp\!\!\!\perp C$; (ii) it is not independent of the treatment, $I \not\perp\!\!\!\perp T$; and (iii) it is independent of the outcome conditional on the treatment and confounder $I \perp\!\!\!\perp Y | \{T, C\}$.

In the above flight ticket example, we might consider shifting supply factors, such as the oil price, as a valid instrument, as it only affects sales via price, thereby identifying the customers' demand [463].

Sometimes, we observe treatment corrupted by measurement error, e.g., when we do not observe it directly. For instance, in the setting of medical treatments, suppose patients are asked to take a drug at home instead of a hospital with supervision. Some of them may not comply, and a self-reported measurement is erroneous due to the patient lying or being forgetful. Zhu et al. [467] propose a method to deal with such measurement error scenarios assuming access to an instrument.

11.2.2 Causal Discovery

Causal Discovery is an umbrella term for methods that try to recover the underlying causal structure of a DGP from observational and/or interventional data. Depending on the assumptions one is willing to make, the output ranges from partial node orderings to the full SCM (assuming linear structural equations).

Generally, causal discovery is difficult due to (i) structure identifiability and (ii) computational complexity. First, the identifiability challenge means that, given observational data only, the causal DAG \mathcal{G} is typically not identifiable, as a set of possible graphs could have generated the data [12]. Secondly, due to the combinatorial nature of the solution space, its size grows super-exponentially with the number of variables, rendering naive methods computationally infeasible [468].

We point keen readers looking for more details on these methods to primarily three classes: (i) combinatorial methods [469, 470, 471, 472, 473, 474, 475], (ii) continuous relaxations [476, 477, 478, 479, 480, 481, 482, 483], and (iii) permutation-based methods [484, 485, 486, 487]. Further, Squires and Uhler [472], Vowels et al. [488] provide excellent surveys.

Recovering the full causal graph is especially difficult in high dimensions. Hence, one line of work focused on more specialized problems, e.g., more coarse-grained [489] ancestral structures, partial ancestral graphs [490], or the causal order among a subgraph of variables known to descend from some set of confounding covariates [491]. Another line of work aims to perform causal discovery actively to improve data efficiency and minimize the number of required interventions [492, 493, 494, 495].

11.2.3 Granger Causality

In many scientific fields, such as neuroscience, econometrics, and civil engineering, it is important to understand causal relationships found in time series data. In neuroscience, for example, researchers study whether activity in one brain region correlates with activity in another [496]. An econometrician may be interested in which macroeconomic indicators predict one another [497]. Civil engineers are interested in understanding differences in traffic across highways to build well-utilized transportation infrastructure [498].

Granger causality [499] is a framework for time series structure discovery that quantifies the extent to which the past of one time series aids in predicting the future evolution of another time series. It is based on the simple assumption that causes precede their effects.

Def.: 11.2.1: Granger Causality [7, 499]

Whenever past observations of X help in predicting the future of time-series Y , then X Granger-causes Y . Formally, we write

$$X \text{ Granger-causes } Y : \iff Y_t \not\perp X_{\text{past } (t)} \mid Y_{\text{past } (t)}. \quad (11.1)$$

If our goal is to detect the underlying causal DAG of multiple variables observed over time, we can use Granger causality only sometimes. Specifically, we can recover the DAG if we observe all relevant variables (implying no unobserved confounding, see Sec. 2.6) and no *instantaneous* effects exist, i.e., when no variable causes another at the same time step. This condition and other limitations are explained in more detail by Peters et al. [7, Chapter 10].

There are several interesting extensions of Granger causality. Originally, Granger causality was defined using linear relations. Tank et al. [500] extended it to the non-linear case.

Wu et al. [501] proposes to combine the benefits of Granger causality with deep learning models. They propose relational learning with *minimum predictive information* (MPI) regularization. This MPI regularizer quantifies the directional predictive strength between each pair of time series.

Khanna and Tan [502] utilize Statistical Recurrent Units (SRUs) to model the network topology of Granger causal relations. Specifically, they explain that the structured sparse estimate of the internal parameters of the SRU networks, trained to predict the stochastic processes' time series measurements, relate to the Granger causal relations. Moreover, by regularizing a low-parameterized variant of SRUs, called *economy-SRU*, towards interpretability of its learned predictive features, this model is less likely to overfit the time series data.

Löwe et al. [503] point out that most causal discovery methods based on Granger causality fit a new model whenever they encounter samples from a new underlying causal graph. However, time series with different underlying causal graphs may

share relevant information, e.g., when inferring synaptic brain connections between neurons based on their spiking behavior: test subjects may have varying brain connectivity but the same underlying neurochemistry. Hence, they propose *Amortized Causal Discovery*, an approach that separates the causal relation prediction from modeling the time series dynamics, permitting us to aggregate statistical strength across samples.

12

Conclusion

We summarize some of our key findings:

1. Causal Inference (Chapter 2) encodes causal assumptions about a system of interest. Therefore, as opposed to conventional statistical or probabilistic inference, it allows us to reason about interventional and counterfactual estimands.
2. We argue that these estimands benefit specific areas of ML research, namely:
 - (a) **Causal Supervised Learning** (Chapter 3) improves predictive generalization by learning invariant features or mechanisms, both aiming at deconfounding models' reliance w.r.t. reliance on spurious associations. Future work should investigate targeted benchmarks testing for learned invariances, connections to adversarial robustness and meta-learning, and the potential utilization of additional supervision signals.
 - (b) **Causal Generative Modeling** (Chapter 4) supports sampling from interventional or counterfactual distributions, naturally performing principled controllable generation or sample editing tasks, respectively. All existing methods learn structural assignments; some infer the causal structure from data. However, it is underexplored what levels of abstractions should be considered for different applications, how to scale assignment learning up to larger graphs, and when counterfactually-generated data augmentations are effective (and when not).
 - (c) **Causal Explanations** (Chapter 5) explain model predictions while accounting for the causal structure of either the model mechanics or the data-generating process. We divide these methods into (i) feature attributions, which quantify the causal impact of input features, and (ii) contrastive explanations, representing altered instances achieving the desired outcome. So far, it is unclear how to unify both classes of methodology best, scale up explanations, make them robust against distribution shifts, secure and private against attackers, and circumvent the inevitable trade-off of their robustness against recourse sensitivity.
 - (d) **Causal Fairness** (Chapter 6) paves the way for criteria that assess a model's fairness as well as mitigate harmful disparities w.r.t. causal relationships of the underlying data. However, the criteria rely on either counterfactual or interventional distributions. Future work should shed light on alternatives to equality, fairness outside standard prediction settings, weaker observability

assumptions (e.g., hidden confounding), and the validity of interventionist views on social categories.

- (e) **Causal Reinforcement Learning** (Chapter 7) describes RL methods taking the explicit causal structure of the decision-making environment into account. We categorize these methods into eight categories and observe that their claimed benefits over non-causal methods include deconfounding (resulting in better generalization), intrinsic rewards, and data efficiency. Open problems suggest that some formalisms might be unifiable, deconfounding of offline data has been largely unaddressed in offline RL sections, and agents making decisions based on counterfactuals might offer further benefits.
 - (f) **Modality-Applications:** We review how previously introduced, and modality-specific principles provide opportunities to improve computer vision, natural language processing, and graph representation learning settings.
3. We review existing **Causal Benchmarks** designed for CAUSALML methods, including ground-truth causal interventions and counterfactuals.
 4. We discussed some of the benefits of using CAUSALML methodology, as well as issues around untestable assumptions, lack of benchmarks, lack of software, and lack of comparisons to non-causal methods.
 5. We touched upon two categories of previous related work; other CAUSALML surveys and research that utilizes machine learning to infer causal estimands.

Acknowledgements: In alphabetical order, we thank Damien Teney, David Watson, Jakob Zeitler, Limor Gultchin, Phillip Lippe, Stefano Blumberg, and Tom Everitt for helpful feedback.

Bibliography

- [1] A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman *et al.*, “Underspecification presents challenges for credibility in modern machine learning,” *arXiv preprint arXiv:2011.03395*, 2020.
- [2] A. Jahanian*, L. Chai*, and P. Isola, “On the "steerability" of generative adversarial networks,” in *International Conference on Learning Representations*, 2020. [Online]. Available: <https://openreview.net/forum?id=HylsTT4FvB>
- [3] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.
- [4] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the dangers of stochastic parrots: Can language models be too big?” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 610–623.
- [5] Z. C. Lipton, “The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.” *Queue*, vol. 16, no. 3, pp. 31–57, 2018.
- [6] G. Dulac-Arnold, D. Mankowitz, and T. Hester, “Challenges of real-world reinforcement learning,” *arXiv preprint arXiv:1904.12901*, 2019.
- [7] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference - Foundations and Learning Algorithms*, ser. Adaptive Computation and Machine Learning Series. Cambridge, MA, USA: The MIT Press, 2017.
- [8] A. Goyal and Y. Bengio, “Inductive biases for deep learning of higher-level cognition,” *arXiv preprint arXiv:2011.15091*, 2020.
- [9] J. Pearl, “The seven tools of causal inference, with reflections on machine learning,” *Communications of the ACM*, vol. 62, no. 3, pp. 54–60, 2019.
- [10] O. Ahmed, F. Träuble, A. Goyal, A. Neitz, Y. Bengio, B. Schölkopf, M. Wüthrich, and S. Bauer, “Causalworld: A robotic manipulation benchmark for causal structure and transfer learning,” *arXiv preprint arXiv:2010.04296*, 2020.
- [11] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, “Towards causal representation learning,” *arXiv preprint arXiv:2102.11107*, 2021.

- [12] J. Pearl, *Causality*. Cambridge university press, 2009.
- [13] B. Neal, “Introduction to causal inference,” 2020.
- [14] B. Schölkopf and J. von Kügelgen, “From statistical to causal learning,” 2022. [Online]. Available: <https://arxiv.org/abs/2204.00607>
- [15] A. Van Oord, N. Kalchbrenner, and K. Kavukcuoglu, “Pixel recurrent neural networks,” in *International conference on machine learning*. PMLR, 2016, pp. 1747–1756.
- [16] A. van den Oord, N. Kalchbrenner, L. Espeholt, K. Kavukcuoglu, O. Vinyals, and A. Graves, “Conditional image generation with pixelcnn decoders,” in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, Eds., 2016, pp. 4790–4798. [Online]. Available: <https://proceedings.neurips.cc/paper/2016/hash/b1301141feffabac455e1f90a7de2054-Abstract.html>
- [17] Y. Yu, X. Si, C. Hu, and J. Zhang, “A review of recurrent neural networks: Lstm cells and network architectures,” *Neural computation*, vol. 31, no. 7, pp. 1235–1270, 2019.
- [18] J. Kaddour, Y. Zhu, Q. Liu, M. J. Kusner, and R. Silva, “Causal effect inference for structured treatments,” in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021, pp. 24 841–24 854. [Online]. Available: <https://proceedings.neurips.cc/paper/2021/hash/d02e9bdc27a894e882fa0c9055c99722-Abstract.html>
- [19] F. Huszár, “Ml beyond curve fitting: An intro to causal inference and do-calculus,” 2018. [Online]. Available: <https://www.inference.vc/untitled/>
- [20] M. Glymour, J. Pearl, and N. P. Jewell, *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.
- [21] E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard, “On pearl’s hierarchy and the foundations of causal inference,” in *Probabilistic and Causal Inference: The Works of Judea Pearl*, 2022, pp. 507–556.
- [22] V. Papyan, X. Han, and D. L. Donoho, “Prevalence of neural collapse during the terminal phase of deep learning training,” *Proceedings of the National Academy of Sciences*, vol. 117, no. 40, pp. 24 652–24 663, 2020.
- [23] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, “Shortcut learning in deep neural networks,” *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020.

- [24] Y. Wang and M. I. Jordan, “Desiderata for representation learning: A causal perspective,” *arXiv preprint arXiv:2109.03795*, 2021.
- [25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [26] S. Singla and S. Feizi, “Salient imangenet: How to discover spurious features in deep learning?” in *International Conference on Learning Representations*, 2022. [Online]. Available: <https://openreview.net/forum?id=XVPqLyNxSyh>
- [27] I. Shpitser and J. Pearl, “Complete identification methods for the causal hierarchy,” *Journal of Machine Learning Research*, vol. 9, no. 64, pp. 1941–1979, 2008. [Online]. Available: <http://jmlr.org/papers/v9/shpitser08a.html>
- [28] E. Perković, J. Textor, M. Kalisch, and M. H. Maathuis, “Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs,” *Journal of Machine Learning Research*, vol. 18, no. 220, pp. 1–62, 2018. [Online]. Available: <http://jmlr.org/papers/v18/16-319.html>
- [29] L. Gultchin, M. J. Kusner, V. Kanade, and R. Silva, “Differentiable causal backdoor discovery,” in *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, ser. Proceedings of Machine Learning Research, S. Chiappa and R. Calandra, Eds., vol. 108. PMLR, 2020, pp. 3970–3979. [Online]. Available: <http://proceedings.mlr.press/v108/gultchin20a.html>
- [30] M. A. Hernán and J. M. Robins, “Causal inference,” 2010.
- [31] D. Janzing, D. Balduzzi, M. Grosse-Wentrup, and B. Schölkopf, “Quantifying causal influences,” *The Annals of Statistics*, vol. 41, no. 5, pp. 2324–2358, 2013.
- [32] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, “Invariant risk minimization,” *CoRR*, vol. abs/1907.02893, 2019. [Online]. Available: <http://arxiv.org/abs/1907.02893>
- [33] Z. Shen, J. Liu, Y. He, X. Zhang, R. Xu, H. Yu, and P. Cui, “Towards out-of-distribution generalization: A survey,” *arXiv preprint arXiv:2108.13624*, 2021.
- [34] S. Beery, G. Van Horn, and P. Perona, “Recognition in terra incognita,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 456–473.
- [35] J. B. Tenenbaum and W. T. Freeman, “Separating style and content,” in *Advances in Neural Information Processing Systems 9, NIPS, Denver, CO, USA, December 2-5, 1996*, M. Mozer, M. I. Jordan, and T. Petsche, Eds. MIT Press, 1996, pp. 662–668. [Online]. Available: <http://papers.nips.cc/paper/1290-separating-style-and-content>

- [36] M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour, and B. Schölkopf, “Domain adaptation with conditional transferable components,” in *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, ser. JMLR Workshop and Conference Proceedings, M. Balcan and K. Q. Weinberger, Eds., vol. 48. JMLR.org, 2016, pp. 2839–2848. [Online]. Available: <http://proceedings.mlr.press/v48/gong16.html>
- [37] C. Heinze-Deml and N. Meinshausen, “Conditional variance penalties and domain shift robustness,” *Mach. Learn.*, vol. 110, no. 2, pp. 303–348, 2021. [Online]. Available: <https://doi.org/10.1007/s10994-020-05924-1>
- [38] J. Peters, P. Bühlmann, and N. Meinshausen, “Causal inference by using invariant prediction: identification and confidence intervals,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 78, no. 5, pp. 947–1012, 2016.
- [39] J. Mitrovic, B. McWilliams, J. C. Walker, L. H. Buesing, and C. Blundell, “Representation learning via invariant causal mechanisms,” in *International Conference on Learning Representations*, 2021.
- [40] M. Ilse, J. M. Tomczak, and P. Forré, “Selecting data augmentation for simulating interventions,” in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 4555–4562. [Online]. Available: <http://proceedings.mlr.press/v139/ilse21a.html>
- [41] J. V. Kügelgen, Y. Sharma, L. Gresele, W. Brendel, B. Schölkopf, M. Besserve, and F. Locatello, “Self-supervised learning with data augmentations provably isolates content from style,” in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: https://openreview.net/forum?id=4pf_pOo0Dt
- [42] D. Kaushik, A. Setlur, E. H. Hovy, and Z. C. Lipton, “Explaining the efficacy of counterfactually augmented data,” in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. [Online]. Available: <https://openreview.net/forum?id=HHiiQKWsOcV>
- [43] D. Teney, E. Abbasnejad, and A. van den Hengel, “Learning what makes a difference from counterfactual examples and gradient supervision,” in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part X*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12355. Springer, 2020, pp. 580–599. [Online]. Available: https://doi.org/10.1007/978-3-030-58607-2_34

- [44] C. Mao, A. Cha, A. Gupta, H. Wang, J. Yang, and C. Vondrick, “Generative interventions for causal learning,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 3947–3956. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Mao_Generative_Interventions_for_Causal_Learning_CVPR_2021_paper.html
- [45] E. Hätkönen, A. Hertzmann, J. Lehtinen, and S. Paris, “Ganspace: Discovering interpretable gan controls,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 9841–9850, 2020.
- [46] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, vol. 25, 2012.
- [47] C. Mao, K. Xia, J. Wang, H. Wang, J. Yang, E. Bareinboim, and C. Vondrick, “Causal transportability for visual recognition,” *arXiv preprint arXiv:2204.12363*, 2022.
- [48] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *ICML*, 2020.
- [49] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, “Unsupervised learning of visual features by contrasting cluster assignments,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 9912–9924, 2020.
- [50] K. Xiao, L. Engstrom, A. Ilyas, and A. Madry, “Noise or signal: The role of image backgrounds in object recognition,” *arXiv preprint arXiv:2006.09994*, 2020.
- [51] V. Veitch, A. D’Amour, S. Yadlowsky, and J. Eisenstein, “Counterfactual invariance to spurious correlations in text classification,” in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: <https://openreview.net/forum?id=BdKxQp0iBi8>
- [52] S. C. Mouli and B. Ribeiro, “Asymmetry learning for counterfactually-invariant classification in OOD tasks,” in *International Conference on Learning Representations*, 2022. [Online]. Available: <https://openreview.net/forum?id=avgclfZ2211>
- [53] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” in *NeurIPS*, 2020.
- [54] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, “Self-supervised learning for medical image analysis using image context restoration,” *Medical image analysis*, vol. 58, p. 101539, 2019.

- [55] H. Wang, J. Kaddour, S. Liu, J. Tang, M. Kusner, J. Lasenby, and Q. Liu, “Evaluating self-supervised learning for molecular graph embeddings,” 2022. [Online]. Available: <https://arxiv.org/abs/2206.08005>
- [56] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. B. Girshick, “Masked autoencoders are scalable vision learners,” *arXiv:2111.06377*, 2021.
- [57] Z. Wu, Y. Xiong, S. Yu, and D. Lin, “Unsupervised feature learning via non-parametric instance-level discrimination,” 2018. [Online]. Available: <https://arxiv.org/abs/1805.01978>
- [58] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar *et al.*, “Bootstrap your own latent: A new approach to self-supervised learning,” in *NeurIPS*, 2020.
- [59] Y. Tian, D. Krishnan, and P. Isola, “Contrastive multiview coding,” in *European conference on computer vision*. Springer, 2020, pp. 776–794.
- [60] N. Tomasev, I. Bica, B. McWilliams, L. Buesing, R. Pascanu, C. Blundell, and J. Mitrovic, “Pushing the limits of self-supervised resnets: Can we outperform supervised learning without labels on imagenet?” 2022. [Online]. Available: <https://arxiv.org/abs/2201.05119>
- [61] G. Chen, J. Li, J. Lu, and J. Zhou, “Human trajectory prediction via counterfactual analysis,” 2021. [Online]. Available: <https://arxiv.org/abs/2107.14202>
- [62] Y. Niu, K. Tang, H. Zhang, Z. Lu, X.-S. Hua, and J.-R. Wen, “Counterfactual vqa: A cause-effect look at language bias,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 12700–12710.
- [63] Y. Rao, G. Chen, J. Lu, and J. Zhou, “Counterfactual attention learning for fine-grained visual categorization and re-identification,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1025–1034.
- [64] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balasubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao *et al.*, “Wilds: A benchmark of in-the-wild distribution shifts,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 5637–5664.
- [65] E. Rosenfeld, P. K. Ravikumar, and A. Risteski, “The risks of invariant risk minimization,” in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=BbNIbVPJ-42>
- [66] P. Kamath, A. Tangella, D. Sutherland, and N. Srebro, “Does invariant risk minimization capture invariance?” in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 4069–4077.

- [67] K. Ahuja, K. Shanmugam, K. Varshney, and A. Dhurandhar, “Invariant risk minimization games,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 145–155.
- [68] K. Ahuja, E. Caballero, D. Zhang, J.-C. Gagnon-Audet, Y. Bengio, I. Mitliagkas, and I. Rish, “Invariance principle meets information bottleneck for out-of-distribution generalization,” *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [69] N. Tishby, F. C. Pereira, and W. Bialek, “The information bottleneck method,” *arXiv preprint physics/0004057*, 2000.
- [70] D. Krueger, E. Caballero, J. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. L. Priol, and A. C. Courville, “Out-of-distribution generalization via risk extrapolation (rex),” in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 5815–5826. [Online]. Available: <http://proceedings.mlr.press/v139/krueger21a.html>
- [71] Y. Jiang and V. Veitch, “Invariant and transportable representations for anti-causal domain shifts,” *arXiv preprint arXiv:2207.01603*, 2022.
- [72] H. Wang, H. Si, B. Li, and H. Zhao, “Provable domain generalization via invariant-feature subspace recovery,” in *International Conference on Machine Learning*. PMLR, 2022, pp. 23018–23033. [Online]. Available: <https://proceedings.mlr.press/v162/wang22x.html>
- [73] D. Mahajan, S. Tople, and A. Sharma, “Domain generalization using causal matching,” in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 7313–7324. [Online]. Available: <https://proceedings.mlr.press/v139/mahajan21b.html>
- [74] X. Sun, B. Wu, X. Zheng, C. Liu, W. Chen, T. Qin, and T.-Y. Liu, “Recovering latent causal factor for generalization to distributional shifts,” in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: <https://openreview.net/forum?id=go3GvM7aFD>
- [75] C. Liu, X. Sun, J. Wang, T. Li, T. Qin, W. Chen, and T.-Y. Liu, “Learning causal semantic representation for out-of-distribution prediction,” 11 2020.
- [76] C. Lu, Y. Wu, J. M. Hernández-Lobato, and B. Schölkopf, “Invariant causal representation learning for out-of-distribution generalization,” in *International Conference on Learning Representations*, 2022. [Online]. Available: <https://openreview.net/forum?id=-e4EXDWXnSn>
- [77] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman, *Causation, prediction, and search*. MIT press, 2000.

- [78] Y. Atzmon, F. Kreuk, U. Shalit, and G. Chechik, “A causal view of compositional zero-shot recognition,” *arXiv:2006.14610 [cs]*, Nov. 2020, arXiv: 2006.14610. [Online]. Available: <http://arxiv.org/abs/2006.14610>
- [79] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, “Clevr: A diagnostic dataset for compositional language and elementary visual reasoning,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2901–2910.
- [80] G. Parascandolo, N. Kilbertus, M. Rojas-Carulla, and B. Schölkopf, “Learning independent causal mechanisms,” in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 4036–4044. [Online]. Available: <https://proceedings.mlr.press/v80/parascandolo18a.html>
- [81] A. Goyal, A. Lamb, J. Hoffmann, S. Sodhani, S. Levine, Y. Bengio, and B. Schölkopf, “Recurrent independent mechanisms,” in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=mLcmdlEUxy->
- [82] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [83] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2017. [Online]. Available: <https://arxiv.org/abs/1706.03762>
- [84] K. Madan, N. R. Ke, A. Goyal, B. Schölkopf, and Y. Bengio, “Fast and slow learning of recurrent independent mechanisms,” 2021. [Online]. Available: <https://openreview.net/forum?id=Lc28QAB4ypz>
- [85] Z. Yue, Q. Sun, X.-S. Hua, and H. Zhang, “Transporting causal mechanisms for unsupervised domain adaptation,” 2021. [Online]. Available: <https://arxiv.org/abs/2107.11055>
- [86] J. Pearl and E. Bareinboim, “External validity: From do-calculus to transportability across populations,” *Statistical Science*, vol. 29, no. 4, nov 2014.
- [87] T. Teshima, I. Sato, and M. Sugiyama, “Few-shot domain adaptation by causal mechanism transfer,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 9458–9469.
- [88] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.
- [89] E. Grant, C. Finn, S. Levine, T. Darrell, and T. L. Griffiths, “Recasting gradient-based meta-learning as hierarchical bayes,” in *6th International*

- Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings.* OpenReview.net, 2018. [Online]. Available: https://openreview.net/forum?id=BJ_UL-k0b
- [90] V. Mikulik, G. Delétang, T. McGrath, T. Genewein, M. Martic, S. Legg, and P. Ortega, “Meta-trained agents implement bayes-optimal agents,” *Advances in neural information processing systems*, vol. 33, pp. 18 691–18 703, 2020.
 - [91] J. Kaddour, S. Sæmundsson, and M. P. Deisenroth, “Probabilistic active meta-learning,” in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/hash/ef0d17b3bdb4ee2aa741ba28c7255c53-Abstract.html>
 - [92] K. Sohn, H. Lee, and X. Yan, “Learning structured output representation using deep conditional generative models,” in *NIPS*, 2015.
 - [93] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, “Score-based generative modeling through stochastic differential equations,” *arXiv preprint arXiv:2011.13456*, 2020.
 - [94] N. Pawlowski, D. Coelho de Castro, and B. Glocker, “Deep structural causal models for tractable counterfactual inference,” *Advances in Neural Information Processing Systems*, vol. 33, 2020.
 - [95] A. Sauer and A. Geiger, “Counterfactual generative networks,” in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=BXewfAYMmJw>
 - [96] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem, “Challenging common assumptions in the unsupervised learning of disentangled representations,” in *international conference on machine learning*. PMLR, 2019, pp. 4114–4124.
 - [97] I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen, “Variational autoencoders and nonlinear ica: A unifying framework,” in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 2207–2217.
 - [98] P. Sanchez-Martin, M. Rateike, and I. Valera, “Vaca: Design of variational graph autoencoders for interventional and counterfactual queries.” *arXiv*, 2021. [Online]. Available: <https://arxiv.org/abs/2110.14690>
 - [99] H. Kim, S. Shin, J. Jang, K. Song, W. Joo, W. Kang, and I.-C. Moon, “Counterfactual fairness with disentangled causal effect variational autoencoder,” 2020. [Online]. Available: <https://arxiv.org/abs/2011.11878>
 - [100] X. Shen, F. Liu, H. Dong, Q. Lian, Z. Chen, and T. Zhang, “Disentangled generative causal representation learning,” *arXiv preprint arXiv:2010.02637*, 2020.

- [101] P. Sanchez and S. A. Tsaftaris, “Diffusion causal models for counterfactual estimation,” *arXiv preprint arXiv:2202.10166*, 2022.
- [102] F. Locatello, B. Poole, G. Rätsch, B. Schölkopf, O. Bachem, and M. Tschannen, “Weakly-supervised disentanglement without compromises,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 6348–6359.
- [103] I. Kobyzev, S. J. Prince, and M. A. Brubaker, “Normalizing flows: An introduction and review of current methods,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 3964–3979, nov 2021.
- [104] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” 2013. [Online]. Available: <https://arxiv.org/abs/1312.6114>
- [105] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Advances in neural information processing systems*, vol. 27, 2014.
- [106] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.
- [107] P. Dhariwal and A. Nichol, “Diffusion models beat gans on image synthesis,” *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [108] A. Brock, J. Donahue, and K. Simonyan, “Large scale gan training for high fidelity natural image synthesis,” *arXiv preprint arXiv:1809.11096*, 2018.
- [109] M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao, and J. Wang, “Causalvae: Disentangled representation learning via neural structural causal models,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 9593–9602. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Yang_CausalVAE_Disentangled_Representation_Learning_via_Neural_Structural_Causal_Models_CVPR_2021_paper.html
- [110] J. Brehmer, P. De Haan, P. Lippe, and T. Cohen, “Weakly supervised causal representation learning,” *arXiv preprint arXiv:2203.16437*, 2022.
- [111] P. Lippe, S. Magliacane, S. Löwe, Y. M. Asano, T. Cohen, and E. Gavves, “Citris: Causal identifiability from temporal intervened sequences,” *arXiv preprint arXiv:2202.03169*, 2022.
- [112] P. Lippe, S. Magliacane, S. Löwe, Y. M. Asano, T. Cohen, and E. Gavves, “icitrис: Causal representation learning for instantaneous temporal effects,” 2022. [Online]. Available: <https://arxiv.org/abs/2206.06169>
- [113] K. Chalupka, F. Eberhardt, and P. Perona, “Causal feature learning: an overview,” *Behaviormetrika*, vol. 44, no. 1, pp. 137–164, 2017.

- [114] S. Beckers and J. Y. Halpern, “Abstracting causal models,” in *Proceedings of the aaai conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 2678–2685.
- [115] S. Beckers, F. Eberhardt, and J. Y. Halpern, “Approximate causal abstractions,” in *Uncertainty in Artificial Intelligence*. PMLR, 2020, pp. 606–615.
- [116] D. Kinney and D. Watson, “Causal feature learning for utility-maximizing agents,” in *International conference on probabilistic graphical models*. PMLR, 2020, pp. 257–268.
- [117] L. Gresele, P. K. Rubenstein, A. Mehrjou, F. Locatello, and B. Schölkopf, “The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ica,” in *Uncertainty in Artificial Intelligence*. PMLR, 2020, pp. 217–227.
- [118] T. Cohen, “Towards a grounded theory of causation for embodied ai,” *arXiv preprint arXiv:2206.13973*, 2022.
- [119] K. Sachs, O. Perez, D. Pe’er, D. A. Lauffenburger, and G. P. Nolan, “Causal protein-signaling networks derived from multiparameter single-cell data,” *Science*, vol. 308, no. 5721, pp. 523–529, 2005.
- [120] N. Guelzim, S. Bottani, P. Bourgine, and F. Képès, “Topological and causal structure of the yeast transcriptional regulatory network,” *Nature genetics*, vol. 31, no. 1, pp. 60–63, 2002.
- [121] D. Chen, Y. Lin, W. Li, P. Li, J. Zhou, and X. Sun, “Measuring and relieving the over-smoothing problem for graph neural networks from the topological view,” 2019. [Online]. Available: <https://arxiv.org/abs/1909.03211>
- [122] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.
- [123] A. Hernández-García and P. König, “Data augmentation instead of explicit regularization,” 2018. [Online]. Available: <https://arxiv.org/abs/1806.03852>
- [124] X. Chen, C.-J. Hsieh, and B. Gong, “When vision transformers outperform resnets without pre-training or strong data augmentations,” 2021. [Online]. Available: <https://arxiv.org/abs/2106.01548>
- [125] J. Kaddour, L. Liu, R. Silva, and M. J. Kusner, “A fair comparison of two popular flat minima optimizers: Stochastic weight averaging vs. sharpness-aware minimization,” 2022. [Online]. Available: <https://arxiv.org/abs/2202.00661>
- [126] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, “Autoaugment: Learning augmentation policies from data,” *arXiv preprint arXiv:1805.09501*, 2018.

- [127] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, “RandAugment: Practical automated data augmentation with a reduced search space,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 702–703.
- [128] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” *Artificial intelligence*, vol. 267, pp. 1–38, 2019.
- [129] C. Molnar, *Interpretable Machine Learning*, 2nd ed., 2022. [Online]. Available: christophm.github.io/interpretable-ml-book/
- [130] N. Burkart and M. F. Huber, “A survey on the explainability of supervised machine learning,” *Journal of Artificial Intelligence Research*, vol. 70, pp. 245–317, 2021.
- [131] F. K. Došilović, M. Brčić, and N. Hlupić, “Explainable artificial intelligence: A survey,” in *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE, 2018, pp. 0210–0215.
- [132] S. Verma, J. P. Dickerson, and K. Hines, “Counterfactual explanations for machine learning: A review,” *CoRR*, vol. abs/2010.10596, 2020. [Online]. Available: <https://arxiv.org/abs/2010.10596>
- [133] A. Karimi, G. Barthe, B. Schölkopf, and I. Valera, “A survey of algorithmic recourse: definitions, formulations, solutions, and prospects,” *CoRR*, vol. abs/2010.04050, 2020. [Online]. Available: <https://arxiv.org/abs/2010.04050>
- [134] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” *arXiv preprint arXiv:1312.6034*, 2013.
- [135] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, “Sanity checks for saliency maps,” *Advances in neural information processing systems*, vol. 31, 2018.
- [136] P. Schwab and W. Karlen, “Cxplain: Causal explanations for model interpretation under uncertainty,” in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 10 220–10 230. [Online]. Available: <https://proceedings.neurips.cc/paper/2019/hash/3ab6be46e1d6b21d59a3c3a0b9d0f6ef-Abstract.html>
- [137] M. R. O’Shaughnessy, G. Canal, M. Connor, C. Rozell, and M. A. Davenport, “Generative causal explanations of black-box classifiers,” in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds.,

2020. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/hash/3a93a609b97ec0ab0ff5539eb79ef33a-Abstract.html>
- [138] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in neural information processing systems*, vol. 30, 2017.
- [139] C. Frye, C. Rowat, and I. Feige, “Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability,” in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/hash/0d770c496aa3da6d2c3f2bd19e7b9d6b-Abstract.html>
- [140] D. Janzing, L. Minorics, and P. Blöbaum, “Feature relevance quantification in explainable AI: A causal problem,” in *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, ser. Proceedings of Machine Learning Research, S. Chiappa and R. Calandra, Eds., vol. 108. PMLR, 2020, pp. 2907–2916. [Online]. Available: <http://proceedings.mlr.press/v108/janzing20a.html>
- [141] T. Heskes, E. Sijben, I. G. Bucur, and T. Claassen, “Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models,” in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/hash/32e54441e6382a7fbacbbaf3c450059-Abstract.html>
- [142] H. Chen, J. D. Janizek, S. Lundberg, and S.-I. Lee, “True to the model or true to the data?” 2020. [Online]. Available: <https://arxiv.org/abs/2006.16234>
- [143] Y. Jung, S. Kasiviswanathan, J. Tian, D. Janzing, P. Bloebaum, and E. Bareinboim, “On measuring causal contributions via do-interventions,” in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 10476–10501. [Online]. Available: <https://proceedings.mlr.press/v162/jung22a.html>
- [144] P. Lipton, “Contrastive explanation,” *Royal Institute of Philosophy Supplements*, vol. 27, pp. 247–266, 1990.
- [145] A. Jacovi, S. Swayamdipta, S. Ravfogel, Y. Elazar, Y. Choi, and Y. Goldberg, “Contrastive explanations for model interpretability,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds.

- Association for Computational Linguistics, 2021, pp. 1597–1611. [Online]. Available: <https://doi.org/10.18653/v1/2021.emnlp-main.120>
- [146] S. Verma, J. Dickerson, and K. Hines, “Counterfactual explanations for machine learning: Challenges revisited,” *arXiv preprint arXiv:2106.07756*, 2021.
- [147] A. Karimi, B. Schölkopf, and I. Valera, “Algorithmic recourse: from counterfactual explanations to interventions,” in *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, M. C. Elish, W. Isaac, and R. S. Zemel, Eds. ACM, 2021, pp. 353–362. [Online]. Available: <https://doi.org/10.1145/3442188.3445899>
- [148] M. T. Ribeiro, S. Singh, and C. Guestrin, “" why should i trust you?" explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.
- [149] S. Wachter, B. D. Mittelstadt, and C. Russell, “Counterfactual explanations without opening the black box: Automated decisions and the GDPR,” *CoRR*, vol. abs/1711.00399, 2017. [Online]. Available: <http://arxiv.org/abs/1711.00399>
- [150] A. Karimi, B. J. von Kügelgen, B. Schölkopf, and I. Valera, “Algorithmic recourse under imperfect causal knowledge: a probabilistic approach,” in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/hash/02a3c7fb3f489288ae6942498498db20-Abstract.html>
- [151] S. Joshi, O. Koyejo, W. Vijitbenjaronk, B. Kim, and J. Ghosh, “Towards realistic individual recourse and actionable explanations in black-box decision making systems,” *CoRR*, vol. abs/1907.09615, 2019. [Online]. Available: <http://arxiv.org/abs/1907.09615>
- [152] L. Schut, O. Key, R. McGrath, L. Costabello, B. Sacaleanu, M. Corcoran, and Y. Gal, “Generating interpretable counterfactual explanations by implicit minimisation of epistemic and aleatoric uncertainties,” in *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, A. Banerjee and K. Fukumizu, Eds., vol. 130. PMLR, 2021, pp. 1756–1764. [Online]. Available: <http://proceedings.mlr.press/v130/schut21a.html>
- [153] A. Abid, M. Yuksekgonul, and J. Zou, “Meaningfully explaining model mistakes using conceptual counterfactuals,” 2021. [Online]. Available: <https://arxiv.org/abs/2106.12723>

- [154] D. Mahajan, C. Tan, and A. Sharma, “Preserving causal constraints in counterfactual explanations for machine learning classifiers,” *CoRR*, vol. abs/1912.03277, 2019. [Online]. Available: <http://arxiv.org/abs/1912.03277>
- [155] B. Ustun, A. Spangher, and Y. Liu, “Actionable recourse in linear classification,” in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ser. FAT* ’19. New York, NY, USA: Association for Computing Machinery, 2019, p. 10–19. [Online]. Available: <https://doi.org/10.1145/3287560.3287566>
- [156] J. von Kügelgen, N. Agarwal, J. Zeitler, A. Mastouri, and B. Schölkopf, “Algorithmic recourse in partially and fully confounded settings through bounding counterfactual effects,” 2021. [Online]. Available: <https://arxiv.org/abs/2106.11849>
- [157] R. Kommiya Mothilal, D. Mahajan, C. Tan, and A. Sharma, “Towards unifying feature attribution and counterfactual explanations: Different means to the same end,” in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, pp. 652–663.
- [158] S. Galhotra, R. Pradhan, and B. Salimi, “Feature attribution and recourse via probabilistic contrastive counterfactuals,” 2021.
- [159] E. Albini, J. Long, D. Dervovic, and D. Magazzeni, “Counterfactual shapley additive explanations,” *arXiv preprint arXiv:2110.14270*, 2021.
- [160] D. S. Watson, L. Gultchin, A. Taly, and L. Floridi, “Local explanations via necessity and sufficiency: unifying theory and practice,” in *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI 2021, Virtual Event, 27-30 July 2021*, ser. Proceedings of Machine Learning Research, C. P. de Campos, M. H. Maathuis, and E. Quaeghebeur, Eds., vol. 161. AUAI Press, 2021, pp. 1382–1392. [Online]. Available: <https://proceedings.mlr.press/v161/watson21a.html>
- [161] D. Watson, “Rational shapley values,” in *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1083–1094.
- [162] H. Fokkema, R. de Heide, and T. van Erven, “Attribution-based explanations that provide recourse cannot be robust,” 2022. [Online]. Available: <https://arxiv.org/abs/2205.15834>
- [163] A. Artelt and B. Hammer, “On the computation of counterfactual explanations—a survey,” *arXiv preprint arXiv:1911.07749*, 2019.
- [164] D. Mahajan, C. Tan, and A. Sharma, “Preserving causal constraints in counterfactual explanations for machine learning classifiers,” *arXiv preprint arXiv:1912.03277*, 2019.

- [165] M. Pawelczyk, T. Datta, J. van-den Heuvel, G. Kasneci, and H. Lakkaraju, “Algorithmic recourse in the face of noisy human responses,” 2022. [Online]. Available: <https://arxiv.org/abs/2203.06768>
- [166] K. Rawal, E. Kamar, and H. Lakkaraju, “Algorithmic recourse in the wild: Understanding the impact of data and model shifts,” *arXiv preprint arXiv:2012.11788*, 2020.
- [167] A. Ghorbani, M. Kim, and J. Zou, “A distributional framework for data valuation,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 3535–3544.
- [168] U. Aïvodji, A. Bolot, and S. Gambs, “Model extraction from counterfactual explanations,” *arXiv preprint arXiv:2009.01884*, 2020.
- [169] R. Shokri, M. Strobel, and Y. Zick, “On the privacy risks of model explanations,” in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, pp. 231–241.
- [170] Y.-B. He and Z. Geng, “Active learning of causal networks with intervention experiments and optimal designs,” *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2523–2547, 2008.
- [171] D. Slack, A. Hilgard, H. Lakkaraju, and S. Singh, “Counterfactual explanations can be manipulated,” *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [172] R. Naidu, A. Priyanshu, A. Kumar, S. Kotti, H. Wang, and F. Mireshghallah, “When differential privacy meets interpretability: A case study,” *arXiv preprint arXiv:2106.13203*, 2021.
- [173] K. Makhlof, S. Zhioua, and C. Palamidessi, “Survey on causal-based machine learning fairness notions,” *arXiv preprint arXiv:2010.09553*, 2020.
- [174] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. fairmlbook.org, 2019, <http://www.fairmlbook.org>.
- [175] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva, “Counterfactual fairness,” in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 4066–4076. [Online]. Available: <https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html>
- [176] R. Nabi and I. Shpitser, “Fair inference on outcomes,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.
- [177] S. Chiappa, “Path-specific counterfactual fairness,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 7801–7808.

- [178] Y. Wu, L. Zhang, X. Wu, and H. Tong, “Pc-fairness: A unified framework for measuring causality-based fairness,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [179] J. Zhang and E. Bareinboim, “Fairness in decision-making—the causal explanation formula,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.
- [180] T. J. VanderWeele and W. R. Robinson, “On causal interpretation of race in regressions adjusting for confounding and mediating variables,” *Epidemiology (Cambridge, Mass.)*, vol. 25, no. 4, p. 473, 2014.
- [181] L. Hu and I. Kohler-Hausmann, “What’s sex got to do with machine learning?” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ser. FAT* ’20. New York, NY, USA: Association for Computing Machinery, 2020, p. 513. [Online]. Available: <https://doi.org/10.1145/3351095.3375674>
- [182] N. Kilbertus, “Beyond traditional assumptions in fair machine learning,” Ph.D. dissertation, University of Cambridge, UK, Sep. 2020, (Cambridge-Tübingen-Fellowship).
- [183] N. Kilbertus, M. Rojas Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, “Avoiding discrimination through causal reasoning,” *Advances in neural information processing systems*, vol. 30, 2017.
- [184] B. Salimi, L. Rodriguez, B. Howe, and D. Suciu, “Interventional fairness: Causal database repair for algorithmic fairness,” in *Proceedings of the 2019 International Conference on Management of Data*, 2019, pp. 793–810.
- [185] H. Singh, R. Singh, V. Mhasawade, and R. Chunara, “Fairness violations and mitigation under covariate shift,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 3–13.
- [186] J. Schrouff, N. Harris, O. Koyejo, I. Alabdulmohsin, E. Schnider, K. Opsahl-Ong, A. Brown, S. Roy, D. Mincu, C. Chen *et al.*, “Maintaining fairness across distribution shift: do we have viable solutions for real-world applications?” *arXiv preprint arXiv:2202.01034*, 2022.
- [187] D. Slack, S. A. Friedler, and E. Givental, “Fairness warnings and fair-maml: learning fairly with minimal data,” in *FAT* ’20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, M. Hildebrandt, C. Castillo, L. E. Celis, S. Ruggieri, L. Taylor, and G. Zanfir-Fortuna, Eds. ACM, 2020, pp. 200–209. [Online]. Available: <https://doi.org/10.1145/3351095.3372839>
- [188] A. Subbaswamy, R. Adams, and S. Saria, “Evaluating model robustness and stability to dataset shift,” in *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021*,

- Virtual Event*, ser. Proceedings of Machine Learning Research, A. Banerjee and K. Fukumizu, Eds., vol. 130. PMLR, 2021, pp. 2611–2619. [Online]. Available: <http://proceedings.mlr.press/v130/subbaswamy21a.html>
- [189] M. E. Guy and S. A. McCandless, “Social equity: Its legacy, its promise,” *Public Administration Review*, vol. 72, no. s1, pp. S5–S13, 2012.
 - [190] J. McGinnis Johnson, “Race and social equity: A nervous area of government,” *Equality, Diversity and Inclusion: An International Journal*, vol. 34, pp. 262–264, 03 2015.
 - [191] J. G. Richens, R. Beard, and D. H. Thompson, “First do no harm: counterfactual objective functions for safe & ethical ai,” 2022. [Online]. Available: <https://arxiv.org/abs/2204.12993>
 - [192] V. Gupta, P. Nokhiz, C. D. Roy, and S. Venkatasubramanian, “Equalizing recourse across groups,” 2019. [Online]. Available: <https://arxiv.org/abs/1909.03166>
 - [193] J. von Kügelgen, A.-H. Karimi, U. Bhatt, I. Valera, A. Weller, and B. Schölkopf, “On the fairness of causal algorithmic recourse,” *arXiv preprint arXiv:2010.06529*, 2020.
 - [194] W. Huan, Y. Wu, L. Zhang, and X. Wu, “Fairness through equality of effort,” in *Companion Proceedings of the Web Conference 2020*, 2020, pp. 743–751.
 - [195] Y. Wu, L. Zhang, and X. Wu, “Counterfactual fairness: Unidentification, bound and algorithm,” in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.
 - [196] I. Kohler-Hausmann, “Eddie murphy and the dangers of counterfactual causal thinking about detecting racial discrimination,” *Nw. UL Rev.*, vol. 113, p. 1163, 2018.
 - [197] L. Hu, “Causation in the social world,” Ph.D. dissertation, Harvard University, 2022.
 - [198] L. Hu and I. Kohler-Hausmann, “What’s sex got to do with machine learning?” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ser. FAT* ’20. New York, NY, USA: Association for Computing Machinery, 2020, p. 513. [Online]. Available: <https://doi.org/10.1145/3351095.3375674>
 - [199] A. Kasirzadeh and A. Smart, “The use and misuse of counterfactuals in ethical machine learning,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 228–236.
 - [200] J. Kleinberg, S. Mullainathan, and M. Raghavan, “Inherent trade-offs in the fair determination of risk scores,” *arXiv preprint arXiv:1609.05807*, 2016.

- [201] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, “Algorithmic decision making and the cost of fairness,” in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 797–806.
- [202] A. Chouldechova, “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments,” *Big data*, vol. 5, no. 2, pp. 153–163, 2017.
- [203] L. Gultchin, V. Cohen-Addad, S. Giffard-Roisin, V. Kanade, and F. Mallmann-Trenn, “Beyond impossibility: Balancing sufficiency, separation and accuracy,” *CoRR*, vol. abs/2205.12327, 2022. [Online]. Available: <https://doi.org/10.48550/arXiv.2205.12327>
- [204] H. Nilforoshan, J. D. Gaebler, R. Shroff, and S. Goel, “Causal conceptions of fairness and their consequences,” in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 16848–16887. [Online]. Available: <https://proceedings.mlr.press/v162/nilforoshan22a.html>
- [205] C. Ashurst, R. Carey, S. Chiappa, and T. Everitt, “Why fair labels can yield unfair predictions: Graphical conditions for introduced unfairness,” in *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelfth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 2022, pp. 9494–9503. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/21182>
- [206] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: A Bradford Book, 2018.
- [207] S. Zhu, I. Ng, and Z. Chen, “Causal discovery with reinforcement learning,” *arXiv preprint arXiv:1906.04477*, 2019.
- [208] L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson, “Counterfactual reasoning and learning systems: The example of computational advertising.” *Journal of Machine Learning Research*, vol. 14, no. 11, 2013.
- [209] E. Bareinboim, A. Forney, and J. Pearl, “Bandits with unobserved confounders: A causal approach,” *Advances in Neural Information Processing Systems*, vol. 28, 2015.
- [210] S. J. Gershman, “Reinforcement learning and causal models,” *The Oxford handbook of causal reasoning*, vol. 1, p. 295, 2017.
- [211] C. Szepesvari, “Causality from the perspective of reinforcement learning,” 2018, machine Learning for Causal Inference, Counterfactual Prediction, and

- Autonomous Action (CausalML). [Online]. Available: <https://sites.google.com/site/faim18wscausalml/invited-talks>
- [212] J. Bannon, B. Windsor, W. Song, and T. Li, “Causality and batch reinforcement learning: Complementary approaches to planning in unknown domains,” *arXiv preprint arXiv:2006.02579*, 2020.
 - [213] E. Bareinboim, “Causal reinforcement learning,” 2020, iCML. [Online]. Available: <https://crl.causalai.net/>
 - [214] S. Weichwald, S. W. Mogensen, T. E. Lee, D. Baumann, O. Kroemer, I. Guyon, S. Trimpe, J. Peters, and N. Pfister, “Learning by doing: Controlling a dynamical system using causality, control, and reinforcement learning,” *arXiv preprint arXiv:2202.06052*, 2022.
 - [215] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel *et al.*, “Mastering chess and shogi by self-play with a general reinforcement learning algorithm,” *arXiv preprint arXiv:1712.01815*, 2017.
 - [216] R. S. Sutton, “The bitter lesson,” 2019. [Online]. Available: <http://www.incompleteideas.net/IncIdeas/BitterLesson.html>
 - [217] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, “Double/debiased machine learning for treatment and structural parameters,” *The Econometrics Journal*, vol. 21, no. 1, pp. C1–C68, 01 2018.
 - [218] E. H. Kennedy, “Optimal doubly robust estimation of heterogeneous causal effects,” *arXiv preprint arXiv:2004.14497*, 2020.
 - [219] S. Levine, A. Kumar, G. Tucker, and J. Fu, “Offline reinforcement learning: Tutorial, review, and perspectives on open problems,” *arXiv preprint arXiv:2005.01643*, 2020.
 - [220] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.
 - [221] F. Lattimore, T. Lattimore, and M. D. Reid, “Causal bandits: Learning good interventions via causal inference,” in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, Eds., 2016, pp. 1181–1189. [Online]. Available: <https://proceedings.neurips.cc/paper/2016/hash/b4288d9c0ec0a1841b3b3728321e7088-Abstract.html>
 - [222] S. Bubeck, R. Munos, and G. Stoltz, “Pure exploration in multi-armed bandits problems,” in *International conference on Algorithmic learning theory*. Springer, 2009, pp. 23–37.

- [223] V. Gabillon, M. Ghavamzadeh, and A. Lazaric, “Best arm identification: A unified approach to fixed budget and fixed confidence,” *Advances in Neural Information Processing Systems*, vol. 25, 2012.
- [224] J.-Y. Audibert, S. Bubeck, and R. Munos, “Best arm identification in multi-armed bandits.” in *COLT*. Citeseer, 2010, pp. 41–53.
- [225] S. Lee and E. Bareinboim, “Structural causal bandits: Where to intervene?” in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018. [Online]. Available: <https://proceedings.neurips.cc/paper/2018/file/c0a271bc0ecb776a094786474322cb82-Paper.pdf>
- [226] A. A. W. M. de Kroon, D. Belgrave, and J. M. Mooij, “Causal discovery for causal bandits utilizing separating sets,” 2020. [Online]. Available: <https://arxiv.org/abs/2009.07916>
- [227] Y. Lu, A. Meisami, and A. Tewari, “Causal bandits with unknown graph structure,” in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: <https://openreview.net/forum?id=9-XhLobA4z>
- [228] R. Silva, “Observational-interventional priors for dose-response learning,” in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, Eds., 2016, pp. 1561–1569. [Online]. Available: <https://proceedings.neurips.cc/paper/2016/hash/aff1621254f7c1be92f64550478c56e6-Abstract.html>
- [229] V. Aglietti, X. Lu, A. Paleyes, and J. González, “Causal bayesian optimization,” in *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, ser. Proceedings of Machine Learning Research, S. Chiappa and R. Calandra, Eds., vol. 108. PMLR, 2020, pp. 3155–3164. [Online]. Available: <http://proceedings.mlr.press/v108/aglietti20a.html>
- [230] D. R. Jones, M. Schonlau, and W. J. Welch, “Efficient global optimization of expensive black-box functions,” *Journal of Global optimization*, vol. 13, no. 4, pp. 455–492, 1998.
- [231] S. Saengkyongam, N. Thams, J. Peters, and N. Pfister, “Invariant policy learning: A causal perspective,” *CoRR*, vol. abs/2106.00808, 2021. [Online]. Available: <https://arxiv.org/abs/2106.00808>
- [232] Y. Lu, A. Meisami, and A. Tewari, “Efficient reinforcement learning with prior causal knowledge,” in *First Conference on Causal Learning and Reasoning*, 2022. [Online]. Available: <https://openreview.net/forum?id=xzJWqJadOG>
- [233] T. M. Moerland, J. Broekens, and C. M. Jonker, “Model-based reinforcement learning: A survey,” *arXiv preprint arXiv:2006.16712*, 2020.

- [234] D. Ha and J. Schmidhuber, “World models,” *arXiv preprint arXiv:1803.10122*, 2018.
- [235] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel *et al.*, “Mastering atari, go, chess and shogi by planning with a learned model,” *Nature*, vol. 588, no. 7839, pp. 604–609, 2020.
- [236] D. J. Rezende, I. Danihelka, G. Papamakarios, N. R. Ke, R. Jiang, T. Weber, K. Gregor, H. Merzic, F. Viola, J. Wang *et al.*, “Causally correct partial models for reinforcement learning,” *arXiv preprint arXiv:2002.02836*, 2020.
- [237] M. Li, M. Yang, F. Liu, X. Chen, Z. Chen, and J. Wang, “Causal world models by unsupervised deconfounding of physical dynamics,” *arXiv preprint arXiv:2012.14228*, 2020.
- [238] D. Chapman and L. P. Kaelbling, “Input generalization in delayed reinforcement learning: An algorithm and performance comparisons.” in *Ijcai*, vol. 91. Citeseer, 1991, pp. 726–731.
- [239] A. K. McCallum, *Reinforcement learning with selective perception and hidden state*. University of Rochester, 1996.
- [240] Z. Wang, X. Xiao, Z. Xu, Y. Zhu, and P. Stone, “Causal dynamics learning for task-independent state abstraction,” 2022. [Online]. Available: <https://arxiv.org/abs/2206.13452>
- [241] A. A. Mastakouri, B. Schölkopf, and D. Janzing, “Necessary and sufficient conditions for causal feature selection in time series with latent common causes,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 7502–7511.
- [242] L. Weng, “Meta reinforcement learning,” *lilianweng.github.io*, 2019. [Online]. Available: <https://lilianweng.github.io/posts/2019-06-23-meta-rl/>
- [243] S. A. Sontakke, A. Mehrjou, L. Itti, and B. Schölkopf, “Causal curiosity: RL agents discovering self-supervised experiments for causal representation learning,” in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 9848–9858. [Online]. Available: <http://proceedings.mlr.press/v139/sontakke21a.html>
- [244] F. Doshi-Velez and G. Konidaris, “Hidden parameter markov decision processes: A semiparametric regression approach for discovering latent task parametrizations,” in *IJCAI: proceedings of the conference*, vol. 2016. NIH Public Access, 2016, p. 1432.

- [245] K. Rakelly, A. Zhou, C. Finn, S. Levine, and D. Quillen, “Efficient off-policy meta-reinforcement learning via probabilistic context variables,” in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 2019, pp. 5331–5340. [Online]. Available: <http://proceedings.mlr.press/v97/rakelly19a.html>
- [246] A. Zhang, C. Lyle, S. Sodhani, A. Filos, M. Kwiatkowska, J. Pineau, Y. Gal, and D. Precup, “Invariant causal prediction for block mdps,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 11214–11224.
- [247] M. Tomar, A. Zhang, R. Calandra, M. E. Taylor, and J. Pineau, “Model-invariant state abstractions for model-based reinforcement learning,” 2021.
- [248] K. Kansky, T. Silver, D. A. Mély, M. Eldawy, M. Lázaro-Gredilla, X. Lou, N. Dorfman, S. Sidor, S. Phoenix, and D. George, “Schema networks: Zero-shot transfer with a generative causal model of intuitive physics,” 2017.
- [249] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, “Human-level control through deep reinforcement learning,” *nature*, vol. 518, no. 7540, pp. 529–533, 2015.
- [250] M. Mutti, R. De Santi, E. Rossi, J. F. Calderon, M. Bronstein, and M. Restelli, “Provably efficient causal model-based reinforcement learning for systematic generalization,” *arXiv preprint arXiv:2202.06545*, 2022.
- [251] A. Lei, B. Schölkopf, and I. Posner, “Variational causal dynamics: Discovering modular world models from interventions,” 2022. [Online]. Available: <https://arxiv.org/abs/2206.11131>
- [252] S. Parbhoo, S. Joshi, and F. Doshi-Velez, “Generalizing off-policy evaluation from a causal perspective for sequential decision-making,” 2022.
- [253] L. Buesing, T. Weber, Y. Zwols, N. Heess, S. Racaniere, A. Guez, and J.-B. Lespiau, “Woulda, coulda, shoulda: Counterfactually-guided policy search,” in *International Conference on Learning Representations*, 2019. [Online]. Available: <https://openreview.net/forum?id=BJG0voC9YQ>
- [254] M. Oberst and D. A. Sontag, “Counterfactual off-policy evaluation with gumbel-max structural causal models,” in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 2019, pp. 4881–4890. [Online]. Available: <http://proceedings.mlr.press/v97/oberst19a.html>
- [255] N. Kallus and A. Zhou, “Confounding-robust policy improvement,” in *Advances in Neural Information Processing Systems 31: Annual Conference*

- on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018, pp. 9289–9299. [Online]. Available: <https://proceedings.neurips.cc/paper/2018/hash/3a09a524440d44d7f19870070a5ad42f-Abstract.html>
- [256] N. Kallus and A. Zhou, “Confounding-robust policy evaluation in infinite-horizon reinforcement learning,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 22 293–22 304. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/file/fd4f21f2556dad0ea8b7a5c04eabebda-Paper.pdf>
 - [257] A. Bennett, N. Kallus, L. Li, and A. Mousavi, “Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders,” in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 1999–2007.
 - [258] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, “Continuous control with deep reinforcement learning,” *arXiv preprint arXiv:1509.02971*, 2015.
 - [259] A. Charpentier, R. Elie, and C. Remlinger, “Reinforcement learning in economics and finance,” *Computational Economics*, pp. 1–38, 2021.
 - [260] H. Namkoong, R. Keramati, S. Yadlowsky, and E. Brunskill, “Off-policy policy evaluation for sequential decisions under unobserved confounding,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 819–18 831, 2020.
 - [261] G. Tennenholz, U. Shalit, and S. Mannor, “Off-policy evaluation in partially observable environments,” in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 2020, pp. 10 276–10 283. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/6590>
 - [262] A. Bennett and N. Kallus, “Proximal reinforcement learning: Efficient off-policy evaluation in partially observed markov decision processes,” *CoRR*, vol. abs/2110.15332, 2021. [Online]. Available: <https://arxiv.org/abs/2110.15332>
 - [263] M. Gasse, D. Grasset, G. Gaudron, and P.-Y. Oudeyer, “Causal reinforcement learning using observational and interventional data,” *arXiv preprint arXiv:2106.14421*, 2021.
 - [264] L. Wang, Z. Yang, and Z. Wang, “Provably efficient causal reinforcement learning with confounded observational data,” in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: <https://openreview.net/forum?id=tUeeRzMJZ>

- [265] Y. Chen, L. Xu, C. Gulcehre, T. L. Paine, A. Gretton, N. de Freitas, and A. Doucet, “On instrumental variable regression for deep offline policy evaluation,” 2021. [Online]. Available: <https://arxiv.org/abs/2105.10148>
- [266] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne, “Imitation learning: A survey of learning methods,” *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, pp. 1–35, 2017.
- [267] P. de Haan, D. Jayaraman, and S. Levine, “Causal confusion in imitation learning,” in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 11 693–11 704. [Online]. Available: <https://proceedings.neurips.cc/paper/2019/hash/947018640bf36a2bb609d3557a285329-Abstract.html>
- [268] J. Tien, J. Z.-Y. He, Z. Erickson, A. D. Dragan, and D. Brown, “A study of causal confusion in preference-based reward learning,” *arXiv preprint arXiv:2204.06601*, 2022.
- [269] P. A. Ortega, M. Kunesch, G. Delétang, T. Genewein, J. Grau-Moya, J. Veness, J. Buchli, J. Degrave, B. Piot, J. Perolat, T. Everitt, C. Tallec, E. Parisotto, T. Erez, Y. Chen, S. Reed, M. Hutter, N. de Freitas, and S. Legg, “Shaking the foundations: delusions in sequence models for interaction and control,” 2021.
- [270] J. Zhang, D. Kumor, and E. Bareinboim, “Causal imitation learning with unobserved confounders,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 263–12 274. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/file/8fdd149fcaa7058caccc9c4ad5b0d89a-Paper.pdf>
- [271] D. Kumor, J. Zhang, and E. Bareinboim, “Sequential causal imitation learning with unobserved confounders,” in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: <https://openreview.net/forum?id=o6-k168bBD8>
- [272] I. Bica, D. Jarrett, and M. van der Schaar, “Invariant causal imitation learning for generalizable policies,” in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: <https://openreview.net/forum?id=715E7e6j4gU>
- [273] G. Swamy, S. Choudhury, J. A. Bagnell, and Z. S. Wu, “Causal imitation learning under temporally correlated noise,” 2022. [Online]. Available: <https://arxiv.org/abs/2202.01312>
- [274] A. Hefny, C. Downey, and G. J. Gordon, “Supervised learning for dynamical system learning,” *Advances in neural information processing systems*, vol. 28, 2015.

- [275] A. Y. Ng, H. J. Kim, M. I. Jordan, S. Sastry, and S. Ballianda, “Autonomous helicopter flight via reinforcement learning.” in *NIPS*, vol. 16. Citeseer, 2003.
- [276] P. G. Wright, *Tariff on animal and vegetable oils*. Macmillan Company, New York, 1928.
- [277] T. Desautels, R. Das, J. Calvert, M. Trivedi, C. Summers, D. J. Wales, and A. Ercole, “Prediction of early unplanned intensive care unit readmission in a uk tertiary care hospital: a cross-sectional machine learning approach,” *BMJ open*, vol. 7, no. 9, p. e017199, 2017.
- [278] M. Minsky, “Steps toward artificial intelligence,” *Proceedings of the IRE*, vol. 49, no. 1, pp. 8–30, 1961.
- [279] T. Mesnard, T. Weber, F. Viola, S. Thakoor, A. Saade, A. Harutyunyan, W. Dabney, T. S. Stepleton, N. Heess, A. Guez, E. Moulines, M. Hutter, L. Buesing, and R. Munos, “Counterfactual credit assignment in model-free reinforcement learning,” in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 7654–7664. [Online]. Available: <http://proceedings.mlr.press/v139/mesnard21a.html>
- [280] M. Seitzer, B. Schölkopf, and G. Martius, “Causal influence detection for improving efficiency in reinforcement learning,” *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [281] G. Weiss, *Multiagent systems: a modern approach to distributed artificial intelligence*. MIT press, 1999.
- [282] S. Gronauer and K. Diepold, “Multi-agent deep reinforcement learning: a survey,” *Artificial Intelligence Review*, vol. 55, no. 2, pp. 895–943, 2022.
- [283] Y.-H. Chang, T. Ho, and L. Kaelbling, “All learning is local: Multi-agent learning in global reward games,” *Advances in neural information processing systems*, vol. 16, 2003.
- [284] J. N. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, “Counterfactual multi-agent policy gradients,” in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, S. A. McIlraith and K. Q. Weinberger, Eds. AAAI Press, 2018, pp. 2974–2982. [Online]. Available: <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17193>
- [285] N. Jaques, A. Lazaridou, E. Hughes, Ç. Gülcühre, P. A. Ortega, D. Strouse, J. Z. Leibo, and N. de Freitas, “Social influence as intrinsic motivation for multi-agent deep reinforcement learning,” in *Proceedings of the 36th*

- International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 2019, pp. 3040–3049. [Online]. Available: <http://proceedings.mlr.press/v97/jaques19a.html>
- [286] S. Pitis, E. Creager, and A. Garg, “Counterfactual data augmentation using locally factored dynamics,” in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/hash/294e09f267683c7ddc6cc5134a7e68a8-Abstract.html>
 - [287] L. P. Kaelbling, “Learning to achieve goals,” in *IN PROC. OF IJCAI-93*. Morgan Kaufmann, 1993, pp. 1094–1098.
 - [288] M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, O. Pieter Abbeel, and W. Zaremba, “Hindsight experience replay,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: <https://proceedings.neurips.cc/paper/2017/file/453fadbd8a1a3af50a9df4df899537b5-Paper.pdf>
 - [289] M. Andrychowicz, A. Raichuk, P. Stańczyk, M. Orsini, S. Girgin, R. Marinier, L. Huszenot, M. Geist, O. Pietquin, M. Michalski *et al.*, “What matters in on-policy reinforcement learning? a large-scale empirical study,” *arXiv preprint arXiv:2006.05990*, 2020.
 - [290] M. Laskin, K. Lee, A. Stooke, L. Pinto, P. Abbeel, and A. Srinivas, “Reinforcement learning with augmented data,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 19 884–19 895. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/file/e615c82aba461681ade82da2da38004a-Paper.pdf>
 - [291] R. S. Sutton, “Dyna, an integrated architecture for learning, planning, and reacting,” *ACM Sigart Bulletin*, vol. 2, no. 4, pp. 160–163, 1991.
 - [292] M. Janner, J. Fu, M. Zhang, and S. Levine, “When to trust your model: Model-based policy optimization,” in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 12 498–12 509. [Online]. Available: <https://proceedings.neurips.cc/paper/2019/hash/5faf461eff3099671ad63c6f3f094f7f-Abstract.html>

- [293] C. Lu, B. Huang, K. Wang, J. M. Hernández-Lobato, K. Zhang, and B. Schölkopf, “Sample-efficient reinforcement learning via counterfactual-based data augmentation,” in *Offline Reinforcement Learning - Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS)*, 2020. [Online]. Available: <https://offline-rl-neurips.github.io/pdf/34.pdf>
- [294] A. Jaiswal, W. AbdAlmageed, Y. Wu, and P. Natarajan, “Bidirectional conditional generative adversarial networks,” in *Asian Conference on Computer Vision*. Springer, 2018, pp. 216–232.
- [295] N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*, 1st ed. USA: Oxford University Press, Inc., 2014.
- [296] T. Everitt, G. Lea, and M. Hutter, “AGI safety literature review,” in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, J. Lang, Ed. ijcai.org, 2018, pp. 5441–5449. [Online]. Available: <https://doi.org/10.24963/ijcai.2018/768>
- [297] E. Yudkowsky, “The ai alignment problem: why it is hard, and where to start,” *Symbolic Systems Distinguished Speaker*, 2016.
- [298] B. Christian, *The alignment problem: How can machines learn human values?* Atlantic Books, 2021.
- [299] J. Miller, S. Milli, and M. Hardt, “Strategic classification is causal modeling in disguise,” 2019. [Online]. Available: <https://arxiv.org/abs/1910.10362>
- [300] T. Everitt, R. Carey, L. Hammond, J. Fox, E. Langlois, and S. Legg, “Progress on causal influence diagrams,” Jun 2021. [Online]. Available: <https://deepmindsafetyresearch.medium.com/progress-on-causal-influence-diagrams-a7a32180b0d1#4e50>
- [301] T. Everitt, R. Carey, E. D. Langlois, P. A. Ortega, and S. Legg, “Agent incentives: A causal perspective,” in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 2021, pp. 11487–11495. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/17368>
- [302] A. Jern and C. Kemp, “Capturing mental state reasoning with influence diagrams,” in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 33, no. 33, 2011.
- [303] M. Kleiman-Weiner, T. Gerstenberg, S. Levine, and J. B. Tenenbaum, “Inference of intention and permissibility in moral decision making.” in *CogSci*, 2015.

- [304] R. A. Howard, “Information value theory,” *IEEE Transactions on systems science and cybernetics*, vol. 2, no. 1, pp. 22–26, 1966.
- [305] C. Evans and A. Kasirzadeh, “User tampering in reinforcement learning recommender systems,” *CoRR*, vol. abs/2109.04083, 2021. [Online]. Available: <https://arxiv.org/abs/2109.04083>
- [306] A. P. Dawid, “Influence diagrams for causal modelling and inference,” *International Statistical Review*, vol. 70, no. 2, pp. 161–189, 2002.
- [307] E. Fagiuoli and M. Zaffalon, “A note about redundancy in influence diagrams,” *International Journal of Approximate Reasoning*, vol. 19, no. 3-4, pp. 351–365, 1998.
- [308] R. D. Shachter, “Evaluating influence diagrams,” *Operations research*, vol. 34, no. 6, pp. 871–882, 1986.
- [309] W. Saunders, G. Sastry, A. Stuhlmüller, and O. Evans, “Trial without error: Towards safe reinforcement learning via human intervention,” in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2018, Stockholm, Sweden, July 10-15, 2018*, E. André, S. Koenig, M. Dastani, and G. Sukthankar, Eds. International Foundation for Autonomous Agents and Multiagent Systems Richland, SC, USA / ACM, 2018, pp. 2067–2069. [Online]. Available: <http://dl.acm.org/citation.cfm?id=3238074>
- [310] F. Berkenkamp, M. Turchetta, A. Schoellig, and A. Krause, “Safe model-based reinforcement learning with stability guarantees,” *Advances in neural information processing systems*, vol. 30, 2017.
- [311] E. D. Langlois and T. Everitt, “How rl agents behave when their actions are modified,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 13, 2021, pp. 11586–11594.
- [312] T. Everitt, M. Hutter, R. Kumar, and V. Krakovna, “Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective,” *Synthese*, vol. 198, no. 27, pp. 6435–6467, 2021.
- [313] S. Farquhar, R. Carey, and T. Everitt, “Path-specific objectives for safer agent incentives,” 2022. [Online]. Available: <https://arxiv.org/abs/2204.10018>
- [314] A. Guez, M. Mirza, K. Gregor, R. Kabra, S. Racanière, T. Weber, D. Raposo, A. Santoro, L. Orseau, T. Eccles, G. Wayne, D. Silver, and T. P. Lillicrap, “An investigation of model-free planning,” in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 2019, pp. 2464–2473. [Online]. Available: <http://proceedings.mlr.press/v97/guez19a.html>

- [315] K. Gregor, D. Jimenez Rezende, F. Besse, Y. Wu, H. Merzic, and A. van den Oord, “Shaping belief states with generative environment models for rl,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [316] F. Baradel, N. Neverova, J. Mille, G. Mori, and C. Wolf, “Cophy: Counterfactual learning of physical dynamics,” in *International Conference on Learning Representations*, 2020. [Online]. Available: <https://openreview.net/forum?id=SkeyppEFvS>
- [317] Wikipedia contributors, “Reinventing the wheel,” 2022, [Online; accessed 10-May-2022]. [Online]. Available: https://en.wikipedia.org/wiki/Reinventing_the_wheel
- [318] J. Zhang and E. Bareinboim, “Markov decision processes with unobserved confounders: A causal approach,” Technical report, Technical Report R-23, Purdue AI Lab, Tech. Rep., 2016.
- [319] T. Wang, C. Zhou, Q. Sun, and H. Zhang, “Causal attention for unbiased visual recognition,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
- [320] A. Hallak, D. Di Castro, and S. Mannor, “Contextual markov decision processes,” *arXiv preprint arXiv:1502.02259*, 2015.
- [321] A. Guez, D. Silver, and P. Dayan, “Efficient bayes-adaptive reinforcement learning using sample-based search,” *Advances in neural information processing systems*, vol. 25, 2012.
- [322] J. Zhang and E. Bareinboim, “Can humans be out of the loop?” in *First Conference on Causal Learning and Reasoning*, 2022. [Online]. Available: <https://openreview.net/forum?id=P0f91v5fTK>
- [323] L. Bottou, F. E. Curtis, and J. Nocedal, “Optimization methods for large-scale machine learning,” *Siam Review*, vol. 60, no. 2, pp. 223–311, 2018.
- [324] K. Tang, J. Huang, and H. Zhang, “Long-tailed classification by keeping the good and removing the bad momentum causal effect,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1513–1524. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/file/1091660f3dff84fd648efe31391c5524-Paper.pdf>
- [325] W. J. Reed, “The pareto, zipf and other power laws,” *Economics letters*, vol. 74, no. 1, pp. 15–19, 2001.
- [326] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, “Generalizing from a few examples: A survey on few-shot learning,” *ACM computing surveys (csur)*, vol. 53, no. 3, pp. 1–34, 2020.

- [327] K. P. Murphy, *Probabilistic Machine Learning: An introduction.* MIT Press, 2022. [Online]. Available: probml.ai
- [328] Z. Yue, H. Zhang, Q. Sun, and X. Hua, “Interventional few-shot learning,” in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/hash/1cc8a8ea51cd0addf5dab504a285915-Abstract.html>
- [329] X. Li, Z. Zhang, G. Wei, C. Lan, W. Zeng, X. Jin, and Z. Chen, “Confounder identification-free causal visual feature learning,” *CoRR*, vol. abs/2111.13420, 2021. [Online]. Available: <https://arxiv.org/abs/2111.13420>
- [330] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.
- [331] Y. Liu, R. Cadei, J. Schweizer, S. Bahmani, and A. Alahi, “Towards robust and adaptive motion forecasting: A causal representation perspective,” 2021. [Online]. Available: <https://arxiv.org/abs/2111.14820>
- [332] D. Zhang, H. Zhang, J. Tang, X. Hua, and Q. Sun, “Causal intervention for weakly-supervised semantic segmentation,” in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/hash/07211688a0869d995947a8fb11b215d6-Abstract.html>
- [333] J. Huang, Y. Qin, J. Qi, Q. Sun, and H. Zhang, “Deconfounded visual grounding,” 2021. [Online]. Available: <https://arxiv.org/abs/2112.15324>
- [334] Y. Wang and D. M. Blei, “The blessings of multiple causes,” 2018. [Online]. Available: <https://arxiv.org/abs/1805.06826>
- [335] E. L. Ogburn, I. Shpitser, and E. J. T. Tchetgen, “Comment on “blessings of multiple causes”,” *Journal of the American Statistical Association*, vol. 114, no. 528, pp. 1611–1615, 2019.
- [336] A. D’Amour, “On multi-cause approaches to causal inference with unobserved confounding: Two cautionary failure cases and a promising alternative,” in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 3478–3486.
- [337] M. Otani, Y. Nakashima, E. Rahtu, and J. Heikkilä, “Uncovering hidden challenges in query-based video moment retrieval,” 2020. [Online]. Available: <https://arxiv.org/abs/2009.00325>

- [338] G. Nan, R. Qiao, Y. Xiao, J. Liu, S. Leng, H. Zhang, and W. Lu, "Interventional video grounding with dual contrastive learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2765–2775.
- [339] X. Yang, F. Feng, W. Ji, M. Wang, and T.-S. Chua, "Deconfounded video moment retrieval with causal intervention," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1–10.
- [340] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual visual explanations," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 2376–2384. [Online]. Available: <https://proceedings.mlr.press/v97/goyal19a.html>
- [341] L. A. Hendricks, R. Hu, T. Darrell, and Z. Akata, "Generating counterfactual explanations with natural language," *CoRR*, vol. abs/1806.09809, 2018. [Online]. Available: <http://arxiv.org/abs/1806.09809>
- [342] Z. Yue, T. Wang, H. Zhang, Q. Sun, and X.-S. Hua, "Counterfactual Zero-Shot and Open-Set Visual Recognition," *arXiv:2103.00887 [cs]*, Mar. 2021, arXiv: 2103.00887. [Online]. Available: <http://arxiv.org/abs/2103.00887>
- [343] J. Pearl, "Brief report: On the consistency rule in causal inference: "axiom, definition, assumption, or theorem?"," *Epidemiology*, vol. 21, no. 6, pp. 872–875, 2010. [Online]. Available: <http://www.jstor.org/stable/20788241>
- [344] X. Zhang, Y. Wong, X. Wu, J. Lu, M. Kankanhalli, X. Li, and W. Geng, "Learning causal representation for training cross-domain pose estimator via generative interventions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11270–11280.
- [345] Y. Li, A. Torralba, A. Anandkumar, D. Fox, and A. Garg, "Causal discovery in physical systems from videos," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9180–9192, 2020.
- [346] T. D. Kulkarni, A. Gupta, C. Ionescu, S. Borgeaud, M. Reynolds, A. Zisserman, and V. Mnih, "Unsupervised learning of object keypoints for perception and control," *Advances in neural information processing systems*, vol. 32, 2019.
- [347] S. Janny, F. Baradel, N. Neverova, M. Nadri, G. Mori, and C. Wolf, "Filtered-cophy: Unsupervised learning of counterfactual physics in pixel space," *arXiv preprint arXiv:2202.00368*, 2022.
- [348] D. Maliniak, R. Powers, and B. F. Walter, "The gender citation gap in international relations," *International Organization*, vol. 67, no. 4, pp. 889–922, 2013.

- [349] A. Feder, K. A. Keith, E. Manzoor, R. Pryzant, D. Sridhar, Z. Wood-Doughty, J. Eisenstein, J. Grimmer, R. Reichart, M. E. Roberts, B. M. Stewart, V. Veitch, and D. Yang, “Causal inference in natural language processing: Estimation, prediction, interpretation and beyond,” *CoRR*, vol. abs/2109.00725, 2021.
- [350] C. Fong and J. Grimmer, “Discovery of treatments from text corpora,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 1600–1609.
- [351] R. Pryzant, K. Shen, D. Jurafsky, and S. Wagner, “Deconfounded lexicon induction for interpretable social science,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 1615–1625.
- [352] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, “Beyond accuracy: Behavioral testing of NLP models with checklist,” *CoRR*, vol. abs/2005.04118, 2020.
- [353] A. Ross, T. Wu, H. Peng, M. E. Peters, and M. Gardner, “Tailor: Generating and perturbing text with semantic controls,” *CoRR*, vol. abs/2107.07150, 2021.
- [354] R. Adragna, E. Creager, D. Madras, and R. S. Zemel, “Fairness and robustness in invariant learning: A case study in toxicity classification,” *CoRR*, vol. abs/2011.06485, 2020.
- [355] S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, and A. Beutel, “Counterfactual fairness in text classification through robustness,” *CoRR*, vol. abs/1809.10610, 2018.
- [356] Q. Liu, M. J. Kusner, and P. Blunsom, “Counterfactual data augmentation for neural machine translation,” in *NAACL-HLT*. Association for Computational Linguistics, 2021, pp. 187–197.
- [357] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine, “The hateful memes challenge: Detecting hate speech in multimodal memes,” 2020. [Online]. Available: <https://arxiv.org/abs/2005.04790>
- [358] E. Abbasnejad, D. Teney, A. Parvaneh, J. Shi, and A. van den Hengel, “Counterfactual Vision and Language Learning,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, Jun. 2020, pp. 10 041–10 051. [Online]. Available: <https://ieeexplore.ieee.org/document/9156448/>
- [359] M. Alzantot, Y. Sharma, A. Elgohary, B.-J. Ho, M. Srivastava, and K.-W. Chang, “Generating natural language adversarial examples,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 2890–2896. [Online]. Available: <https://aclanthology.org/D18-1316>

- [360] M. Iyyer, J. Wieting, K. Gimpel, and L. Zettlemoyer, “Adversarial example generation with syntactically controlled paraphrase networks,” 2018. [Online]. Available: <https://arxiv.org/abs/1804.06059>
- [361] J. Ebrahimi, A. Rao, D. Lowd, and D. Dou, “Hotflip: White-box adversarial examples for text classification,” 2017. [Online]. Available: <https://arxiv.org/abs/1712.06751>
- [362] M. Ye, C. Gong, and Q. Liu, “Safer: A structure-free approach for certified robustness to adversarial word substitutions,” 2020. [Online]. Available: <https://arxiv.org/abs/2005.14424>
- [363] M. Mozes, M. Bartolo, P. Stenetorp, B. Kleinberg, and L. D. Griffin, “Contrasting human- and machine-generated word-level adversarial examples for text classification,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds. Association for Computational Linguistics, 2021, pp. 8258–8270. [Online]. Available: <https://doi.org/10.18653/v1/2021.emnlp-main.651>
- [364] H. Zhao, C. Ma, X. Dong, A. T. Luu, Z.-H. Deng, and H. Zhang, “Certified robustness against natural language attacks by causal intervention,” *arXiv preprint arXiv:2205.12331*, 2022.
- [365] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, “Learning word vectors for sentiment analysis,” in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 2011, pp. 142–150.
- [366] A. Feder, N. Oved, U. Shalit, and R. Reichart, “Causalm: Causal model explanation through counterfactual language models,” *Comput. Linguistics*, vol. 47, no. 2, pp. 333–386, 2021. [Online]. Available: https://doi.org/10.1162/coli_a_00404
- [367] M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, and A. T. Kalai, “Bias in bios: A case study of semantic representation bias in a high-stakes setting,” in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ser. FAT* ’19. New York, NY, USA: Association for Computing Machinery, 2019, p. 120–128. [Online]. Available: <https://doi.org/10.1145/3287560.3287572>
- [368] J. Vig, S. Gehrmann, Y. Belinkov, S. Qian, D. Nevo, Y. Singer, and S. M. Shieber, “Investigating gender bias in language models using causal mediation analysis,” in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html>

- [369] S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, and A. Beutel, “Counterfactual fairness in text classification through robustness,” in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019, Honolulu, HI, USA, January 27-28, 2019*, V. Conitzer, G. K. Hadfield, and S. Vallor, Eds. ACM, 2019, pp. 219–226. [Online]. Available: <https://doi.org/10.1145/3306618.3317950>
- [370] X. Zeng, Y. Li, Y. Zhai, and Y. Zhang, “Counterfactual generator: A weakly-supervised method for named entity recognition,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, 2020, pp. 7270–7280. [Online]. Available: <https://doi.org/10.18653/v1/2020.emnlp-main.590>
- [371] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, “Man is to computer programmer as woman is to homemaker? debiasing word embeddings,” in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016. [Online]. Available: <https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf>
- [372] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, “Men also like shopping: Reducing gender bias amplification using corpus-level constraints,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 2979–2989. [Online]. Available: <https://aclanthology.org/D17-1323>
- [373] R. Zmigrod, S. J. Mielke, H. Wallach, and R. Cotterell, “Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 1651–1661. [Online]. Available: <https://aclanthology.org/P19-1161>
- [374] J. Nivre, M. Abrams, Ž. Agić, L. Ahrenberg, L. Antonsen, K. Aplonova, M. J. Aranzabe, G. Arutie, M. Asahara, L. Ateyah, M. Attia, A. Atutxa, L. Augustinus, E. Badmaeva, M. Ballesteros, E. Banerjee, S. Bank, V. Barbu Mititelu, V. Basmov, J. Bauer, S. Bellato, K. Bengoetxea, Y. Berzak, I. A. Bhat, R. A. Bhat, E. Biagetti, E. Bick, R. Blokland, V. Bobicev, C. Börstell, C. Bosco, G. Bouma, S. Bowman, A. Boyd, A. Burchardt, M. Candito, B. Caron, G. Caron, G. Cebiroğlu Eryigit, F. M. Cecchini, G. G. A. Celano, S. Čéplö, S. Cetin, F. Chalub, J. Choi, Y. Cho, J. Chun, S. Cinková, A. Collomb, Ç. Çöltekin, M. Connor, M. Courtin, E. Davidson, M.-C. de Marneffe, V. de Paiva, A. Diaz de Ilarrazza, C. Dickerson, P. Dirix, K. Dobrovoljc, T. Dozat, K. Droganova, P. Dwivedi, M. Eli, A. Elkahky, B. Ephrem, T. Erjavec, A. Etienne,

R. Farkas, H. Fernandez Alcalde, J. Foster, C. Freitas, K. Gajdošová, D. Galbraith, M. Garcia, M. Gärdenfors, S. Garza, K. Gerdes, F. Ginter, I. Goenaga, K. Gojenola, M. Gökirmak, Y. Goldberg, X. Gómez Guinovart, B. González Saavedra, M. Grioni, N. Grūžītis, B. Guillaume, C. Guillot-Barbance, N. Habash, J. Hajič, J. Hajič jr., L. Hà Mý, N.-R. Han, K. Harris, D. Haug, B. Hladká, J. Hlaváčová, F. Hociung, P. Hohle, J. Hwang, R. Ion, E. Irimia, Ö. Ishola, T. Jelínek, A. Johannsen, F. Jørgensen, H. Kaşikara, S. Kahane, H. Kanayama, J. Kanerva, B. Katz, T. Kayadelen, J. Kenney, V. Kettnerová, J. Kirchner, K. Kopacewicz, N. Kotsyba, S. Krek, S. Kwak, V. Laippala, L. Lambertino, L. Lam, T. Lando, S. D. Larasati, A. Lavrentiev, J. Lee, P. Lê Hồng, A. Lenci, S. Lertpradit, H. Leung, C. Y. Li, J. Li, K. Li, K. Lim, N. Ljubešić, O. Loginova, O. Lyashevskaya, T. Lynn, V. Macketanz, A. Makazhanov, M. Mandl, C. Manning, R. Manurung, C. Mărănduc, D. Mareček, K. Marheinecke, H. Martínez Alonso, A. Martins, J. Mašek, Y. Matsumoto, R. McDonald, G. Mendonça, N. Miekka, M. Misirpashayeva, A. Missilä, C. Mititelu, Y. Miyao, S. Montemagni, A. More, L. Moreno Romero, K. S. Mori, S. Mori, B. Mortensen, B. Moskalevskyi, K. Muischnek, Y. Murawaki, K. Müürisepp, P. Nainwani, J. I. Navarro Horňiacek, A. Nedoluzhko, G. Nešpore-Běrzkalne, L. Nguyẽn Thị, H. Nguyẽn Thị Minh, V. Nikolaev, R. Nitisoroj, H. Nurmi, S. Ojala, A. Olúòkun, M. Omura, P. Osenova, R. Östling, L. Øvreliid, N. Partanen, E. Pascual, M. Passarotti, A. Patejuk, G. Paulino-Passos, S. Peng, C.-A. Perez, G. Perrier, S. Petrov, J. Piitulainen, E. Pitler, B. Plank, T. Poibeau, M. Popel, L. Pretkalniņa, S. Prévost, P. Prokopidis, A. Przepiórkowski, T. Puolakainen, S. Pyysalo, A. Rääbis, A. Rademaker, L. Ramasamy, T. Rama, C. Ramisch, V. Ravishankar, L. Real, S. Reddy, G. Rehm, M. Rießler, L. Rinaldi, L. Rituma, L. Rocha, M. Romanenko, R. Rosa, D. Rovati, V. Roșca, O. Rudina, J. Rueter, S. Sadde, B. Sagot, S. Saleh, T. Samardžić, S. Samson, M. Sanguinetti, B. Saulīte, Y. Sawanakunanon, N. Schneider, S. Schuster, D. Seddah, W. Seeker, M. Seraji, M. Shen, A. Shimada, M. Shohibussirri, D. Sichinava, N. Silveira, M. Simi, R. Simionescu, K. Simkó, M. Šimková, K. Simov, A. Smith, I. Soares-Bastos, C. Spadine, A. Stella, M. Straka, J. Strnadová, A. Suhr, U. Sulubacak, Z. Szántó, D. Taji, Y. Takahashi, T. Tanaka, I. Tellier, T. Trosterud, A. Trukhina, R. Tsarfaty, F. Tyers, S. Uematsu, Z. Urešová, L. Uria, H. Uszkoreit, S. Vajjala, D. van Niekerk, G. van Noord, V. Varga, E. Villemonte de la Clergerie, V. Vincze, L. Wallin, J. X. Wang, J. N. Washington, S. Williams, M. Wirén, T. Woldemariam, T.-s. Wong, C. Yan, M. M. Yavrumyan, Z. Yu, Z. Žabokrtský, A. Zeldes, D. Zeman, M. Zhang, and H. Zhu, “Universal dependencies 2.3,” 2018, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. [Online]. Available: <http://hdl.handle.net/11234/1-2895>

- [375] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. M. Mooij, “On causal and anticausal learning,” in *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress, 2012. [Online]. Available: <http://icml.cc/2012/papers/625.pdf>
- [376] N. Kilbertus*, G. Parascandolo*, and B. Schölkopf*, “Generalization in anti-causal learning,” in *NeurIPS 2018 Workshop on Critiquing and Correcting Trends in Machine Learning*, Dec. 2018, *authors are listed in alphabetical order. [Online]. Available: <https://ml-critique-correct.github.io/>
- [377] Z. Jin, J. von Kügelgen, J. Ni, T. Vaidhya, A. Kaushal, M. Sachan, and B. Schölkopf, “Causal direction of data collection matters: Implications of causal and anticausal learning for nlp,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Nov. 2021, pp. 9499–9513, *equal contribution. [Online]. Available: <https://aclanthology.org/2021.emnlp-main.748>
- [378] W. Fan, Y. Ma, Q. Li, Y. He, E. Zhao, J. Tang, and D. Yin, “Graph neural networks for social recommendation,” in *The world wide web conference*, 2019, pp. 417–426.
- [379] X. Jing and J. Xu, “Fast and effective protein model refinement using deep graph neural networks,” *Nature computational science*, vol. 1, no. 7, pp. 462–469, 2021.
- [380] W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, and J. Leskovec, “Open graph benchmark: Datasets for machine learning on graphs,” *Advances in neural information processing systems*, vol. 33, pp. 22118–22133, 2020.
- [381] Y. Wu, X. Wang, A. Zhang, X. He, and T.-S. Chua, “Discovering invariant rationales for graph neural networks,” in *International Conference on Learning Representations*, 2022. [Online]. Available: <https://openreview.net/forum?id=hGXij5rfiHw>
- [382] Y. Chen, Y. Zhang, H. Yang, K. Ma, B. Xie, T. Liu, B. Han, and J. Cheng, “Invariance principle meets out-of-distribution generalization on graphs,” *arXiv preprint arXiv:2202.05441*, 2022.
- [383] Y. Sui, X. Wang, J. Wu, X. He, and T.-S. Chua, “Deconfounded training for graph neural networks,” *ArXiv*, vol. abs/2112.15089, 2021.
- [384] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad, “Collective classification in network data,” *AI magazine*, vol. 29, no. 3, pp. 93–93, 2008.
- [385] T. Zhao, G. Liu, D. Wang, W. Yu, and M. Jiang, “Learning from counterfactual links for link prediction,” 2021. [Online]. Available: <https://arxiv.org/abs/2106.02172>

- [386] D. Chen, Y. Lin, W. Li, P. Li, J. Zhou, and X. Sun, “Measuring and relieving the over-smoothing problem for graph neural networks from the topological view,” in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 2020, pp. 3438–3445. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/5747>
- [387] F. Feng, W. Huang, X. He, X. Xin, Q. Wang, and T.-S. Chua, “Should graph convolution trust neighbors? a simple causal inference method,” in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1208–1218.
- [388] M. Zečević, D. S. Dhami, P. Veličković, and K. Kersting, “Relating graph neural networks to structural causal models,” 2021.
- [389] E. Todorov, T. Erez, and Y. Tassa, “Mujoco: A physics engine for model-based control,” in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 5026–5033.
- [390] M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling, “The arcade learning environment: An evaluation platform for general agents,” *Journal of Artificial Intelligence Research*, vol. 47, pp. 253–279, jun 2013.
- [391] N. R. Ke, A. Didolkar, S. Mittal, A. Goyal, G. Lajoie, S. Bauer, D. Rezende, Y. Bengio, M. Mozer, and C. Pal, “Systematic evaluation of causal discovery in visual model based reinforcement learning,” 2021. [Online]. Available: <https://arxiv.org/abs/2107.00848>
- [392] J. X. Wang, M. King, N. P. M. Porcel, Z. Kurth-Nelson, T. Zhu, C. Deck, P. Choy, M. Cassin, M. Reynolds, H. F. Song, G. Buttimore, D. P. Reichert, N. C. Rabinowitz, L. Matthey, D. Hassabis, A. Lerchner, and M. Botvinick, “Alchemy: A benchmark and analysis toolkit for meta-reinforcement learning agents,” in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. [Online]. Available: <https://openreview.net/forum?id=eZu4BZxlRnX>
- [393] D. McDuff, Y. Song, J. Lee, V. Vineet, S. Vemprala, N. A. Gyde, H. Salman, S. Ma, K. Sohn, and A. Kapoor, “Causalcity: Complex simulations with agency for causal discovery and reasoning,” in *First Conference on Causal Learning and Reasoning*, 2022. [Online]. Available: <https://openreview.net/forum?id=YWRhER626PX>
- [394] K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, and J. B. Tenenbaum, “Clevrer: Collision events for video representation and reasoning,” in *International Conference on Learning Representations*, 2020. [Online]. Available: <https://openreview.net/forum?id=HkxYzANYDB>

- [395] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” in *NAACL-HLT (1)*. Association for Computational Linguistics, 2019, pp. 4171–4186.
- [396] D. Kaushik, E. Hovy, and Z. Lipton, “Learning the difference that makes a difference with counterfactually-augmented data,” in *International Conference on Learning Representations*, 2020. [Online]. Available: <https://openreview.net/forum?id=SkIgs0NFvr>
- [397] J. Frohberg and F. Binder, “Crass: A novel data set and benchmark to test counterfactual reasoning of large language models,” 2021.
- [398] A. Srivastava, A. Rastogi, A. B. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, A. Kluska, A. Lewkowycz, A. Agarwal, A. Power, A. Ray, A. Warstadt, A. W. Kocurek, A. Safaya, A. Tazarv, A. Xiang, A. Parrish, A. Nie, A. Hussain, A. Askell, A. Dsouza, A. A. Rahane, A. S. Iyer, A. J. Andreassen, A. Santilli, A. Stuhlmuller, A. M. Dai, A. D. La, A. K. Lampinen, A. Zou, A. Jiang, A. Chen, A. Vuong, A. Gupta, A. Gottardi, A. Norelli, A. Venkatesh, A. Gholamidavoodi, A. Tabassum, A. Menezes, A. Kirubarajan, A. Mullokandov, A. Sabharwal, A. Herrick, A. Efrat, A. Erdem, A. Karakacs, B. R. Roberts, B. S. Loe, B. Zoph, B. Bojanowski, B. Ozyurt, B. Hedayatnia, B. Neyshabur, B. Inden, B. Stein, B. Ekmekci, B. Y. Lin, B. S. Howald, C. Diao, C. Dour, C. Stinson, C. Argueta, C. F. Ram’irez, C. Singh, C. Rathkopf, C. Meng, C. Baral, C. Wu, C. Callison-Burch, C. Waites, C. Voigt, C. D. Manning, C. Potts, C. T. Ramirez, C. Rivera, C. Siro, C. Raffel, C. Ashcraft, C. Garbacea, D. Sileo, D. H. Garrette, D. Hendrycks, D. Kilman, D. Roth, D. Freeman, D. Khashabi, D. Levy, D. Gonz’alez, D. Hernandez, D. Chen, D. Ippolito, D. Gilboa, D. Dohan, D. Drakard, D. Jurgens, D. Datta, D. Ganguli, D. Emelin, D. Kleyko, D. Yuret, D. Chen, D. Tam, D. Hupkes, D. Misra, D. Buzan, D. C. Mollo, D. Yang, D.-H. Lee, E. Shutova, E. D. Cubuk, E. Segal, E. Hagerman, E. Barnes, E. P. Donoway, E. Pavlick, E. Rodolà, E. F. Lam, E. Chu, E. Tang, E. Erdem, E. Chang, E. A. Chi, E. Dyer, E. Jerzak, E. Kim, E. E. Manyasi, E. Zheltonozhskii, F. Xia, F. Siar, F. Mart’inez-Plumed, F. Happ’e, F. Chollet, F. Rong, G. Mishra, G. I. Winata, G. de Melo, G. Kruszewski, G. Parascanolo, G. Mariani, G. Wang, G. Jaimovich-L’opez, G. Betz, G. Gur-Ari, H. Galijasevic, H. S. Kim, H. Rashkin, H. Hajishirzi, H. Mehta, H. Bogar, H. Shevlin, H. Schütze, H. Yakura, H. Zhang, H. Wong, I. A.-S. Ng, I. Noble, J. Jumelet, J. Geissinger, J. Kernion, J. Hilton, J. Lee, J. F. Fisac, J. B. Simon, J. Koppel, J. Zheng, J. Zou, J. Koco’n, J. Thompson, J. Kaplan, J. Radom, J. Sohl-Dickstein, J. Phang, J. Wei, J. Yosinski, J. Novikova, J. Bosscher, J. Marsh, J. Kim, J. Taal, J. Engel, J. O. Alabi, J. Xu, J. Song, J. Tang, J. W. Waweru, J. Burden, J. Miller, J. U. Balis, J. Berant, J. Frohberg, J. Rozen, J. Hernández-Orallo, J. Boudeman, J. Jones, J. B. Tenenbaum, J. S. Rule, J. Chua, K. Kanclerz, K. Livescu, K. Krauth, K. Gopalakrishnan, K. Ignatyeva, K. Markert, K. D. Dhole, K. Gimpel, K. O. Omondi, K. W. Math-

ewson, K. Chiaffullo, K. Shkaruta, K. Shridhar, K. McDonell, K. Richardson, L. Reynolds, L. Gao, L. Zhang, L. Dugan, L. Qin, L. Contreras-Ochando, L.-P. Morency, L. Moschella, L. Lam, L. Noble, L. Schmidt, L. He, L. O. Col'on, L. Metz, L. K. cSenel, M. Bosma, M. Sap, M. ter Hoeve, M. Andrea, M. S. Farooqi, M. Faruqui, M. Mazeika, M. Baturan, M. Marelli, M. Maru, M. Quintana, M. Tolkiehn, M. Giulianelli, M. Lewis, M. Potthast, M. Leavitt, M. Hagen, M. Schubert, M. Baitemirova, M. Arnaud, M. A. McElrath, M. A. Yee, M. Cohen, M. Gu, M. I. Ivanitskiy, M. Starritt, M. Strube, M. Swkedrowski, M. Bevilacqua, M. Yasunaga, M. Kale, M. Cain, M. Xu, M. Suzgun, M. Tiwari, M. Bansal, M. Aminnaseri, M. Geva, M. Gheini, T. MukundVarma, N. Peng, N. Chi, N. Lee, N. G.-A. Krakover, N. Cameron, N. S. Roberts, N. Doiron, N. Nangia, N. Deckers, N. Muennighoff, N. S. Keskar, N. Iyer, N. Constant, N. Fiedel, N. Wen, O. Zhang, O. Agha, O. Elbaghdadi, O. Levy, O. Evans, P. Casares, P. Doshi, P. Fung, P. P. Liang, P. Vicol, P. Alipoor-molabashi, P. Liao, P. Liang, P. W. Chang, P. Eckersley, P. M. Htut, P.-B. Hwang, P. Milkowski, P. S. Patil, P. Pezeshkpour, P. Oli, Q. Mei, Q. LYU, Q. Chen, R. Banjade, R. E. Rudolph, R. Gabriel, R. Habacker, R. R. Delgado, R. Millière, R. Garg, R. Barnes, R. A. Saurous, R. Arakawa, R. Raymaekers, R. Frank, R. Sikand, R. Novak, R. Sitelew, R. Lebras, R. Liu, R. Jacobs, R. Zhang, R. Salakhutdinov, R. Chi, R. Lee, R. Stovall, R. Teehan, R. Yang, S. J. Singh, S. M. Mohammad, S. Anand, S. Dillavou, S. Shleifer, S. Wiseman, S. Gruetter, S. Bowman, S. S. Schoenholz, S. Han, S. Kwatra, S. A. Rous, S. Ghazarian, S. Ghosh, S. Casey, S. Bischoff, S. Gehrmann, S. Schuster, S. Sadeghi, S. S. Hamdan, S. Zhou, S. Srivastava, S. Shi, S. Singh, S. Asaadi, S. S. Gu, S. Pachchigar, S. Toshniwal, S. Upadhyay, S. Debnath, S. Shakeri, S. Thormeyer, S. Melzi, S. Reddy, S. P. Makini, S. hwan Lee, S. B. Torene, S. Hatwar, S. Dehaene, S. Divic, S. Ermon, S. R. Biderman, S. C. Lin, S. Prasad, S. T. Piantadosi, S. M. Shieber, S. Mishergi, S. Kiritchenko, S. Mishra, T. Linzen, T. Schuster, T. Li, T. Yu, T. A. Ali, T. Hashimoto, T.-L. Wu, T. Desbordes, T. Rothschild, T. Phan, T. Wang, T. Nkinyili, T. Schick, T. N. Kornev, T. Telleen-Lawton, T. Tunduny, T. Gerstenberg, T. Chang, T. Neeraj, T. Khot, T. O. Shultz, U. Shaham, V. Misra, V. Demberg, V. Nyamai, V. Raunak, V. V. Ramasesh, V. U. Prabhu, V. Padmakumar, V. Sriku-mar, W. Fedus, W. Saunders, W. Zhang, W. Vossen, X. Ren, X. F. Tong, X. Wu, X. Shen, Y. Yaghoobzadeh, Y. Lakretz, Y. Song, Y. Bahri, Y. J. Choi, Y. Yang, Y. Hao, Y. Chen, Y. Belinkov, Y. Hou, Y. Hou, Y. Bai, Z. Seid, Z. Xinran, Z. Zhao, Z. F. Wang, Z. J. Wang, Z. Wang, Z. Wu, S. Singh, and U. Shaham, “Beyond the imitation game: Quantifying and extrapolating the capabilities of language models,” *ArXiv*, vol. abs/2206.04615, 2022.

- [399] L. Yang, Z. Wang, Y. Wu, J. Yang, and Y. Zhang, “Towards fine-grained causal reasoning and qa,” *arXiv preprint arXiv:2204.07408*, 2022.
- [400] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in pytorch,” 2017.

- [401] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, “{TensorFlow}: a system for {Large-Scale} machine learning,” in *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 2016, pp. 265–283.
- [402] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, “JAX: composable transformations of Python+NumPy programs,” 2018. [Online]. Available: <http://github.com/google/jax>
- [403] J. Heek, A. Levskaya, A. Oliver, M. Ritter, B. Rondepierre, A. Steiner, and M. van Zee, “Flax: A neural network library and ecosystem for JAX,” 2020. [Online]. Available: <http://github.com/google/flax>
- [404] T. Hennigan, T. Cai, T. Norman, and I. Babuschkin, “Haiku: Sonnet for JAX,” 2020. [Online]. Available: <http://github.com/deepmind/dm-haiku>
- [405] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, “Transformers: State-of-the-art natural language processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: <https://www.aclweb.org/anthology/2020.emnlp-demos.6>
- [406] R. Wightman, “Pytorch image models,” <https://github.com/rwightman/pytorch-image-models>, 2019.
- [407] M. Fey and J. E. Lenssen, “Fast graph representation learning with PyTorch Geometric,” in *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.
- [408] K. Xia, K.-Z. Lee, Y. Bengio, and E. Bareinboim, “The causal-neural connection: Expressiveness, learnability, and inference,” *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [409] M. Scutari, “Learning bayesian networks with the bnlearn r package,” *arXiv preprint arXiv:0908.3817*, 2009.
- [410] M. Kalisch, M. Mächler, D. Colombo, M. H. Maathuis, and P. Bühlmann, “Causal inference using graphical models with the r package pcalg,” *Journal of statistical software*, vol. 47, pp. 1–26, 2012.
- [411] A. Sharma and E. Kiciman, “Dowhy: An end-to-end library for causal inference,” *arXiv preprint arXiv:2011.04216*, 2020.
- [412] H. Chen, T. Harinen, J.-Y. Lee, M. Yung, and Z. Zhao, “Causalml: Python package for causal machine learning,” *arXiv preprint arXiv:2002.11631*, 2020.

- [413] P. Bach, V. Chernozhukov, M. S. Kurz, and M. Spindler, “Doubleml—an object-oriented implementation of double machine learning in r,” *arXiv preprint arXiv:2103.09603*, 2021.
- [414] James Fox, Tom Everitt, Ryan Carey, Eric Langlois, Alessandro Abate, and Michael Wooldridge, “PyCID: A Python Library for Causal Influence Diagrams,” in *Proceedings of the 20th Python in Science Conference*, Meghann Agarwal, Chris Calloway, Dillon Niederhut, and David Shupe, Eds., 2021, pp. 43 – 51.
- [415] I. Gulrajani and D. Lopez-Paz, “In search of lost domain generalization,” in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. [Online]. Available: <https://openreview.net/forum?id=lQdXeXDoWtI>
- [416] J. Wang, C. Lan, C. Liu, Y. Ouyang, W. Zeng, and T. Qin, “Generalizing to unseen domains: A survey on domain generalization,” *arXiv preprint arXiv:2103.03097*, 2021.
- [417] L. Zintgraf, K. Shiarlis, M. Igl, S. Schulze, Y. Gal, K. Hofmann, and S. Whiteson, “Varibad: A very good method for bayes-adaptive deep rl via meta-learning,” *arXiv preprint arXiv:1910.08348*, 2019.
- [418] A. Gupta, R. Mendonca, Y. Liu, P. Abbeel, and S. Levine, “Meta-reinforcement learning of structured exploration strategies,” *Advances in neural information processing systems*, vol. 31, 2018.
- [419] L. Cheng, R. Guo, R. Moraffah, P. Sheth, K. S. Candan, and H. Liu, “Evaluation methods and measures for causal learning algorithms,” *IEEE Transactions on Artificial Intelligence*, 2022.
- [420] D. B. Rubin, “Estimating causal effects of treatments in randomized and non-randomized studies.” *Journal of educational Psychology*, vol. 66, no. 5, p. 688, 1974.
- [421] K. Quach, “Openai shuts down robotics team because it doesn’t have enough data yet,” 2021, [Online; accessed 30-May-2022]. [Online]. Available: https://www.theregister.com/2021/07/18/in_brief_ai/
- [422] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. P. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, “Mastering the game of go with deep neural networks and tree search,” *Nat.*, vol. 529, no. 7587, pp. 484–489, 2016. [Online]. Available: <https://doi.org/10.1038/nature16961>
- [423] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. P. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, “Mastering the

- game of go without human knowledge,” *Nat.*, vol. 550, no. 7676, pp. 354–359, 2017. [Online]. Available: <https://doi.org/10.1038/nature24270>
- [424] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. A. Riedmiller, “Playing atari with deep reinforcement learning,” *CoRR*, vol. abs/1312.5602, 2013. [Online]. Available: <http://arxiv.org/abs/1312.5602>
- [425] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev *et al.*, “Grandmaster level in starcraft ii using multi-agent reinforcement learning,” *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.
- [426] M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal, “The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care,” *Nature medicine*, vol. 24, no. 11, pp. 1716–1720, 2018.
- [427] M. G. Bellemare, S. Candido, P. S. Castro, J. Gong, M. C. Machado, S. Moitra, S. S. Ponda, and Z. Wang, “Autonomous navigation of stratospheric balloons using reinforcement learning,” *Nature*, vol. 588, no. 7836, pp. 77–82, 2020.
- [428] J. Pearl, “Invited commentary: understanding bias amplification,” *American journal of epidemiology*, vol. 174, no. 11, pp. 1223–1227, 2011.
- [429] M. L Mitchell and J. M Jolley, *Research design explained*, 2010.
- [430] W. H. Jefferys and J. O. Berger, “Ockham’s razor and bayesian analysis,” *American scientist*, vol. 80, no. 1, pp. 64–72, 1992.
- [431] B. Schölkopf, “Causality for machine learning,” 2019.
- [432] A. Feder, K. A. Keith, E. Manzoor, R. Pryzant, D. Sridhar, Z. Wood-Doughty, J. Eisenstein, J. Grimmer, R. Reichart, M. E. Roberts, B. M. Stewart, V. Veitch, and D. Yang, “Causal inference in natural language processing: Estimation, prediction, interpretation and beyond,” 2021.
- [433] T. McCoy, E. Pavlick, and T. Linzen, “Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3428–3448. [Online]. Available: <https://aclanthology.org/P19-1334>
- [434] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A survey of methods for explaining black box models,” *ACM Comput. Surv.*, vol. 51, no. 5, aug 2018. [Online]. Available: <https://doi.org/10.1145/3236009>
- [435] L. Cheng, A. Mosallanezhad, P. Sheth, and H. Liu, “Causal learning for socially responsible ai,” 2021. [Online]. Available: <https://arxiv.org/abs/2104.12278>

- [436] Y. Liu, Y. Wei, H. Yan, G. Li, and L. Lin, “Causal reasoning meets visual representation learning: A prospective study,” 2022. [Online]. Available: <https://arxiv.org/abs/2204.12037>
- [437] P. Sanchez, J. P. Voisey, T. Xia, H. I. Watson, A. Q. O’Neil, and S. A. Tsaftaris, “Causal machine learning for healthcare and precision medicine,” 2022. [Online]. Available: <https://arxiv.org/abs/2205.11402>
- [438] A. Vlontzos, D. Rueckert, and B. Kainz, “A review of causality for learning algorithms in medical image analysis,” 2022. [Online]. Available: <https://arxiv.org/abs/2206.05498>
- [439] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, “Ai for radiographic covid-19 detection selects shortcuts over signal,” *Nature Machine Intelligence*, vol. 3, no. 7, pp. 610–619, 2021.
- [440] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu, “Metalearners for estimating heterogeneous treatment effects using machine learning,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 10, pp. 4156–4165, 2019.
- [441] U. Shalit, F. D. Johansson, and D. Sontag, “Estimating individual treatment effect: generalization bounds and algorithms,” in *International Conference on Machine Learning*. PMLR, 2017, pp. 3076–3085.
- [442] C. Shi, D. Blei, and V. Veitch, “Adapting neural networks for the estimation of treatment effects,” in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.
- [443] A. Caron, I. Manolopoulou, and G. Baio, “Estimating individual treatment effects using non-parametric regression models: a review,” *arXiv preprint arXiv:2009.06472*, 2020.
- [444] X. Nie and S. Wager, “Quasi-oracle estimation of heterogeneous treatment effects,” *Biometrika*, 09 2020.
- [445] A. Curth and M. van der Schaar, “Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms,” in *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, A. Banerjee and K. Fukumizu, Eds., vol. 130. PMLR, 2021, pp. 1810–1818.
- [446] L. Nie, M. Ye, qiang liu, and D. Nicolae, “{VCN}et and functional targeted regularization for learning causal effects of continuous treatments,” in *International Conference on Learning Representations*, 2021.
- [447] Y.-F. Zhang, H. Zhang, Z. C. Lipton, L. E. Li, and E. P. Xing, “Exploring transformer backbones for heterogeneous treatment effect estimation,” 2022. [Online]. Available: <https://arxiv.org/abs/2202.01336>

- [448] A. Malek and S. Chiappa, “Asymptotically best causal effect identification with multi-armed bandits,” in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: <https://openreview.net/forum?id=1dqrBgHYC0d>
- [449] C. F. Manski, “Nonparametric bounds on treatment effects,” *The American Economic Review*, vol. 80, no. 2, pp. 319–323, 1990.
- [450] C. Manski, *Partial identification of probability distributions*. Springer, 2003, vol. 5.
- [451] K. Imai, L. Keele, and T. Yamamoto, “Identification, inference and sensitivity analysis for causal mediation effects,” *Statistical science*, vol. 25, no. 1, pp. 51–71, 2010.
- [452] C. Cinelli, D. Kumor, B. Chen, J. Pearl, and E. Bareinboim, “Sensitivity analysis of linear structural causal models,” in *International conference on machine learning*. PMLR, 2019, pp. 1252–1261.
- [453] M. Baiocchi, J. Cheng, and D. S. Small, “Instrumental variable methods for causal inference,” *Statistics in medicine*, vol. 33, no. 13, pp. 2297–2340, 2014.
- [454] W. Miao, Z. Geng, and E. J. Tchetgen Tchetgen, “Identifying causal effects with proxy variables of an unmeasured confounder,” *Biometrika*, vol. 105, no. 4, pp. 987–993, 2018.
- [455] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling, “Causal effect inference with deep latent-variable models,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.
- [456] L. Xu, H. Kanagawa, and A. Gretton, “Deep proxy causal learning and its application to confounded bandit policy evaluation,” *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [457] L. Gultchin, D. S. Watson, M. J. Kusner, and R. Silva, “Operationalizing complex causes: A pragmatic view of mediation,” in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 3875–3885. [Online]. Available: <http://proceedings.mlr.press/v139/gultchin21a.html>
- [458] A. Mastouri, Y. Zhu, L. Gultchin, A. Korba, R. Silva, M. Kusner, A. Gretton, and K. Muandet, “Proximal causal learning with kernels: Two-stage estimation and moment restriction,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 7512–7523.

- [459] A. Ghassami, A. Ying, I. Shpitser, and E. T. Tchetgen, “Minimax kernel machine learning for a class of doubly robust functionals with application to proximal causal inference,” 2021. [Online]. Available: <https://arxiv.org/abs/2104.02929>
- [460] A. Ying, Y. Cui, and E. J. T. Tchetgen, “Proximal causal inference for marginal counterfactual survival curves,” 2022. [Online]. Available: <https://arxiv.org/abs/2204.13144>
- [461] A. Ghassami, I. Shpitser, and E. T. Tchetgen, “Proximal causal inference with hidden mediators: Front-door and related mediation problems,” 2021. [Online]. Available: <https://arxiv.org/abs/2111.02927>
- [462] J. H. Stock and F. Trebbi, “Retrospectives: Who invented instrumental variable regression?” *Journal of Economic Perspectives*, vol. 17, no. 3, pp. 177–194, 2003.
- [463] R. Singh, M. Sahani, and A. Gretton, “Kernel instrumental variable regression,” in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 4595–4607. [Online]. Available: <https://proceedings.neurips.cc/paper/2019/hash/17b3c7061788dbe82de5abe9f6fe22b3-Abstract.html>
- [464] A. Bennett, N. Kallus, and T. Schnabel, “Deep generalized method of moments for instrumental variable analysis,” *Advances in neural information processing systems*, vol. 32, 2019.
- [465] K. Muandet, A. Mehrjou, S. K. Lee, and A. Raj, “Dual instrumental variable regression,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 2710–2721, 2020.
- [466] N. Dikkala, G. Lewis, L. Mackey, and V. Syrgkanis, “Minimax estimation of conditional moment models,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 248–12 262, 2020.
- [467] Y. Zhu, L. Gultchin, A. Gretton, M. Kusner, and R. Silva, “Causal inference with treatment measurement error: A nonparametric instrumental variable approach,” in *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022. [Online]. Available: <https://openreview.net/forum?id=SLcxbOUi9gq>
- [468] D. M. Chickering, *Learning Bayesian Networks is NP-Complete*. New York, NY: Springer New York, 1996, pp. 121–130.
- [469] A. P. Singh and A. W. Moore, *Finding optimal Bayesian networks by dynamic programming*. Citeseer, 2005.

- [470] J. Xiang and S. Kim, “A* lasso for learning a sparse bayesian network structure for continuous variables,” in *Advances in Neural Information Processing Systems*, vol. 26, 2013.
- [471] J. Cussens, “Bayesian network learning with cutting planes,” in *Uncertainty in Artificial Intelligence*, 2011.
- [472] C. Squires and C. Uhler, “Causal structure learning: a combinatorial perspective,” 2022. [Online]. Available: <https://arxiv.org/abs/2206.01152>
- [473] M. Scanagatta, C. P. de Campos, G. Corani, and M. Zaffalon, “Learning bayesian networks with thousands of variables,” in *Advances in Neural Information Processing Systems*, vol. 28, 2015.
- [474] B. Aragam and Q. Zhou, “Concave penalized estimation of sparse gaussian bayesian networks,” *The Journal of Machine Learning Research*, vol. 16, 2015.
- [475] J. D. Ramsey, M. Glymour, R. Sanchez-Romero, and C. Glymour, “A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images,” *International Journal of Data Science and Analytics*, vol. 3, 2017.
- [476] X. Zheng, B. Aragam, P. Ravikumar, and E. P. Xing, “Dags with NO TEARS: continuous optimization for structure learning,” in *Advances in Neural Information Processing Systems*, 2018.
- [477] Y. Yu, J. Chen, T. Gao, and M. Yu, “DAG-GNN: DAG structure learning with graph neural networks,” in *ICML*, vol. 97, 2019.
- [478] X. Zheng, C. Dan, B. Aragam, P. Ravikumar, and E. Xing, “Learning sparse nonparametric dags,” in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 3414–3425.
- [479] N. R. Ke, O. Bilaniuk, A. Goyal, S. Bauer, H. Larochelle, B. Schölkopf, M. C. Mozer, C. Pal, and Y. Bengio, “Learning neural causal models from unknown interventions,” 2019. [Online]. Available: <https://arxiv.org/abs/1910.01075>
- [480] I. Ng, A. Ghassami, and K. Zhang, “On the role of sparsity and DAG constraints for learning linear dags,” in *NeurIPS*, 2020.
- [481] P. Brouillard, S. Lachapelle, A. Lacoste, S. Lacoste-Julien, and A. Drouin, “Differentiable causal discovery from interventional data,” in *NeurIPS*, 2020.
- [482] Y. He, P. Cui, Z. Shen, R. Xu, F. Liu, and Y. Jiang, “DARING: differentiable causal discovery with residual independence,” in *KDD*, 2021.
- [483] P. Lippe, T. Cohen, and E. Gavves, “Efficient neural causal discovery without acyclicity constraints,” in *International Conference on Learning Representations*, 2022. [Online]. Available: <https://openreview.net/forum?id=eYciPrLuUhG>

- [484] N. Friedman and D. Koller, “Being bayesian about network structure. a bayesian approach to structure discovery in bayesian networks,” *Machine learning*, vol. 50, 2003.
- [485] M. Gao, Y. Ding, and B. Aragam, “A polynomial-time algorithm for learning nonparametric causal graphs,” in *Advances in Neural Information Processing Systems*, 2020.
- [486] C. Cundy, A. Grover, and S. Ermon, “BCD nets: Scalable variational approaches for bayesian causal discovery,” in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021.
- [487] V. Zantedeschi, J. Kaddour, L. Franceschi, M. Kusner, and V. Niculae, “DAG learning on the permutohedron,” in *ICLR2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality*, 2022.
- [488] M. J. Vowels, N. C. Camgoz, and R. Bowden, “D’ya like dags? a survey on structure learning and causal discovery,” *arXiv preprint arXiv:2103.02582*, 2021.
- [489] S. Magliacane, T. Claassen, and J. M. Mooij, “Ancestral causal inference,” in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, Eds., 2016, pp. 4466–4474. [Online]. Available: <https://proceedings.neurips.cc/paper/2016/hash/f3d9de86462c28781cbe5c47ef22c3e5-Abstract.html>
- [490] J. M. Mooij and T. Claassen, “Constraint-based causal discovery using partial ancestral graphs in the presence of cycles,” in *Conference on Uncertainty in Artificial Intelligence*. PMLR, 2020, pp. 1159–1168.
- [491] D. Watson and R. Silva, “Causal discovery under a confounder blanket,” in *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022. [Online]. Available: <https://openreview.net/forum?id=S0eeRL8jcec>
- [492] F. Eberhardt, “Causal discovery as a game,” in *Causality: Objectives and Assessment*. PMLR, 2010, pp. 87–96.
- [493] A. Hyttinen, F. Eberhardt, and P. O. Hoyer, “Experiment selection for causal discovery,” *Journal of Machine Learning Research*, vol. 14, pp. 3041–3071, 2013.
- [494] J. von Kügelgen, P. K. Rubenstein, B. Schölkopf, and A. Weller, “Optimal experimental design via bayesian optimization: active causal structure learning for gaussian process networks,” *arXiv preprint arXiv:1910.03962*, 2019.
- [495] N. Scherrer, O. Bilaniuk, Y. Annadani, A. Goyal, P. Schwab, B. Schölkopf, M. C. Mozer, Y. Bengio, S. Bauer, and N. R. Ke, “Learning neural

- causal models with active interventions,” 2021. [Online]. Available: <https://arxiv.org/abs/2109.02429>
- [496] P. A. Stokes and P. L. Purdon, “A study of problems encountered in granger causality analysis from a neuroscience perspective,” *Proceedings of the National Academy of Sciences*, vol. 114, no. 34, pp. E7063–E7072, 2017.
- [497] H. Lütkepohl and M. Krätsig, *Applied time series econometrics*. Cambridge university press, 2004.
- [498] C. Krupitzer, M. Pfannemüller, J. Kaddour, and C. Becker, “Satisfy: Towards a self-learning analyzer for time series forecasting in self-improving systems,” in *2018 IEEE 3rd International Workshops on Foundations and Applications of Self* Systems (FAS* W)*. IEEE, 2018, pp. 182–189.
- [499] C. W. Granger, “Investigating causal relations by econometric models and cross-spectral methods,” *Econometrica: journal of the Econometric Society*, pp. 424–438, 1969.
- [500] A. Tank, I. Covert, N. Foti, A. Shojaie, and E. Fox, “Neural granger causality,” *arXiv preprint arXiv:1802.05842*, 2018.
- [501] T. Wu, T. Breuel, M. Skuhersky, and J. Kautz, “Discovering nonlinear relations with minimum predictive information regularization,” 2020. [Online]. Available: <https://arxiv.org/abs/2001.01885>
- [502] S. Khanna and V. Y. F. Tan, “Economy statistical recurrent units for inferring nonlinear granger causality,” in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: <https://openreview.net/forum?id=SyxV9ANFDH>
- [503] S. Löwe, D. Madras, R. Zemel, and M. Welling, “Amortized causal discovery: Learning to infer causal graphs from time-series data,” *arXiv preprint arXiv:2006.10833*, 2020.