FROM STATISTICAL TO CAUSAL LEARNING

BERNHARD SCHÖLKOPF AND
JULIUS VON KÜGELGEN

ABSTRACT

We describe basic ideas underlying research to build and understand artificially intelligent systems: from symbolic approaches via statistical learning to interventional models relying on concepts of causality. Some of the hard open problems of machine learning and AI are intrinsically related to causality, and progress may require advances in our understanding of how to model and infer causality from data.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 68T05; Secondary 68Q32, 68T01, 68T10, 68T30, 68T37

KEYWORDS

Causal inference, machine learning

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5540–5593
DOI 10.4171/ICM2022/173

1. INTRODUCTION

In 1958, the New York Times reported on a new machine called the *perceptron*. Frank Rosenblatt, its inventor, demonstrated that the perceptron was able to learn from experience. He predicted that later perceptrons would be able to recognize people, or instantly translate spoken language. Now a reality, this must have sounded like distant science fiction at the time. In hindsight, we may consider it the birth of machine learning, the field fueling most of the current advances in artificial intelligence (AI).

Around the same time, another equally revolutionary development took place: scientists understood that computers could do more than compute numbers: they can process symbols. Although this insight was also motivated by artificial intelligence, in hindsight it was the birth of the field of computer science. There was great optimism that the manipulation of symbols, in programs written by humans, implementing rules designed by humans, should be enough to generate intelligence. Below, we shall refer to this as the *symbol–rule hypothesis*.[1]

There was initially encouraging progress on seemingly hard problems such as automatic theorem proving and computer chess. One of the fathers of the field, Herb Simon, predicted in 1956 that "machines will be capable, within twenty years, of doing any work a man can do." However, problems that appeared simple, such as most things animals could do, turned out to be hard. This came to be known as *Moravec's paradox*. When IBM's *Deep Blue* chess computer beat Garry Kasparov in 1997, Kasparov was physically facing a human during the match: while Deep Blue was capable of analyzing the game's search tree in unprecedented detail, it was unable to recognize and physically move chess pieces, so this task had to be relegated to a human, in an inversion of the famous *mechanical turk*.[2]

In the years to follow, the field of AI entered what came to be known as the *AI winter*. The community got disillusioned with the lack of progress and prospects, and interest greatly declined. However, largely independently of the field of classic AI, *machine learning* eventually started to boom. Like Rosenblatt's early work, it was built on the observation that all existing examples of truly intelligent systems—i.e., animals, including humans—were not built on the symbol–rule hypothesis: both the representations and the rules implemented by natural intelligent systems are acquired from experience, through processes of evolution and learning.

Rather than exploring the well-known dichotomy between rule- and learning-based approaches, we will explore the less known questions of causality and interventions. While the field of causality in computer science was initially strongly linked to classic AI, recent years have witnessed great interest in connecting it to machine learning [111]. Below, we explore some of these connections, drawing from [125, 133]. We will argue that the causal view is relevant when it comes to addressing crucial open problems of machine learning, related to notions of robustness and generalization beyond the training distribution.

---

1 The term should be taken with a grain of salt, since it suggests a separation between representations and computations which is hard to uphold in practice.

2 https://en.wikipedia.org/wiki/Mechanical_Turk.

Overview. In statistical learning, our starting point is a joint distribution $p(\mathbf{X})$ generating the observable data. Here, $\mathbf{X}$ is a random vector, and we are usually given a dataset $\mathbf{x}_1, \ldots, \mathbf{x}_m$ sampled i.i.d. from p. We are often interested in estimating properties of conditionals of some components of $\mathbf{X}$ given others, e.g., a classifier (which may be obtained by thresholding a conditional at 0.5). This is a nontrivial inverse problem, giving rise to statistical learning theory (Section 2).

Causal learning is motivated by shortcomings of statistical learning (Section 3). Its starting point is a structural causal model (SCM) [104] (Section 4). In an SCM, the components $X_1, \ldots, X_n$ of $\mathbf{X}$ are identified with vertices of a directed graph whose arrows represent direct causal influences, and there is a random variable U_i for each vertex, along with a function f_i which computes X_i from its graph parents $\mathbf{PA}_i$ and U_i, i.e.,

$$X_i := f_i(\mathbf{PA}_i, U_i). \tag{1.1}$$

Given a distribution over the U_i, which are assumed independent, this also gives rise to a probabilistic model $p(\mathbf{X})$. However, the model in (1.1) is more structured: the graph connectivity and the functions f_i create particular dependences between the observables. Moreover, it describes how the system behaves under intervention: by replacing functions by constants, we can compute the effect of setting some variables to specific values.

Causal learning builds on assumptions different from standard machine learning (Section 5), and addresses a different level in the modeling hierarchy (Section 6). It also comes with new problems, such as causal discovery, where we seek to infer properties of graph and functions from data (Section 7). In some cases, conditional independences among the X_i contain information about the graph [144]; but novel assumptions let us handle some cases that were previously unsolvable [68]. Those assumptions have nontrivial implications for machine learning tasks such as semisupervised learning, covariate shift adaptation and transfer learning [128] (Section 8). Once provided with a causal model, causal reasoning (Section 9) allows us to identify and estimate certain causal queries of interest from observational data. We conclude with a list of some current and open problems (Section 10), with a particular emphasis on the topic of causal representation learning.

The presentation and notation will be somewhat informal in several respects. We generally assume that all distributions possess densities (with respect to a suitable reference measure). We sometimes write $p(x)$ for the distribution (or density) of a random variable X. Accordingly, the same p can denote another distribution $p(y)$, distinguished by the argument of $p(\cdot)$. We also sometimes use summation for marginalization which supposes discrete variables; the corresponding expressions for continuous quantities would use integrals.

2. STATISTICAL LEARNING THEORY

Suppose we have measured two statistically dependent observables and found the points to lie approximately on a straight line. An empirical scientist might be willing to hypothesize a corresponding law of nature (see Figure 1). However, already Leibniz pointed out that if we scatter spots of ink randomly on a piece of paper by shaking a quill pen, we

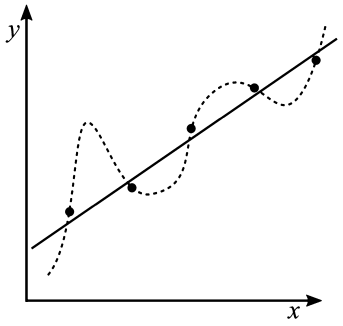

FIGURE 1
Given a small number of observations, how do we find a law underlying them? Leibniz argued that even if we generate a random set of points, we can always find a mathematical equation satisfied by these points.

can also find a mathematical equation satisfied by these points [81]. He argued that we would not call this a law of nature, because no matter how the points are distributed, there always exists such an equation; we would only call it a law of nature only if the equation is simple. This raises the question of what makes an equation simple. The physicist Rutherford took the pragmatic view that if there is a law, it should be directly evident from the data: "if your experiment needs statistics, you ought to have done a better experiment."[3] This view may have been a healthy one when faced with low-dimensional inference problems where regularities are immediately obvious; however, modern AI is facing inference problems that are harder: they are often high-dimensional and nonlinear, yet we may have little prior knowledge about the underlying regularity (e.g., for medical data, we usually do not have a mechanistic model).

Statistical learning theory studies the problem of how to still perform valid inference, provided that we have sufficiently large datasets and the computational means to process them. Let us look at some theoretical results for the simplest learning scenario, drawing from [130]; for details, see [153]. Suppose we are given empirical observations,

$$(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X} \times \mathcal{Y}, \tag{2.1}$$

where $\mathcal{X}$ is some nonempty set from which the *inputs* come, and $\mathcal{Y} = \{\pm 1\}$ is the *output* set, in our case consisting of just two *classes*. This situation is called *pattern recognition*, and our goal is to use the *training data* (2.1) to infer a function $f : \mathcal{X} \to \{\pm 1\}$ (from some function class chosen a priori) which will produce the correct output for a new input x which we may not have seen before. To formalize what we mean by "correct," we make the assumption that all observations (x_i, y_i) have been generated independently by performing a random experiment described by an unknown probability distribution $p(x, y)$—a setting referred to as *i.i.d. (independent and identically distributed) data*. Our goal will be to minimize the

3 Cited after http://www.warwick.ac.uk/statsdept/staff/JEHS/data/jehsquot.pdf.

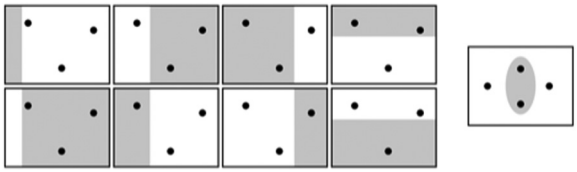

FIGURE 2

Using straight lines, we can separate three points in all possible ways; we cannot do this for four points, no matter how they are placed. The class of linear separations is not "falsifiable" using three points, but it becomes falsifiable once we have four or more points.

expected error (or risk)

$$R[f] = \int_{\mathcal{X}\times\mathcal{Y}} c(y, f(x))dp(x, y), \tag{2.2}$$

where c is a so-called loss function, e.g., the misclassification error $c(y, f(x)) = \frac{1}{2}|f(x) - y|$ taking the value 0 whenever $f(x) = y$ and 1 otherwise.

The difficulty of the task stems from the fact that we are trying to minimize a quantity that we cannot evaluate: since we do not know p, we cannot compute (2.2). We do know, however, the training data (2.1) sampled from p. We can thus try to infer a function f from the training sample whose risk is close to the minimum of (2.2). To this end, we need what is called an *induction principle*.

One way to proceed is to use the training sample to approximate (2.2) by a finite sum, referred to as the *empirical risk*

$$R_{\text{emp}}[f] = \frac{1}{m}\sum_{i=1}^{m} c(x_i, y_i, f(x_i)). \tag{2.3}$$

The *empirical risk minimization (ERM) induction principle* recommends that we choose (or "learn") an f that minimizes (2.3). We can then ask whether the ERM principle is statistically *consistent*: in the limit of infinitely many data points, will ERM lead to a solution which will do as well as possible on future data generated by p?

It turns out that if the function class over which we minimize (2.3) is too large, then ERM is not consistent. Hence, we need to suitably restrict the class of possible functions. For instance, ERM is consistent for all probability distributions, provided that the *VC dimension* of the function class is finite. The VC dimension is an example of a *capacity measure*. It is defined as the maximal number of points that can be separated (classified) in all possible ways using functions from the class. For example, using linear classifiers (separating classes by straight lines) on $\mathbb{R}^2$, we can realize all possible classifications for 3 suitably chosen points, but we can no longer do this once we have 4 points, no matter how they are placed (see Figure 2). This means that the VC dimension of this function class is 3. More generally, for linear separations in $\mathbb{R}^d$, the VC dimension is $d + 1$.

Whenever the VC dimension is finite, our class of functions (or explanations) becomes falsifiable in the sense that starting from a certain number of observations, no

longer all possible labelings of the points can be explained (cf. Figure 2). If we can nevertheless explain a sufficiently large set of observed data, we thus have reason to believe that this is a meaningful finding.

Much of machine learning research is concerned with restrictions on classes of functions to make inference possible, be it by imposing prior distributions on function classes, through other constraints, or by designing self-regularizing learning procedures, e.g., gradient descent methods for neural networks [79]. While there is a solid theoretical understanding of supervised machine learning as described above (i.e., function learning from input–output examples), there are still details under investigation, such as the recently observed phenomenon of "double descent" [7].

A popular constraint, implemented in the *Support Vector Machine (SVM)* [130,153], is to consider linear separations with large margin: it turns out that for large margin separations in high-dimensional (or infinite-dimensional) spaces, the capacity can be much smaller than the dimensionality, making learning possible in situations where it would otherwise fail.

For some learning algorithms, including SVMs and nearest neighbor classifiers, there are strong universal consistency results, guaranteeing convergence of the algorithm to the lowest achievable risk, for any problem to be learned [28,130,146,153]. Note, however, that this convergence can be arbitrarily slow.

For a given sample size, it will depend on the problem being learned whether we achieve low expected error. In addition to asymptotic consistency statements, learning theory makes finite sample size statements: one can prove that with probability at least $1-\delta$ (for $\delta > 0$), for all functions f in a class of functions with VC dimension h,

$$R[f] \leq R_{\text{emp}}[f] + \sqrt{\frac{1}{m}\left(h(\log(2m/h)+1) + \log\frac{4}{\delta}\right)}. \tag{2.4}$$

This is an example of a class of results that relate the training error $R_{\text{emp}}[f]$ and the test error $R[f]$ using a confidence interval (the square root term) depending on a capacity measure of a function class (here, its VC dimension h). It says that with high probability, the expected error $R[f]$ on future observations generated by the unknown probability distribution is small, provided the two terms on the right-hand side are small: the training error $R_{\text{emp}}[f]$ (i.e., the error on the examples we have already seen), and the square root term, which will be small whenever the capacity h is small compared to the number of training observations m. If, on the other hand, we try to learn something that may not make sense, such as the mapping from the name of people to their telephone number, we would find that to explain all the training data (i.e., to obtain a small $R_{\text{emp}}[f]$), we need a model whose capacity h is large, and the second term becomes large. In any case, it is crucial for both consistency results and finite sample error bounds such as (2.4) that we have i.i.d. data.

Kernel methods. A symmetric function $k : \mathcal{X}^2 \to \mathbb{R}$, where $\mathcal{X}$ is a nonempty set, is called a positive definite (pd) *kernel* if for arbitrary points $x_1, \ldots, x_m \in \mathcal{X}$ and coefficients $a_1, \ldots, a_m \in \mathbb{R}$:

$$\sum_{i,j} a_i a_j k(x_i, x_j) \geq 0.$$

The kernel is called strictly positive definite if for pairwise distinct points, the implication $\sum_{i,j} a_i a_j k(x_i, x_j) = 0 \Longrightarrow \forall i : a_i = 0$ is valid. Any positive definite kernel induces a mapping

$$\Phi : x \mapsto k(x, \cdot) \tag{2.5}$$

into a *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}$ satisfying

$$\big\langle k(x, \cdot), k(x', \cdot) \big\rangle = k(x, x') \tag{2.6}$$

for all $x, x' \in \mathcal{X}$. Although $\mathcal{H}$ may be infinite-dimensional, we can construct practical classification algorithms in $\mathcal{H}$ provided that all computational steps are carried out in terms of scalar products, since those can be reduced to kernel evaluations (2.6).

In the SVM algorithm, the capacity of the function class is restricted by enforcing a large margin of class separation in $\mathcal{H}$ via a suitable RKHS regularization term. The solution can be shown to take the form

$$f(x) = \operatorname{sgn}\Big(\sum_i \alpha_i k(x_i, x) + b \Big), \tag{2.7}$$

where the learned parameters α_i and b are the solution of a convex quadratic optimization problem. A similar expansion of the solution in terms of kernel functions evaluated at training points holds true for a larger class of kernel algorithms beyond SVMs, regularized by an RKHS norm [126].

In kernel methods, the kernel plays three roles which are crucial for machine learning: it acts as a similarity measure for data points, induces a representation in a linear space[4] via (2.5), and parametrizes the function class within which the solution is sought, cf. (2.7).

Kernel mean embeddings. Consider two sets of points $X := \{x_1, \ldots, x_m\} \subset \mathcal{X}$ and $Y := \{y_1, \ldots, y_n\} \subset \mathcal{X}$. We define the mean map μ as [130]

$$\mu(X) = \frac{1}{m} \sum_{i=1}^{m} k(x_i, \cdot). \tag{2.8}$$

For polynomial kernels $k(x, x') = (\langle x, x' \rangle + 1)^d$, we have $\mu(X) = \mu(Y)$ if all empirical moments up to order d coincide. For strictly pd kernels, the means coincide only if $X = Y$, rendering μ injective [131]. The mean map has some other interesting properties [143], e.g., $\mu(X)$ represents the operation of taking a mean of a function on the sample X:

$$\big\langle \mu(X), f \big\rangle = \left\langle \frac{1}{m} \sum_{i=1}^{m} k(x_i, \cdot), f \right\rangle = \frac{1}{m} \sum_{i=1}^{m} f(x_i).$$

Moreover, we have

$$\big\| \mu(X) - \mu(Y) \big\| = \sup_{\|f\| \le 1} \big| \big\langle \mu(X) - \mu(Y), f \big\rangle \big| = \sup_{\|f\| \le 1} \left| \frac{1}{m} \sum_{i=1}^{m} f(x_i) - \frac{1}{n} \sum_{i=1}^{n} f(y_i) \right|.$$

4 Note that the data domain $\mathcal{X}$ need not have any structure other than being a nonempty set.

If $\mathbb{E}_{x,x'\sim p}[k(x,x')],\ \mathbb{E}_{x,x'\sim q}[k(x,x')] < \infty$, then the above statements, including the injectivity of μ, generalize to Borel measures p, q, if we define the mean map as

$$\mu : p \mapsto \mathbb{E}_{x\sim p}\big[k(x,\cdot)\big],$$

and replace the notion of strictly pd kernels by that of characteristic kernels [33]. This means that we do not lose information when representing a probability distribution in the RKHS. This enables us to work with distributions using Hilbert space methods, and construct practical algorithms analyzing distributions using scalar product evaluations.

Note that the mean map μ can be viewed as a generalization of the *moment generating function* M_p of a random variable x with distribution p,

$$M_p(\cdot) = \mathbb{E}_{x\sim p}\big[e^{\langle x,\cdot\rangle}\big].$$

The map μ has applications in a number of tasks, including computing functions of random variables [129] and testing for homogeneity [41] or independence [43]. The latter will be of particular interest to causal inference: we can develop a kernel-based independence test by computing the distance between sample-based embeddings of a joint distribution $p(X,Y)$ and the product of its marginals $p(X)p(Y)$ [42–44,114,165], and generalize it to conditional independence testing [33,100], as required for certain causal discovery methods (see Section 7).

3. FROM STATISTICAL TO CAUSAL MODELS

Methods relying on i.i.d. data. In current successes of machine learning [79], we generally *(i) have large amounts of data*, often from simulations or large-scale human labeling, *(ii) use high capacity machine learning models* (e.g., neural networks with many adjustable parameters), and *(iii) employ high performance computing*. Statistical learning theory offers a partial explanation for recent successes of learning: huge datasets enable training complex models and thus solving increasingly difficult tasks.

However, a crucial aspect that is often ignored is that we *(iv) assume that the data are i.i.d.* This assumption is crucial for good performance in practice, and it underlies theoretical statements such as (2.4). When faced with problems that violate the i.i.d. assumption, all bets are off. Vision systems can be grossly misled if an object that is normally recognized with high accuracy is placed in a context that *in the training set* may be negatively correlated with the presence of the object. For instance, such a system may fail to recognize a cow standing on the beach. In order to successfully generalize in such settings, we would need to construct systems which do not merely rely on statistical dependences, but instead model mechanisms that are robust across certain violations of the i.i.d. assumption. As we will argue, causality provides a natural framework for capturing such stable mechanisms and reasoning about different types of distribution shifts.

Correlation vs. causation. It is a commonplace that *correlation does not imply causation*. Two popular and illustrative examples are the positive correlation between chocolate consumption and Nobel prizes per capita [91], and that between the number of stork breeding

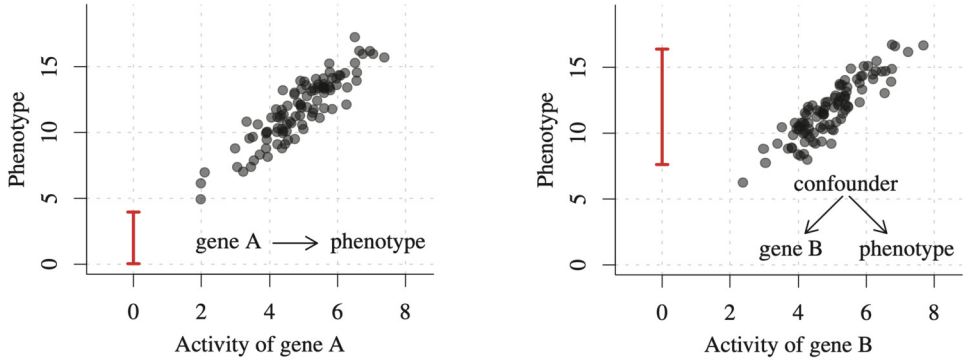

FIGURE 3

Measurements of two genes (x-axis), gene A (left) and gene B (right), show the same strong positive correlation with a phenotype (y-axis). However, this statistical information alone is insufficient to predict the outcome of a knock-out experiment where the activity of a gene is set to zero (vertical lines at $x = 0$). Answering such *interventional* questions requires additional causal knowledge (inset causal graphs): knocking out gene A, which is a direct cause, would lead to a reduction in phenotype, whereas knocking out gene B, which shares a common cause, or confounder, with the phenotype but has no causal effect on it, would leave the phenotype unaffected. This shows that correlation alone is not enough to predict the outcome of perturbations to a system (toy data, figure from [111]).

pairs and human birth rates [89], neither of which admit a sensible interpretation in terms of direct causation. These examples naturally lead to the following questions: What exactly do we mean by "causation"? What is its relationship to correlation? And, if correlation alone is not enough, what is needed to infer causation?

Here, we adopt a notion of causality based on manipulability [159] and intervention [104] which has proven useful in fields such as agriculture [161], econometrics [46, 52], and epidemiology [118].

Definition 3.1 (Causal effect). We say that a random variable X has a causal effect on a random variable Y if there exist $x \neq x'$ such that the distribution of Y after intervening on X and setting it to x differs from the distribution of Y after setting X to x'.

Inherent to the notion of causation, there is a directionality and asymmetry which does not exist for correlation: if X is correlated with Y, then Y is equally correlated with X; but, if X has a causal effect on Y, the converse (in the generic case) does not hold.

We illustrate the intervention-based notion of causation and its difference from correlation (or, more generally, statistical dependence) in Figure 3. Here, knocking out two genes X_A and X_B that are indistinguishable based on their correlation with a phenotype Y would have very different effects. Only intervening on X_A would change the distribution of Y, whereas X_B does not have a causal effect on Y—instead, their correlation arises from a different (confounded) causal structure. Such causal relationships are most commonly represented in the form of *causal graphs* where directed arrows indicate a direct causal effect.

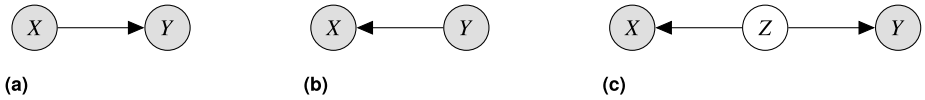

FIGURE 4

Reichenbach's common cause principle **[116]** postulates that statistical dependence between two random variables X and Y has three elementary possible causal explanations shown as causal graphs in (a)–(c). It thus states that association is always induced by an underlying causal process. In (a) the common cause Z coincides with X, and in (b) it coincides with Y. Grey nodes indicate observed and white nodes unobserved variables.

The example in Figure 3 shows that the same correlation can be explained by multiple causal graphs which lead to different experimental outcomes, i.e., *correlation does not imply causation*. However, there is a connection between correlation and causation, expressed by Reichenbach **[116]** as the Common Cause Principle, see Figure 4.

Principle 3.2 (Common cause). *If two random variables X and Y are statistically dependent ($X \not\perp\!\!\!\perp Y$), then there exists a random variable Z which causally influences both of them and which explains all their dependence in the sense of rendering them conditionally independent ($X \perp\!\!\!\perp Y \mid Z$). As a special case, Z may coincide with X or Y.*

According to Principle 3.2, statistical dependence always results from underlying causal relationships by which variables, including potentially unobserved ones, influence each other. Correlation is thus an epiphenomenon, the byproduct of a causal process.

For the example of chocolate consumption (X) and Nobel laureates (Y), common sense suggests that neither of the two variables should have a causal effect on the other, i.e., neither chocolate consumption driving scientific success ($X \to Y$; Figure 4a) nor Nobel laureates increasing chocolate consumption ($Y \to X$; Figure 4b) seem plausible. Principle 3.2 then tells us that the observed correlation must be explained by a common cause Z as in Figure 4c. A plausible candidate for such a confounder could, for example, be economic factors driving both consumer spending and investment in education and science.

Without such background knowledge or additional assumptions, however, we cannot distinguish the three cases in Figure 4 through passive observation, i.e., in a purely data-driven way: the class of observational distributions over X and Y that can be realized by these models is the same in all three cases.

To be clear, this does not mean that correlation cannot be useful, or that causal insight is always required. Both genes in Figure 3 remain useful features for making predictions in a passive, or *observational*, setting in which we measure the activities of certain genes and are asked to predict the phenotype. Similarly, chocolate consumption remains predictive of winning Nobel prizes. However, if we want to answer interventional questions, such as the outcome of a gene-knockout experiment or the effect of a policy enforcing higher chocolate consumption, we need more than correlation: *a causal model*.

4. CAUSAL MODELING FRAMEWORKS

Causal inference has a long history in a variety of disciplines, including statistics, econometrics, epidemiology, and AI. As a result, different frameworks for causal modeling have emerged over the years and coexist today. The first framework described below (CGM) starts from the distribution of the observables, combining it with a directed graph to endow it with causal semantics. The second (SCM) starts from a graph and a set of functional assignments, and generates the observed distribution as the push-forward of an unobserved noise distribution. Finally, we cover a nongraphical approach (PO) popular in statistics.

Causal graphical models (CGMs). The graphical models framework [75, 78] provides a compact way of representing joint probability distributions by encoding the dependence structure between variables in graphical form. Directed graphical models are also known as *Bayesian networks* [101]. While they do not offer a causal interpretation per se—indeed, different graphical models can be compatible with the same distribution (cf. Principle 3.2)—when edges are endowed with the notion of direct causal effect (Definition 3.1), we refer to them as causal graphical models (CGM) [144].

Definition 4.1 (CGM). A CGM $\mathcal{M} = (G, p)$ over n random variables $X_1, \ldots, X_n$ consists of: (i) a directed acyclic graph (DAG) G in which directed edges ($X_j \to X_i$) represent a direct causal effect of X_j on X_i; and (ii) a joint distribution $p(X_1, \ldots, X_n)$ which is Markovian with respect to G:

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i \mid \mathbf{PA}_i) \tag{4.1}$$

where $\mathbf{PA}_i = \{X_j : (X_j \to X_i) \in G\}$ denotes the set of parents, or direct causes, of X_i in G.

We will refer to (4.1) as the *causal (or disentangled) factorization*. While many other *entangled factorizations* are possible, e.g.,

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i \mid X_{i+1}, \ldots, X_n), \tag{4.2}$$

only (4.1) decomposes the joint distribution into causal conditionals, or *causal mechanisms*, $p(X_i \mid \mathbf{PA}_i)$, which can have a meaningful physical interpretation, rather than being mere mathematical objects such as the factors on the RHS of (4.2).

It turns out that (4.1) is equivalent to the following condition.

Definition 4.2 (Causal Markov condition). A distribution p satisfies the causal Markov condition with respect to a DAG G if every variable is conditionally independent of its nondescendants in G given its parents in G.

Definition 4.2 can equivalently be expressed in terms of *d-separation*, a graphical criterion for directed graphs [104], by saying that *d-separation in G implies (conditional) independence in p*. The causal Markov condition thus provides a link between properties of p and G.

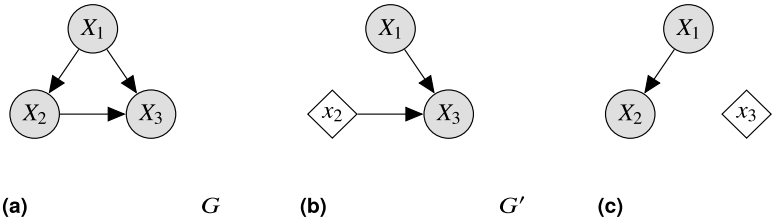

FIGURE 5

(a) A directed acyclic graph (DAG) G over three variables. A causal graphical model (G, p) with causal graph G and observational distribution p can be used to answer interventional queries using the concept of *graph surgery*: when a variable is intervened upon and set to a constant (white diamonds), this removes any influence from other variables, captured graphically by removing all incoming edges. (b) and (c) show postintervention graphs G' and G'' for $\mathrm{do}(X_2 := x_2)$ and $\mathrm{do}(X_3 := x_3)$, respectively. (An intervention on X_1 would leave the graph unaffected.)

What makes CGMs causal is the interpretation of edges as cause–effect relationships which enables reasoning about the outcome of interventions using the *do-operator* [104] and the concept of *graph surgery* [144]. The central idea is that intervening on a variable, say by externally forcing it to take on a particular value, renders it independent of its causes and breaks their causal influence on it, see Figure 5 for an illustration. For example, if a gene is knocked out, it is no longer influenced by other genes that were previously regulating it; instead, its activity is now solely determined by the intervention. This is fundamentally different from conditioning since passively observing the activity of a gene provides information about its driving factors (i.e., its direct causes).

To emphasize this difference between passive observation and active intervention, Pearl [104] introduced the notation $\mathrm{do}(X := x)$ to denote an intervention by which variable X is set to value x. The term *graph surgery* refers to the idea that the effect of such an intervention can be captured in the form of a modification to the original graph by removing all incoming edges to the intervened variable. Interventional queries can then be answered by performing probabilistic inference in the modified postintervention graph which typically implies additional (conditional) independences due to the removed edges.

Example 4.3. The interventional distribution $p(X_3|\,\mathrm{do}(X_2 := x_2))$ for the CGM in Figure 5 is obtained via probabilistic inference with respect to the postintervention graph G' where $X_1 \perp\!\!\!\perp X_2$:

$$p\big(X_3|\,\mathrm{do}(X_2 := x_2)\big) = \sum_{x_1 \in \mathcal{X}_1} p(x_1)p(X_3|x_1, x_2) \tag{4.3}$$

$$\neq \sum_{x_1 \in \mathcal{X}_1} p(x_1|x_2)p(X_3|x_1, x_2) = p(X_3|x_2). \tag{4.4}$$

It differs from the conditional $p(X_3|x_2)$ for which inference in done over G where $X_1 \not\perp\!\!\!\perp X_2$. Note the marginal $p(x_1)$ in (4.3), in contrast to the conditional $p(x_1|x_2)$ in (4.4): this is precisely the link which is broken by the intervention $\mathrm{do}(X_2 := x_2)$, see Figure 5b. The right-

hand side of (4.3) is an example of covariate adjustment: it controls for the confounder X_1 of the causal effect of X_2 on X_3, see Section 9 for more details on adjustment and computing interventions.

CGMs have been widely used in constraint- and score-based approaches to causal discovery [47,144] which we will discuss in Section 7. Due to their conceptual simplicity, they are a useful and intuitive model for reasoning about interventions. However, their capacity as a causal model is limited in that they do not support *counterfactual* reasoning, which is better addressed by the two causal modeling frameworks which we will discuss next.

Structural causal models (SCMs). Structural causal models, also referred to as functional causal models or nonparametric structural equation models, have ties to the graphical approach presented above, but rely on using directed functional parent–child relationships rather than causal conditionals. While conceptually simple in hindsight, this constituted a major step in the understanding of causality, as later expressed by [104, PAGE 104]:

> *"We played around with the possibility of replacing the parents–child relationship* $p(X_i|\mathbf{PA}_i)$ *with its functional counterpart* $X_i = f_i(\mathbf{PA}_i, U_i)$ *and, suddenly, everything began to fall into place: We finally had a mathematical object to which we could attribute familiar properties of physical mechanisms instead of those slippery epistemic probabilities* $p(X_i|\mathbf{PA}_i)$ *with which we had been working so long in the study of Bayesian networks."*

Definition 4.4 (SCM). An SCM $\mathcal{M} = (\mathbf{F}, p_{\mathbf{U}})$ over a set $\mathbf{X}$ of n random variables $X_1, \ldots, X_n$ consists of (i) a set $\mathbf{F}$ of n assignments (the structural equations),

$$\mathbf{F} = \left\{X_i := f_i(\mathbf{PA}_i, U_i)\right\}_{i=1}^{n} \tag{4.5}$$

where f_i are deterministic functions computing each variable X_i from its causal parents $\mathbf{PA}_i \subseteq \mathbf{X} \setminus \{X_i\}$ and an exogenous noise variable U_i; and (ii) a joint distribution $p_{\mathbf{U}}(U_1, \ldots, U_n)$ over the exogenous noise variables.

The paradigm of SCMs views the processes f_i by which each observable X_i is generated from others as a physical mechanism. All randomness comes from the unobserved (also referred to as *unexplained*) noise terms U_i which capture both possible stochasticity of the process, as well as uncertainty due to unmeasured parts of the system.

Note also the assignment symbol ":=" which is used instead of an equality sign to indicate the asymmetry of the causal relationship: the left-hand side quantity is defined to take on the right-hand side value. For example, we cannot simply rewrite a structural equation $X_2 := f_2(X_1, U_2)$ as $X_1 = g(X_2, U_2)$ for some g, as would be the case for a standard (invertible) equation.

In parametric, linear form (i.e., with linear f_i), SCMs are also known as structural equation models and have a long history in path analysis [161] and economics [46,52].

Each SCM induces a corresponding causal graph via the input variables to the structural equations which is useful as a representation and provides a link to CGMs.

Definition 4.5 (Induced causal graph). The causal graph G induced by an SCM $\mathcal{M}$ is the directed graph with vertex set $\mathbf{X}$ and a directed edge from each vertex in $\mathbf{PA}_i$ to X_i for all i.

Example 4.6. Consider an SCM over $\mathbf{X} = \{X_1, X_2, X_3\}$ with some $p_{\mathbf{U}}(U_1, U_2, U_3)$ and

$$X_1 := f_1(U_1), \quad X_2 := f_2(X_1, U_2), \quad X_3 := f_3(X_1, X_2, U_3).$$

Following Definition 4.5, the induced graph then corresponds to G in Figure 5.

Definition 4.4 allows for a rich class of causal models, including those with cyclic causal relations and ones which do not obey the causal Markov condition (Definition 4.2) due to complex covariance structures between the noise terms. While work exists on such cyclic or confounded SCMs [13], it is common to make the following two assumptions.

Assumption 4.7 (Acyclicity). The induced graph G is a DAG: it does not contain cycles.

Assumption 4.8 (Causal sufficiency/no hidden confounders). The U_i are jointly independent, i.e., their distribution factorizes, $p_{\mathbf{U}}(U_1, \ldots, U_n) = p_{U_1}(U_1) \times \cdots \times p_{U_n}(U_n)$.

Assumption 4.7 implies[5] the existence of a well-defined, unique (observational) distribution over $\mathbf{X}$ from which we can draw via *ancestral sampling*:[6] first, we draw the noise variables from $p_{\mathbf{U}}$, and then we iteratively compute the corresponding X_i's in topological order of the induced DAG (i.e., starting at the root node of the graph), substituting previously computed X_i into the structural equations where necessary. Formally, the (observational) distribution $p(X_1, \ldots, X_n)$ induced by an SCM under Assumption 4.7 is defined as the push-forward of the noise distribution $p_{\mathbf{U}}$ through the structural equations $\mathbf{F}$. Under Assumption 4.8, the causal conditionals are thus given by

$$p(X_i | \mathbf{PA}_i = \mathbf{pa}_i) := p_{U_i}\left(f_{\mathbf{pa}_i}^{-1}(X_i)\right) \quad \text{for } i = 1, \ldots, n, \tag{4.6}$$

where $f_{\mathbf{pa}_i}^{-1}(X_i)$ denotes the preimage of X_i under f_i for fixed $\mathbf{PA}_i = \mathbf{pa}_i$.

Assumption 4.8 rules out the existence of hidden confounders because any unmeasured variables affecting more than one of the X_i simultaneously would constitute a dependence between some of the noise terms (which account for any external, or exogenous, influences not explained by the observed X_i). In combination with Assumption 4.7, Assumption 4.8 (also known as *causal sufficiency*) thus ensures that the distribution induced by an SCM factorizes according to its induced causal graph as in (4.1). In other words, it guarantees that the causal Markov condition is satisfied with respect to the induced causal graph [104]. Below, unless explicitly stated otherwise, we will assume causal sufficiency.

Due to the conceptual similarity between interventions and the assignment character of structural equations, the computation of interventional distributions fits in naturally

5 Acyclicity is a sufficient, but not a necessary condition.

6 Since neither $\mathbf{F}$ nor p are known a priori, ancestral sampling should be seen as a hypothetical sampling procedure; inference and learning are still necessary in general.

into the SCM framework. To model an intervention, we simply replace the corresponding structural equation and consider the resulting entailed distribution.

Definition 4.9 (Interventions in SCMs). An intervention $\mathrm{do}(X_i := x_i)$ in an SCM $\mathcal{M} = (\mathbf{F}, p_U)$ is modeled by replacing the ith structural equation in $\mathbf{F}$ by $X_i := x_i$, yielding the intervened SCM $\mathcal{M}^{\mathrm{do}(X_i:=x_i)} = (\mathbf{F}', p_U)$. The interventional distribution $p(\mathbf{X}_{-i} | \mathrm{do}(X_i := x_i))$, where $\mathbf{X}_{-i} = \mathbf{X} \setminus \{X_i\}$, and intervention graph G' are those induced by $\mathcal{M}^{\mathrm{do}(X_i:=x_i)}$.

This way of handling interventions coincides with that for CGMs, e.g., after performing $\mathrm{do}(X_2 := x_2)$ in Example 4.6, X_1 no longer appears in the structural equation for X_2, and the edge $X_1 \to X_2$ hence disappears in the intervened graph, as is the case for G' in Figure 5.

In contrast to CGMs, SCMs also provide a framework for *counterfactual reasoning*. While (i) observations describe what is passively seen or measured and (ii) interventions describe active external manipulation or experimentation, (iii) counterfactuals are statements about what would or could have been, given that something else was in fact observed. These three modes of reasoning are sometimes referred to as the three rungs of the "ladder of causation" [107]. As an example, consider the following counterfactual query:

> *Given that patient X received treatment A and his/her health got worse, what would have happened if he/she had been given treatment B instead, all else being equal?*

The "all else being equal" part highlights the difference between interventions and counterfactuals: observing the factual outcome (i.e., what actually happened) provides information about the background state of the system (as captured by the noise terms in SCMs) which can be used to reason about alternative, counterfactual, outcomes. This differs from an intervention where such background information is not available. For example, observing that treatment A did not work may tell us that the patient has a rare condition and that treatment B would have therefore worked. However, given that treatment A has been prescribed, the patient's condition may have changed, and B may no longer work in a future intervention.

Note that counterfactuals cannot be observed empirically by their very definition and are therefore unfalsifiable. Some therefore consider them unscientific [115] or at least problematic [26]. On the other hand, humans seem to perform counterfactual reasoning in practice, developing this ability in early childhood [14].

Counterfactuals are computed in SCMs through the following three-step procedure:

1. Update the noise distribution to its posterior given the observed evidence ("abduction").
2. Manipulate the structural equations to capture the hypothetical intervention ("action").
3. Use the modified SCM to infer the quantity of interest ("prediction").

Definition 4.10 (Counterfactuals in SCMs). Given evidence $\mathbf{X} = \mathbf{x}$ observed from an SCM $\mathcal{M} = (\mathbf{F}, p_{\mathbf{U}})$, the counterfactual SCM $\mathcal{M}^{\mathbf{X}=\mathbf{x}}$ is obtained by updating $p_{\mathbf{U}}$ with its posterior: $\mathcal{M}^{\mathbf{X}=\mathbf{x}} = (\mathbf{F}, p_{\mathbf{U}|\mathbf{X}=\mathbf{x}})$. Counterfactuals are then computed by performing interventions in the counterfactual SCM $\mathcal{M}^{\mathbf{X}=\mathbf{x}}$, see Definition 4.9.

Note that while computing interventions only involved manipulating the structural equations, counterfactuals also involve updating the noise distribution, highlighting the conceptual difference between the two. Updating p_U requires knowledge of the interaction between noise and observed variables, i.e., of the structural equations, which explains why additional assumptions are necessary. Note that the updated noise variables no longer need to be independent, even if the original system was causally sufficient (Assumption 4.8).

Example 4.11 (Computing counterfactuals with SCMs). Consider an SCM $\mathcal{M}$ defined by

$$X := U_X, \quad Y := 3X + U_Y, \quad U_X, U_Y \sim \mathcal{N}(0, 1). \tag{4.7}$$

Suppose we observe $X = 2$ and $Y = 6.5$ and want to answer the counterfactual "what would Y have been, had $X = 1$?," i.e., we are interested in $p(Y_{X=1}|X = 2, Y = 6.5)$. Updating the noise using the observed evidence via (4.7), we obtain the counterfactual SCM $\mathcal{M}^{X=2,Y=6.5}$,

$$X := U_X, \quad Y := 3X + U_Y, \quad U_X \sim \delta(2), \quad U_Y \sim \delta(0.5), \tag{4.8}$$

where $\delta(\cdot)$ denotes the Dirac delta measure. Performing the intervention $\mathrm{do}(X := 1)$ in (4.8) then gives the result $p(Y_{X=1}|X = 2, Y = 6.5) = \delta(3.5)$, i.e., "$Y$ would have been 3.5." This differs from the interventional distribution $p(Y\,|\,\mathrm{do}(X = 1)) = \mathcal{N}(3, 1)$, since the factual observation helped determine the background state ($U_X = 2$, $U_Y = 0.5$).

The SCM viewpoint is intuitive and lends itself well to studying restrictions on function classes to enable induction (Section 2). For this reason, we will mostly focus on SCMs in the subsequent sections.

Potential outcomes (PO). The potential outcomes framework was initially proposed by Neyman [98] for randomized studies [31], and later popularized and extended to observational settings by Rubin [124] and others. It is popular within statistics and epidemiology and perhaps best understood in the context of the latter. This is also reflected in its terminology: in the most common setting, we consider a binary treatment variable T, with $T = 1$ and $T = 0$ corresponding to treatment and control, respectively, whose causal effect on an outcome variable Y (often a measure of health) is of interest.

One interpretation of the PO framework consistent with its roots in statistics is to view *causal inference as a missing data problem*. In the PO framework, for each individual (or unit) i and treatment value t there is a PO, or potential response, denoted $Y_i(t)$ capturing what would happen if individual i received treatment t. The POs are considered deterministic quantities in the sense that for a given individual i, $Y_i(1)$ and $Y_i(0)$ are fixed and all randomness in the realized outcome Y_i stems from randomness in the treatment assignment,

$$Y_i = TY_i(1) + (1 - T)Y_i(0). \tag{4.9}$$

TABLE 1

Causal inference as a missing data problem: for each individual i (rows), only the PO $Y_i(T_i)$ corresponding to the assigned treatment T_i is observed; the other PO is a counterfactual. Hence, the unit-level causal effect $\tau_i = Y_i(1) - Y_i(0)$ is unidentifiable.

i	T_i	$Y_i(1)$	$Y_i(0)$	τ_i
1	1	7	**?**	?
2	0	**?**	8	?
3	1	3	**?**	?
4	1	6	**?**	?
5	0	**?**	4	?
6	0	**?**	1	?

To decide whether patient i should receive treatment, we need to reason about the *individualized treatment effect* (ITE) τ_i as captured by the difference of the two POs.

Definition 4.12 (ITE). The ITE for individual i under a binary treatment is defined as

$$\tau_i = Y_i(1) - Y_i(0). \tag{4.10}$$

The *"fundamental problem of causal inference"* [51] is that only one of the POs is ever observed for each i. The other, unobserved PO becomes a counterfactual,

$$Y_i^{\text{CF}} = (1 - T)Y_i(1) + TY_i(0). \tag{4.11}$$

Consequently, τ_i can never be measured or computed from data, i.e., it is not identifiable (without further assumptions), as illustrated in Table 1.

Implicit in the form of (4.9) and (4.11) are the following two assumptions.

Assumption 4.13 (Stable unit treatment value; SUTVA). The observation on one unit should be unaffected by the particular assignment of treatments to the other units [23].

Assumption 4.14 (Consistency). If individual i receives treatment t, then the observed outcome is $Y_i = Y_i(t)$, i.e., the potential outcome for t.

Assumption 4.13 is usually understood as (i) units do not interfere, and (ii) there is only one treatment level per group (treated or control) leading to well-defined POs [61]. It can be violated, e.g., through (i) population dynamics such as herd immunity from vaccination or (ii) technical errors or varying within-group dosage, respectively. However, for many situations such as controlled studies it can be a reasonable assumption, and we can then view different units as independent samples from a population.

So far, we have considered POs for a given unit as deterministic quantities. However, most times it is impossible to fully characterize a unit, e.g., when dealing with complex subjects such as humans. Such lack of complete information introduces uncertainty, so that

POs are often instead treated as random variables. This parallels the combination of deterministic structural equations with exogenous noise variables in SCMs.[7] Indeed, there is a equivalence between POs and SCMs [104]:

$$Y_i(t) = Y \mid \mathrm{do}(T := t) \quad \text{in an SCM with } \mathbf{U} = \mathbf{u}_i\,,$$

An individual in the PO framework thus corresponds to a particular instantiation of the U_j in an SCM: the outcome is deterministic given $\mathbf{U}$, but since we do not observe $\mathbf{u}_i$ (nor can we characterize a given individual based on observed covariates), the counterfactual outcome is treated as a random variable. In practice, all we observe is a featurized description $\mathbf{x}_i$ of an individual i and have to reason about expected POs, $\mathbb{E}[Y(1), Y(0)|\mathbf{x}_i]$.

Another common assumption is that of no hidden confounders which we have already encountered in form of the causal Markov condition (Definition 4.2) for CGMs and causal sufficiency (Assumption 4.8) for SCMs. In the PO framework this becomes no hidden confounding between treatment and outcome and is referred to as (conditional) ignorability.

Assumption 4.15 (Conditional ignorability). Given a treatment $T \in \{0, 1\}$, potential outcomes $Y(0), Y(1)$, and observed covariates $\mathbf{X}$, we have

$$Y(0) \perp\!\!\!\perp T \mid \mathbf{X} \quad \text{and} \quad Y(1) \perp\!\!\!\perp T \mid \mathbf{X}. \tag{4.12}$$

The PO framework is tailored toward studying the (confounded) effect of a typically binary treatment variable on an outcome and is mostly used for causal reasoning, i.e., estimating individual and population level causal effects (Section 9). In this context, it is sometimes seen as an advantage that an explicit graphical representation is not needed. At the same time, the lack of a causal graph and the need for special treatment and outcome variables make POs rather unsuitable for causal discovery where other frameworks prevail.

5. INDEPENDENT CAUSAL MECHANISMS

We now return to the disentangled factorization (4.1) of the joint distribution $p(X_1, \ldots, X_n)$. This factorization according to the causal graph is always possible when the U_i are independent, but we will now consider an additional notion of independence relating the factors in (4.1) to one another.

Consider a dataset that consists of altitude A and average annual temperature T of weather stations [111]. Variables A and T are correlated, which we believe is due to the fact that the altitude has a causal effect on the temperature. Suppose we had two such datasets, one for Austria and one for Switzerland. The two joint distributions may be rather different, since the marginal distributions $p(A)$ over altitudes will differ. The conditionals $p(T|A)$, however, may be rather similar, since they follow from physical mechanisms generating temperature from altitude. The causal factorization $p(A)\,p(T|A)$ thus contains a component $p(T|A)$ that

7 When all noise variables in an SCM are fixed, the other variables are uniquely determined; without complete background knowledge, on the other hand, they are random.

generalizes across countries, while the entangled factorization $p(T)p(A|T)$ does not. Cum grano salis, the same applies when we consider interventions in a system. For a model to correctly predict the effect of interventions, it needs to be robust with respect to generalizing from an observational distribution to certain *interventional* distributions.

One can express the above insights as follows [111, 128]:

Principle 5.1 (Independent causal mechanisms (ICM)). *The causal generative process of system's variables is composed of autonomous modules that do not inform or influence each other. In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other mechanisms.*

This principle subsumes several notions important to causality, including separate intervenability of causal variables, modularity and autonomy of subsystems, and invariance [104, 111]. If we have only two variables, it reduces to an independence between the cause distribution and the mechanism producing the effect distribution from the cause distribution.

Applied to the causal factorization (4.1), the principle tells us that the factors should be independent in two senses:

(*influence*) changing (or intervening upon) one mechanism $p(X_i|\mathbf{PA}_i)$ does not change the other mechanisms $p(X_j|\mathbf{PA}_j)$ $(i \neq j)$, and

(*information*) knowing some mechanisms $p(X_i|\mathbf{PA}_i)$ $(i \neq j)$ does not give us information about a mechanism $p(X_j|\mathbf{PA}_j)$.

We view any real-world distribution as a product of causal mechanisms. A change in such a distribution (e.g., when moving from one setting/domain to a related one) will always be due to changes in at least one of those mechanisms. Consistent with Principle 5.1, we hypothesize [133]:

Principle 5.2 (Sparse mechanism shift (SMS)). *Small distribution changes tend to manifest themselves in a sparse or local way in the causal/disentangled factorization* (4.1)*, i.e., they should usually not affect all factors simultaneously.*

In contrast, if we consider a noncausal factorization, e.g., (4.2), then many terms will be affected simultaneously as we change one of the physical mechanisms responsible for a system's statistical dependences. Such a factorization may thus be called *entangled*. The notion of disentanglement has recently gained popularity in machine learning [9, 50, 83, 147], sometimes loosely identified with statistical independence. The notion of invariant, autonomous, and independent mechanisms has appeared in various guises throughout the history of causality research, see [1, 104, 111, 133].

Measures of dependence of mechanisms. Note that the dependence of two mechanisms $p(X_i|\mathbf{PA}_i)$ and $p(X_j|\mathbf{PA}_j)$ does not coincide with the statistical dependence of the random variables X_i and X_j. Indeed, in a causal graph, many of the random variables will be dependent even if all the mechanisms are independent.

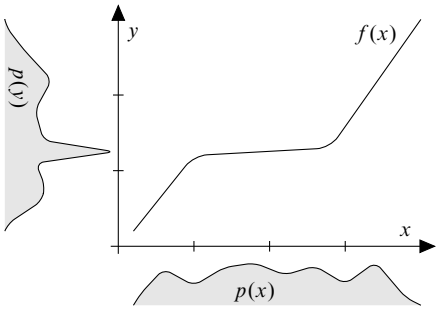

FIGURE 6
If $p(x)$ and f are chosen independently, then peaks of $p(y)$ tend to occur in regions where f has small slope. Hence $p(y)$ contains information about f^{-1} (figure from **[111]**).

Consider two variables and structural assignments $X := U$ and $Y := f(X)$, i.e., the cause X is a noise variable (with density $p(x)$), and the effect Y is a deterministic function of the cause. Let us, moreover, assume that the ranges of X and Y are both $[0, 1]$, and f is strictly monotonically increasing. The ICM principle then reduces to an independence of $p(x)$ and f. Let us consider $p(x)$ and the derivative f' as random variables on the probability space $[0, 1]$ with Lebesgue measure, and use their correlation as a measure of dependence of mechanisms. It can be shown that for $f \neq \mathrm{id}$, independence of $p(x)$ and f' implies dependence between $p(y)$ and $(f^{-1})'$ (see Figure 6). Other measures are possible and admit information-geometric interpretations. Intuitively, under the ICM assumption (Principle 5.1), the "irregularity" of the effect distribution becomes a *sum* of (i) irregularity already present in the input distribution and (ii) irregularity introduced by the mechanism f, i.e., the irregularities of the two mechanisms add up rather than (partly) compensating each other. This would not be the case in the opposite ("anticausal") direction (for details, see **[68]**). Other dependence measures have been proposed for high-dimensional linear settings and time series **[12, 63, 67, 136]**.

Algorithmic independence. So far, we have discussed links between causal and statistical structures. The fundamental of the two is the causal structure, since it captures the physical mechanisms that generate statistical dependences in the first place. The statistical structure is an epiphenomenon that follows if we make the unexplained variables random. It is awkward to talk about the (statistical) information contained in a mechanism, since deterministic functions in the generic case neither generate nor destroy information. This motivated us to devise an algorithmic model of causal structures in terms of Kolmogorov complexity **[65]**. The Kolmogorov complexity (or algorithmic information) of a bit string is essentially the length of its shortest compression on a Turing machine, and thus a measure of its information content. Independence of mechanisms can be defined as vanishing mutual algorithmic information: two conditionals are considered independent if knowing (the shortest compression of) one does not help achieve a shorter compression of the other one.

Algorithmic information theory provides a natural framework for nonstatistical graphical models. Just like statistical CGMs are obtained from SCMs by making the unexplained variables U_i random, we obtain algorithmic CGMs by making the U_i bit strings (jointly independent across nodes) and viewing the node X_i as the output of a fixed Turing machine running program U_i with input $\mathbf{PA}_i$. Similar to the statistical case, one can define a local causal Markov condition, a global one in terms of d-separation, and a decomposition of the joint Kolmogorov complexity in analogy to (4.1), and prove that they are implied by the SCM [65]. This approach shows that causality is not intrinsically bound to statistics: causality is about *mechanisms* governing flow of information which may or may not be statistical.

The assumption of algorithmically independent mechanisms has interesting implications for physics: it implies the second law of thermodynamics (i.e., the arrow of time). Consider a process where an incoming ordered beam of photons (the cause) is scattered by an object (the mechanism). Then the outgoing beam (the effect) contains information about the object. Microscopically, the time evolution is reversible; however, the photons contain information about the object only *after* the scattering. What underlies Loschmidt's paradox [86]?

The asymmetry can be explained by applying the ICM Principle 5.1 to initial state and system dynamics, postulating that the two be algorithmically independent, i.e., knowing one does not allow a shorter description of the other. The Kolmogorov complexity of the system's state can then be shown to be nondecreasing under time evolution [62]. If we view Kolmogorov complexity as a measure of entropy, this means that the entropy of the state can only stay constant or increase, amounting to the second law of thermodynamics.

Note that the resulting state after time evolution is clearly *not* independent of the system dynamic: it is precisely the state that, when fed to the inverse dynamics, would return us to the original (ordered) state.

6. LEVELS OF CAUSAL MODELING

Coupled differential equations are the canonical way of modeling physical phenomena. They allow us to predict the future behavior of a system, to reason about the effect of interventions, and—by suitable averaging procedures—to predict *statistical* dependences that are generated by a coupled time evolution. They also allow us to gain insight into a system, explain its functioning, and, in particular, read off its causal structure.

Consider a coupled set of ordinary differential equations

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d, \tag{6.1}$$

with initial value $\mathbf{x}(t_0) = \mathbf{x}_0$. We assume that they correctly describe the physical mechanisms of a system.[8] The Picard–Lindelöf theorem states that, at least locally, if f is Lipschitz, there

8 In other words, they do not merely phenomenologically describe its time evolution without capturing the underlying mechanisms (e.g., due to unobserved confounding, or a form of coarse-graining that does not preserve the causal structure [123, 133]).

TABLE 2

A simple taxonomy of models. The most detailed model (top) is a mechanistic or physical one, usually in terms of differential equations. At the other end of the spectrum (bottom), we have a purely statistical model; this can be learned from data and is useful for predictions but often provides little insight beyond modeling associations between epiphenomena. Causal models can be seen as descriptions that lie in between, abstracting away from physical realism while retaining the power to answer certain interventional or counterfactual questions.

Model	Predict in i.i.d. setting	Predict under distribution shift/intervention	Answer counterfactual questions	Obtain physical insight	Learn from data
Mechanistic/physical	yes	yes	yes	yes	?
Structural causal	yes	yes	yes	?	?
Causal graphical	yes	yes	no	?	?
Statistical	yes	no	no	no	yes

exists a unique solution $\mathbf{x}(t)$. This implies, in particular, that the immediate future of $\mathbf{x}$ is implied by its past values.

In terms of infinitesimal differentials dt and $d\mathbf{x} = \mathbf{x}(t + dt) - \mathbf{x}(t)$, (6.1) reads

$$\mathbf{x}(t + dt) = \mathbf{x}(t) + dt \cdot f\big(\mathbf{x}(t)\big). \tag{6.2}$$

From this, we can ascertain which entries of the vector $\mathbf{x}(t)$ cause the future of others $\mathbf{x}(t + dt)$, i.e., the causal structure.

Compared to a differential equation, a statistical model derived from the joint distribution of a set of (time-independent) random variables is a rather superficial description of a system. It exploits that some of the variables allow the prediction of others as long as the experimental conditions do not change. If we drive a differential equation system with certain types of noise, or if we average over time, statistical dependences between components of $\mathbf{x}$ may emerge, which can be exploited by machine learning. In contrast to the differential equation model, such a model does not allow us to predict the effect of interventions; however, its strength is that it can often be learned from data.

Causal modeling lies in between these two extremes. It aims to provide understanding and predict the effect of interventions. Causal discovery and learning tries to arrive at such models in a data-driven way, using only weak assumptions (see Table 2, from [111,133]).

While we may naively think that causality is always about time, most existing causal models do not (and need not) consider time. For instance, returning to our example of altitude and temperature, there is an underlying dynamical physical process that results in higher places tending to be colder. On the level of microscopic equations of motion for the involved particles, there is a temporal causal structure. However, when we talk about dependence or causality between altitude and temperature, we need not worry about the details of this temporal structure; we are given a dataset where time does not appear, and we can reason about how that dataset would look if we were to intervene on temperature or altitude.

Some work exists trying to build bridges between these different levels of description. One can derive SCMs that describe the interventional behavior of a coupled system that is in an equilibrium state and perturbed in an adiabatic way [96], with generalizations to oscillatory systems [122]. In this work, an SCM arises as a high-level abstraction of an underlying system of differential equations. It can only be derived if suitable high-level variables can be defined [123], which in practice may well be the exception rather than the rule.

7. CAUSAL DISCOVERY

Sometimes, domain knowledge or the temporal ordering of events can help constrain the causal relationships between variables, e.g., we may know that certain attributes like age or sex are not caused by others; treatments influence health outcomes; and events do not causally influence their past. When such domain knowledge is unavailable or incomplete, we need to perform *causal discovery*: infer which variables causally influence which others, i.e., learn the causal structure (e.g., a DAG) from data. Since experiments are often difficult and expensive to perform while observational (i.e., passively collected) data is abundant, causal discovery from observational data is of particular interest.

As discussed in Section 3 in the context of the Common Cause Principle 3.2, the case where we have two variables is already difficult since the same dependence can be explained by multiple different causal structures. One might thus wonder if the case of more observables is completely hopeless. Surprisingly, this is not the case: the problem becomes easier (in a certain sense) because there are nontrivial conditional independence properties [25, 35, 145] implied by a causal structure. We first review two classical approaches to the multivariate setting before returning to the two-variable case.

Constraint-based methods. Constraint-based approaches to causal discovery test which (conditional) independences can be inferred from the data and then try to find a graph which implies them. They are therefore also known as independence-based methods. Such a procedure requires a way of linking properties of the data distribution p to properties of the underlying causal graph G. This link is known as the faithfulness assumption.

Assumption 7.1 (Faithfulness). The only (conditional) independences satisfied by p are those implied by the causal Markov condition (Definition 4.2).

Faithfulness can be seen as the converse of the causal Markov condition. Together, they constitute a one-to-one correspondence between graphical separation in G and conditional independence in p. While the causal Markov condition is satisfied by construction, faithfulness is an assumption which may be violated. A classical example for a violation of faithfulness is when causal effects along different paths cancel.

Example 7.2 (Violation of faithfulness). Consider the SCM from Example 4.6 and let

$$X_1 := U_1, \quad X_2 := \alpha X_1 + U_2, \quad X_3 := \beta X_1 + \gamma X_2 + U_3$$

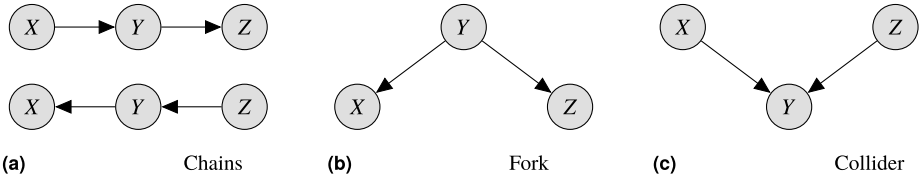

FIGURE 7

Illustration of Markov equivalence using common graph motifs. The chains in (a) and the fork in (b) all imply the relation $X \perp\!\!\!\perp Z \mid Y$ (and no others). They thus form a Markov equivalence class, meaning they cannot be distinguished using conditional independence testing alone. The collider, or v-structure, in (c) implies $X \perp\!\!\!\perp Z$ (but $X \not\perp\!\!\!\perp Z \mid Y$) and forms its own Markov equivalence class, so it can be uniquely identified from observational data. For this reason, v-structures are helpful for causal discovery. It can be shown that two graphs are Markov equivalent if and only if they share the same skeleton and v-structures.

with $U_1, U_2, U_3 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. By substitution, we obtain $X_3 = (\beta + \alpha\gamma)X_1 + \gamma U_2 + U_3$. Hence $X_3 \perp\!\!\!\perp X_1$ whenever $\beta + \alpha\gamma = 0$, even though this independence is not implied by the causal Markov condition over the induced causal graph G, see Figure 5. Here, faithfulness is violated if the direct effect of X_1 on X_3 (β) and the indirect effect via X_2 ($\alpha\gamma$) cancel.

Apart from relying on faithfulness, a fundamental limitation to constraint-based methods is the fact that many different DAGs may encode the same d-separation/independence relations. This is referred to as Markov equivalence and illustrated in Figure 7.

Definition 7.3 (Markov equivalence). Two DAGs are said to be Markov equivalent if they encode the same d-separation statements. The set of all DAGs encoding the same d-separations is called a Markov equivalence class.

Constraint-based algorithms typically first construct an undirected graph, or skeleton, which captures the (conditional) independences found by testing, and then direct as many edges as possible using Meek's orientation rules [90]. The first step carries most of the computational weight and various algorithms have been devised to solve it efficiently.

The simplest procedure is implemented in the IC [109] and SGS [144] algorithms. For each pair of variables (X, Y), these search through all subsets $\mathbf{W}$ of the remaining variables to check whether $X \perp\!\!\!\perp Y \mid \mathbf{W}$. If no such set $\mathbf{W}$ is found, then X and Y are connected with an edge. Since this can be slow due to the large number of subsets, the PC algorithm [144] uses a much more efficient search procedure. It starts from a complete graph and then sequentially test only subsets of the neighbors of X or Y of increasing size, removing an edge when a separating subset is found. This neighbor search is no longer guaranteed to give the right result for causally insufficient systems, i.e., in the presence of hidden confounders. The FCI (short for fast causal inference) algorithm [144] addresses this setting, and produces a partially directed causal graph as output.

Apart from being limited to recovering a Markov equivalence class, constraint-based methods can suffer from statistical issues. In practice, datasets are finite, and conditional

independence testing is a notoriously difficult problem, especially if conditioning sets are continuous and multidimensional. So while, in principle, the conditional independences implied by the causal Markov condition hold true irrespective of the complexity of the functions appearing in an SCM, for finite datasets conditional independence testing is hard without additional assumptions [135]. Recent progress in (conditional) independence testing heavily relies on kernel function classes to represent probability distributions in reproducing kernel Hilbert spaces, see Section 2.

Score-based methods. Score-based approaches to causal discovery assign a score to each graph G from a set of candidate graphs (usually the set of all DAGs). The score S is supposed to reflect how well G explains the observed data $\mathbf{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$, and we choose the graph $\hat{G}$ maximizing this score,

$$\hat{G} = \operatorname*{argmax}_G S(G|\mathbf{D}).$$

Various score functions have been proposed, but most methods assume a parametric model which factorizes according to G, parametrized by $\theta \in \Theta$. Two common choices are multinomial models for discrete data [22] and linear Gaussian models for continuous data [34]. For example, a penalized maximum likelihood approach using the BIC [134] as a score yields

$$S_{\mathrm{BIC}}(G|\mathbf{D}) = \log p(\mathbf{D}|G, \hat{\theta}^{\mathrm{MLE}}) - \frac{k}{2} \log m, \tag{7.1}$$

where k is the number of parameters and $\hat{\theta}^{\mathrm{MLE}}$ is the maximum likelihood estimate for θ to D in G. Note that k generally increases with the number of edges in G so that the second term in (7.1) penalizes complex graphs which do not lead to substantial improvements.

Another choice of score function is the marginal likelihood, or evidence, in a Bayesian approach to causal discovery, which requires specifying prior distributions over graphs and parameters, $p(G, \theta) = p(G)p(\theta|G)$. The score for G is then given by

$$S_{\mathrm{BAYES}}(G|\mathbf{D}) = p(\mathbf{D}|G) = \int_\Theta p(\mathbf{D}|G, \theta) p(\theta|G) \mathrm{d}\theta. \tag{7.2}$$

This integral is intractable in general, but can be computed exactly for some models such as a Dirichlet-multinomial under some mild additional assumptions [47, 48].

A major drawback of score-based approaches is the combinatorial size of the search space. The number of DAGs over n random variables grows superexponentially and can be computed recursively (to account for acyclicity constraints) [119]. For example, the number of DAGs for $n = 5$ and $n = 10$ nodes is 29281 and 4175098976430598143, respectively. Finding the best scoring DAG is NP-hard [20]. To overcome this problem, greedy search techniques can be applied, e.g., greedy equivalence search (GES) [21] which optimizes for the BIC.

In recent years, another class of methods has emerged that is based on assuming particular functional forms for the SCM assignments. Those arose from studying the cause-effect inference problem, as discussed below.

Cause–effect inference. In the case of only two variables, the ternary concept of conditional independences collapses and the causal Markov condition (Definition 4.2) thus has no nontrivial implications. However, we have seen in Section 5 that assuming an independence of mechanisms (Principle 5.1) lets us find asymmetries between cause and effect, and thus address the cause–effect inference problem previously considered unsolvable [68]. It turns out that this problem can be also addressed by making additional assumptions on function classes, as not only the graph topology leaves a footprint in the observational distribution, but so do the functions f_i in an SCM. Such assumptions are typical for machine learning, where it is well known that finite-sample generalization without assumptions on function classes is impossible, and where much attention is devoted to properties of function classes (e.g., priors or capacity measures), as discussed in Section 2.

Let us provide an intuition as to why assumptions on the functions in an SCM should help learn about them from data. Consider a toy SCM with only two observables $X \to Y$. In this case, the structural equations (4.5) turn into

$$X := U, \quad Y := f(X, V) \tag{7.3}$$

with noises $U \perp\!\!\!\perp V$. Now think of V acting as a random selector variable choosing from among a set of functions $\mathcal{F} = \{f_v(x) \equiv f(x, v) \mid v \in \mathrm{supp}(V)\}$. If $f(x, v)$ depends on v in a nonsmooth way, it should be hard to glean information about the SCM from a finite dataset, given that V is not observed and it randomly switches between arbitrarily different f_v.[9] This motivates restricting the complexity with which f depends on V. A natural restriction is to assume an *additive noise model*

$$X := U, \quad Y := f(X) + V. \tag{7.4}$$

If f in (7.3) depends smoothly on V, and if V is relatively well concentrated, this can be motivated by a local Taylor expansion argument. Such assumptions drastically reduce the effective size of the function class—without them, the latter could depend exponentially on the cardinality of the support of V.

Restrictions of function classes can break the symmetry between cause and effect in the two-variable case: one can show that given a distribution over X, Y generated by an additive noise model, one cannot fit an additive noise model in the opposite direction (i.e., with the roles of X and Y interchanged) [6,53,76,95,113]. This is subject to certain genericity assumptions, and notable exceptions include the case where U, V are Gaussian and f is linear. It generalizes results of [139] for linear functions, and it can be generalized to include nonlinear rescaling [164], cycles [94], confounders [64], and multivariable causal discovery [112]. There is now a range of methods that can detect causal direction better than chance [97].

9 Suppose X and Y are binary, and U, V are uniform Bernoulli variables, the latter selecting from $\mathcal{F} = \{\mathrm{id}, \mathrm{not}\}$ (i.e., identity and negation). In this case, the entailed distribution for Y is uniform, *independent* of X, even though we have $X \to Y$. We would be unable to discern $X \to Y$ from data. (This would also constitute a violation of faithfulness (Assumption 7.1).)

We have thus gathered some evidence that ideas from machine learning can help tackle causality problems that were previously considered hard. Equally intriguing, however, is the opposite direction: can causality help us improve machine learning?

Nonstationarity-based methods. The last family of causal discovery approaches we mention is based on ideas of nonstationarity and invariance [128]. These approaches do not apply to purely observational data collected in an i.i.d. setting. In contrast, they aim to leverage heterogeneity of data collected from different environments. The main idea is the following: since causal systems are modular in the sense of the ICM Principle 5.1, changing one of the independent mechanisms should leave the other components, or causal conditionals, unaffected (SMS Principle 5.2). A correct factorization of the joint distribution according to the underlying causal structure should thus be able to explain heterogeneity by localized changes in one (or few) of the mechanisms while the others remain invariant.

One of the first works to use this idea [150] analyzed which causal structures can be distinguished given data resulting from a set of mechanism changes. Recent work [55] additionally aims to learn a low-dimensional representation of the mechanism changes. Other works [110,120] have proposed methods for finding the direct causes of a given target variable. Using a recent result on identifiability of nonlinear ICA [59] which also relies on nonstationarity, a method for learning general nonlinear SCMs was proposed [93]. Here the idea is to train a classifier to discriminate between the true value of some nonstationarity variable (such as a time-stamp or environment indicator) and a shuffled version thereof.

8. IMPLICATIONS FOR MACHINE LEARNING

Semisupervised learning. Suppose our underlying causal graph is $X \to Y$, and we are trying to learn a mapping $X \to Y$. The causal factorization (4.1) for this case is

$$p(X, Y) = p(X)p(Y|X). \tag{8.1}$$

The ICM Principle 5.1 posits that the modules in a joint distribution's causal factorization do not inform or influence each other. This means that, in particular, $p(X)$ should contain no information about $p(Y|X)$, which implies that semisupervised learning [17] should be futile, as it is trying to use additional information about $p(X)$ (from unlabeled data) to improve our estimate of $p(Y|X = x)$. What about the opposite direction, is there hope that semisupervised learning should be possible in that case? It turns out the answer is yes, due to the work on cause–effect inference using the ICM Principle 5.1 [24]. It introduced a measure of dependence between the input and the conditional of output given input, and showed that if this dependence is zero in the causal direction, then it is strictly positive in the opposite direction. Independence of cause and mechanism in the causal direction thus implies that in the backward direction (i.e., for anticausal learning), the distribution of the input variable should contain information about the conditional of output given input, i.e., the quantity that machine learning is usually concerned with. This is exactly the kind of information that semisupervised learning requires when trying to improve the estimate of output given input

by using unlabeled inputs. This suggests that *semisupervised learning should be impossible for causal learning problems, but feasible otherwise*, in particular for anticausal ones. A metaanalysis of published semisupervised learning benchmark studies corroborated this prediction [128], and similar results apply for natural language processing [69]. These findings are intriguing since they provide insight into *physical* properties of learning problems, thus going beyond the methods and applications that machine learning studies usually provide.

Subsequent developments include further theoretical analyses [66,111] and a form of conditional semisupervised learning [156]. The view of semisupervised learning as exploiting dependences between a marginal $p(x)$ and a noncausal conditional $p(y|x)$ is consistent with the common assumptions employed to justify semisupervised learning [17,125].

Invariance and robustness. We have discussed the shortcomings of the i.i.d. assumption, which rarely holds true exactly in practice, and the fact that real-world intelligent agents need to be able to generalize not just within a single i.i.d. setting, but across related problems. This notion has been termed *out-of-distribution (o.o.d.) generalization*, attracting significant attention in recent years [133]. While most work so far has been empirical, statistical bounds would be desirable that generalize (2.4), including additional quantities measuring the distance between training and test distribution, incorporating meaningful assumptions [137]. Such assumptions are necessary [8], and could be causal, or related to invariance properties.

The recent phenomenon of "adversarial vulnerability" [148] shows that minuscule targeted violations of the i.i.d. assumption, generated by adding suitably chosen noise to images (imperceptible to humans), can lead to dangerous errors such as confusion of traffic signs. These examples are compelling as they showcase nonrobustnesses of artificial systems which are not shared by human perception. Our own perception thus exhibits invariance or robustness properties that are not easily learned from a single training set.

Early causal work related to domain shift [128] looked at the problem of learning from multiple cause–effect datasets that share a functional mechanism but differ in noise distributions. More generally, given (data from) multiple distributions, one can try to identify components which are robust, and find means to transfer them across problems [4, 36, 54, 163, 166]. According to the ICM Principle 5.1, invariance of conditionals or functions (also referred to as covariate shift in simple settings) should only hold in the causal direction, a reversal of the impossibility described for SSL.

Building on the work of [110,128], the idea of invariance for prediction has also been used for supervised learning [3, 87, 120]. In particular, "invariant risk minimization" (IRM) was proposed as an alternative to ERM, cf. (2.3).

9. CAUSAL REASONING

In contrast to causal discovery (Section 7), which aims to uncover the causal structure underlying a set of variables, *causal reasoning* starts from a known (or postulated) causal graph and answers causal queries of interest. While causal discovery often looks for qualitative relationships, causal reasoning usually aims to quantify them. This requires two steps:

(i) *identifying* the query, i.e., deriving an estimand for it that only involves observed quantities; and (ii) *estimating* this using data. Often, the quantities of interest can be described as treatment effects, i.e., contrasts between two interventions.

Definition 9.1 (Treatment effects). The conditional average treatment effect (CATE), conditioned on (a subset of) features $\mathbf{x}$, is defined as

$$\tau(\mathbf{x}) := \mathbb{E}\big[Y|\mathbf{x}, \mathrm{do}(T=1)\big] - \mathbb{E}\big[Y|\mathbf{x}, \mathrm{do}(T=0)\big] = \mathbb{E}\big[Y(1) - Y(0)|\mathbf{x}\big]. \tag{9.1}$$

The average treatment effect (ATE) is defined as the population average of the CATE,

$$\tau := \mathbb{E}\big[\tau(\mathbf{X})\big] = \mathbb{E}\big[Y \,|\, \mathrm{do}(T=1)\big] - \mathbb{E}\big[Y \,|\, \mathrm{do}(T=0)\big] = \mathbb{E}\big[Y(1) - Y(0)\big]. \tag{9.2}$$

While ITE (Definition 4.12) and CATE (9.1) are sometimes used interchangeably, there is a conceptual difference: ITE refers to the difference of two POs and is thus bound to an individual, while CATE applies to subpopulations, e.g., the CATE for females in their 40s. Since the ITE is fundamentally impossible to observe, it is often estimated by the CATE conditional on an individual's features $\mathbf{x}_i$ using suitable additional assumptions.

As is clear from Definition 9.1, the treatment effects we want to estimate involve interventional expressions. However, we usually only have access to observational data. Causal reasoning can thus be cast as answering interventional queries using observational data and a causal model. This involves dealing with confounders, both observed and unobserved.

Before discussing how to identify and estimate causal effects, we illustrate why causal assumptions are necessary using a well-known statistical phenomenon.

Simpson's paradox and Covid-19. *Simpson's paradox* refers to the observation that aggregating data across subpopulations may yield opposite trends (and thus lead to reversed conclusions) from considering subpopulations separately [142]. We observed a textbook example of this during the Covid-19 pandemic by comparing case fatality rates (CFRs), i.e., the proportion of confirmed Covid-19 cases which end in fatality, across different countries and age groups as illustrated in Figure 8 [154]: for *all* age groups, CFRs in Italy are *lower* than in China, but the *total* CFR in Italy is *higher*.

How can such a pattern be explained? The case demographic (see Figure 8, right) is rather different across the two countries, i.e., there is a statistical association between country and age. In particular, Italy recorded a much larger proportion of cases in older patients who are generally at higher risk of dying from Covid-19 (see Figure 8, left). While this provides a consistent explanation in a *statistical* sense, the phenomenon may still seem puzzling as it defies our *causal* intuition. Humans appear to naturally extrapolate conditional probabilities to read them as causal effects, which can lead to inconsistent conclusions and may leave one wondering: *how can the disease in Italy be less fatal for the young, less fatal for the old, but more fatal for the people overall?* It is for this reason that the reversal of (conditional) probabilities in Figure 8 is perceived as and referred to as a "paradox" [49, 105].

If we consider the country as treatment whose causal effect on fatality is of interest, then causal assumptions (e.g., in the form of a causal graph) are needed to decide how to

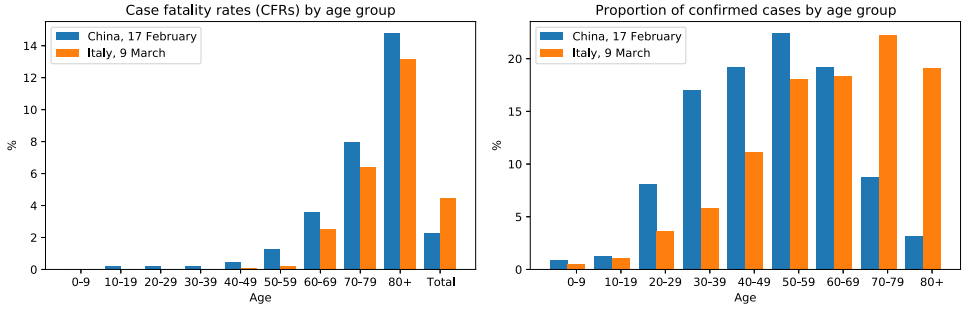

FIGURE 8

(Left) Covid-19 case fatality rates (CFRs) in Italy and China by age and in aggregate ("Total"), including all confirmed cases and fatalities up to the time of reporting in early 2020 (see legend): for all age groups, CFRs in Italy are lower than in China, but the total CFR in Italy is higher, an example of *Simpson's paradox*. (Right) The case demographic differs between countries: in Italy, most cases occurred in the older population (figure from [154]).

handle covariates such as age that are statistically associated with the treatment, e.g., whether to stratify by (i.e., adjust for) age or not. This also explains why randomized controlled trials (RCTs) [31] are the gold standard for causal reasoning: randomizing the assignment breaks any potential links between the treatment variable and other covariates, thus eliminating potential problems of bias. However, RCTs are costly and sometimes unethical to perform, so that causal reasoning often relies on observational data only.[10]

We first consider the simplest setting *without hidden confounders and with overlap*. We start with *identification* of treatment effects on the population level, and then discuss different techniques for *estimating* these from data.

Identification. In absence of unmeasured variables (i.e., without hidden confounding), and provided we know the causal graph, it is straight-forward to compute causal effects by adjusting for covariates. A principled approach to do so for any given graph was proposed by Robins [117] and is known as the *g-computation formula* (where the g stands for general). It is also known as *truncated factorisation* [104] or *manipulation theorem* [144]. It relies on the independence of causal mechanisms (Principle 5.1), i.e., the fact that intervening on a variable leaves the other causal conditionals in (4.1) unaffected:

$$p\big(X_1, \ldots, X_n | \operatorname{do}(X_i := x_i)\big) = \delta(X_i = x_i) \prod_{j \neq i} p(X_j | \mathbf{PA}_j). \tag{9.3}$$

From (9.3) the interventional distribution of interest can then be obtained by marginalization. This is related to the idea of graph surgery (see Figure 5), and leads to a set of three inference rules for manipulating interventional distributions known as *do-calculus* [104] that have been shown to be complete for identifying causal effects [56, 140].

10 For a treatment of more general types of data fusion and transportability of experimental findings across different populations, we refer to [5, 106].

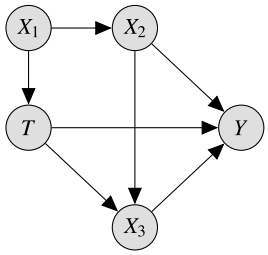

FIGURE 9

Treatment effect estimation with three observed covariates X_1, X_2, X_3: here, the valid adjustment sets for $T \to Y$ (see Proposition 9.3) are $\{X_1\}$, $\{X_2\}$, and $\{X_1, X_2\}$. Including X_3 opens the *nondirected path* $T \to X_3 \leftarrow X_2 \to Y$ and lies on the directed path $T \to X_3 \to Y$, both of which can introduce bias.

Note that covariate adjustment may still be needed, even if there are no clear confounders directly influencing both treatment and outcome, as shown by the example in Figure 9.

Example 9.2. Applying the g-computation formula (9.3) to the setting of Figure 9, we obtain

$$p\big(y|\,\mathrm{do}(t)\big) = \sum_{x_1} p(x_1) \sum_{x_2} p(x_2|x_1) \sum_{x_3} p(x_3|t,x_2)\, p(y|t,x_2,x_3) \tag{9.4}$$

$$= \sum_{x_1} p(x_1) \sum_{x_2} p(x_2|x_1)\, p(y|t,x_2) = \sum_{x_2} p(x_2)\, p(y|t,x_2) \tag{9.5}$$

$$\overset{\text{(a)}}{=} \sum_{x_1,x_2} p(x_1,x_2)\, p(y|t,x_1,x_2) \overset{\text{(b)}}{=} \sum_{x_1} p(x_1)\, p(y|t,x_1), \tag{9.6}$$

where the last line follows by using the following conditional independences implied by the graph: (a) $Y \perp\!\!\!\perp X_1 \mid \{T, X_2\}$, and (b) $X_2 \perp\!\!\!\perp T \mid X_1$.

Note that both the right-hand side in (9.5) and both sides in (9.6) take the form

$$p\big(y|\,\mathrm{do}(t)\big) = \sum_{\mathbf{z}} p(\mathbf{z})\, p(y|t,\mathbf{z}). \tag{9.7}$$

In this case we call $\mathbf{Z}$ a *valid adjustment set* for the effect of T on Y. Here, $\{X_1\}$, $\{X_2\}$, and $\{X_1, X_2\}$ are all valid adjustment sets, but it can be shown that, e.g., $\{X_1, X_3\}$ is not (see Figure 9). As computing the g-formula with many covariates can be cumbersome, graphical criteria for which subsets constitute valid adjustment sets are useful in practice, even in the absence of unobserved confounders.

Proposition 9.3 ([141]). *Under causal sufficiency, a set* $\mathbf{Z}$ *is a valid adjustment set for the causal effect of a singleton treatment* T *on an outcome* Y *(in the sense of* (9.7)*) if and only if the following two conditions hold: (i)* $\mathbf{Z}$ *contains no descendant of any node on a directed path from* T *to* Y *(except for descendants of* T *which are not on a directed path from* T *to* Y*); and (ii)* $\mathbf{Z}$ *blocks all non-directed paths from* T *to* Y*.*

Here, a path is called *directed* if all directed edges on it point in the same direction, and *nondirected* otherwise. A path is *blocked* (by a set of vertices $\mathbf{Z}$) if it contains a triple of

consecutive nodes connected in one of the following three ways: $A \to B \to C$ with $B \in \mathbf{Z}$, $A \leftarrow B \to C$ with $B \in \mathbf{Z}$, or $A \to B \leftarrow C$, where neither B nor any descendant of B is in $\mathbf{Z}$.

Two well-known types of adjustment set implied by Proposition 9.3 are *parent adjustment*, where $\mathbf{Z} = \mathbf{Pa}_T$; and the *backdoor criterion*, where $\mathbf{Z}$ is constrained to contain no descendants of T and to block all "back-door paths" from T to Y ($T \leftarrow \ldots Y$).

Note that Proposition 9.3 only holds singleton treatments (i.e., interventions on a single variable). For treatments $\mathbf{T}$ involving multiple variables, a slightly more complicated version of Proposition 9.3 can be given in terms of proper causal paths, and we refer to [102, 108] for details.

Let us briefly return to our earlier example of Simpson's paradox and Covid-19. Considering a plausible causal graph for this setting [154], we find that age A acts as a *mediator* $C \to A \to F$ of the causal effect of country C on fatality F (there is likely also a direct effect $C \to F$, potentially mediated by other, unobserved variables). If we are interested in the (total) causal effect of C on F (i.e., the overall influence of country on fatality), A should not be included for adjustment according to Proposition 9.3, and, subject to causal sufficiency, the total CFRs can be interpreted causally.[11] For another classic example of Simpson's paradox in the context of kidney stone treatment [18], on the other hand, the size of the stone acts as a *confounder* and thus needs to be adjusted for to obtain sound causal conclusions.

Valid covariate adjustment and the g-formula tell us how to compute interventions from the observational distribution when there are no hidden confounders. To actually identify causal effects from data, however, we need to also be able to *estimate* the involved quantities in (9.7). This is a problem if a subgroup of the population never (or always) receives a certain treatment. We thus need the additional assumption of a nonzero probability of receiving each possible treatment, referred to as *overlap*, or common support.

Assumption 9.4 (Overlap/common treatment support). For any treatment t and any configuration of features $\mathbf{x}$, it holds that $0 < p(T = t|\mathbf{X} = \mathbf{x}) < 1$.

The combination of overlap and ignorability (that is, no hidden confounders—see Assumption 4.15) is also referred to as *strong ignorability* and is a sufficient condition for identifying ATE and CATE: the absence of hidden confounders guarantees the existence of a valid adjustment set $\mathbf{Z} \subseteq \mathbf{X}$ for which $p(Y \,|\, \mathrm{do}(T = t), \mathbf{Z}) = p(Y|T = t, \mathbf{Z})$, and overlap guarantees that we can actually estimate the latter term for any $\mathbf{z}$ occurring with nonzero probability.[12]

Regression adjustment. Having identified a valid adjustment set (using Proposition 9.3), *regression adjustment* works by fitting a regression function $\hat{f}$ to $\mathbb{E}[Y|\mathbf{Z} = \mathbf{z}, T = t] = f(\mathbf{z}, t)$ using an observational sample $\{(y_i, t_i, \mathbf{z}_i)\}_{i=1}^m$. We can then use $\hat{f}$ to impute counterfactual outcomes as $\hat{y}_i^{\mathrm{CF}} = \hat{f}(\mathbf{z}_i, 1 - t_i)$ in order to estimate the CATE. The ATE is then

11 Mediation analysis [103] provides tools to tease apart and quantify the direct and indirect effects; the age-specific CFRs in Figure 8 then correspond to *controlled direct effects* [154].

12 The overlap assumption can thus be relaxed to hold for at least one valid adjustment set.

given by the population average and can be estimated as

$$\hat{\tau}_{\text{regression-adj.}} = \frac{1}{m_1} \sum_{i:t_i=1} \big(y_i - \hat{f}(\mathbf{z}_i, 0)\big) + \frac{1}{m_0} \sum_{i:t_i=0} \big(\hat{f}(\mathbf{z}_i, 1) - y_i\big), \tag{9.8}$$

where m_1 and m_0 are the number of observations from the treatment and control groups, respectively. Note the difference to the RCT estimator where no adjustment is necessary,

$$\hat{\tau}_{\text{RCT}} = \frac{1}{m_1} \sum_{i:t_i=1} y_i - \frac{1}{m_0} \sum_{i:t_i=0} y_i. \tag{9.9}$$

Matching and weighting approaches. While regression adjustment indirectly estimates ATE via CATE, matching and weighting approaches aim to estimate ATE directly. The general is idea to emulate the conditions of an RCT as well as possible.

Matching approaches work by splitting the population into subgroups based on feature similarity. This can be done on an individual level (so-called one-to-one or nearest neighbor matching) by matching each individual i with the most similar one, $j(i)$, from the opposite treatment group (i.e., $t_i \neq t_{j(i)}$). The difference of their outcomes, $y_i - y_{j(i)}$, is then considered as a sample of the ATE, and their average taken as an estimate thereof,

$$\hat{\tau}_{\text{NN-matching}} = \frac{1}{m_1} \sum_{i:t_i=1} (y_i - y_{j(i)}) + \frac{1}{m_0} \sum_{i:t_i=0} (y_{j(i)} - y_i). \tag{9.10}$$

Alternatively, the population can be split into larger subgroups with similar features (so-called strata). Each stratum is then treated as an independent RCT. If there are K strata containing $m_1, \ldots, m_K$ observations each, the stratified ATE estimator is

$$\hat{\tau}_{\text{stratified}} = \frac{\sum_{k=1}^{K} m_k \hat{\tau}_{\text{RCT}}^{(k)}}{\sum_{k=1}^{K} m_k}, \tag{9.11}$$

where $\hat{\tau}_{\text{RCT}}^{(k)}$ is the estimator from (9.9) applied to observation in the kth stratum.

Weighting approaches, on the other hand, aim to counteract the confounding bias by reweighting each observation to make the population more representative of an RCT. This means that underrepresented treatment groups are upweighted and overrepresented ones downweighted. An example is the inverse probability weighting (IPW) estimator,

$$\hat{\tau}_{\text{IPW}} = \frac{1}{m_1} \sum_{i:t_i=1} \frac{y_i}{p(T=1|\mathbf{Z}=\mathbf{z}_i)} - \frac{1}{m_0} \sum_{i:t_i=0} \frac{y_i}{p(T=0|\mathbf{Z}=\mathbf{z}_i)}. \tag{9.12}$$

The treatment probability $p(T = 1|\mathbf{Z})$ is also known as *propensity score*. While from a theoretical point of view $\mathbf{Z}$ should be a valid adjustment set, practitioners sometimes use all covariates to construct a propensity score.

Propensity score methods. To overcome the curse of dimensionality and gain statistical efficiency in high-dimensional, low-data regimes, propensity scores can be a useful tool, because covariates and treatment are rendered conditionally independent, $T \perp\!\!\!\perp \mathbf{Z} \mid s(\mathbf{z})$, by the propensity score $s(\mathbf{z}) := p(T = 1|\mathbf{Z} = \mathbf{z})$ [121]. Instead of adjusting for large feature sets or performing matching in high-dimensional spaces, the scalar propensity score can be used instead. Applying this idea to the above methods gives rise to *propensity score adjustment*

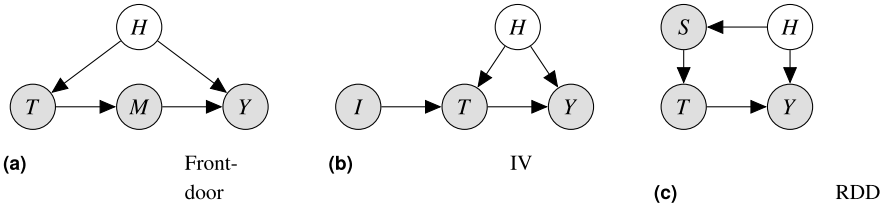

FIGURE 10

Overview of special settings which allow estimating causal effects of treatment T on outcome Y when the strong ignorability assumption (no hidden confounding or overlap) does not hold. In (a) the hidden confounder H is dealt with by means of an observed mediator M, while (b) relies on an instrumental variable (IV) which is independent of H. (c) In a regression discontinuity design (RDD), treatment assignment is a threshold function of some observed decision score S so that there is no overlap between treatment groups.

and *propensity score matching*. For the latter, the difference in propensity scores is used as similarity between instances to find nearest neighbors or to define strata.

While simplifying in one respect, the propensity score needs to be estimated from data which is an additional source of error. The standard approach for this is to estimate $s(\mathbf{z})$ by logistic regression, but more sophisticated methods are also possible. However, propensity score methods still rely on having identified a valid adjustment set $\mathbf{Z}$ to give unbiased results. Using all covariates to estimate s, without checking for validity as an adjustment set, can thus lead to wrong results.

Next, we consider the case of causal reasoning with *unobserved confounders*. While it is not possible to identify causal effects in the general case, we will discuss two particular situations in which ATE can still be estimated. These are shown in Figures 10a and 10b.

Front-door adjustment. The first situation in which identification is possible even though a hidden variable H confounds the effect between treatment and outcome is known as *front-door adjustment*. The corresponding causal graph is shown in Figure 10a. Front-door adjustment relies on the existence of an observed variable M which blocks all directed paths from T to Y, so that T only causally influences Y through M. For this reason M is also called a *mediator*. The other important assumption is that the hidden confounder does not influence the mediator other than through the treatment T, i.e., $M \perp\!\!\!\perp H \mid T$. In this case, and provided $p(t, m) > 0$ for all t and m, the causal effect of T on Y is identifiable and is given by the following.

Proposition 9.5 (Front-door adjustment). *For the causal graph in Figure 10a it holds that*

$$p\big(y|\,\mathrm{do}(t)\big) = \sum_m p(m|t) \sum_{t'} p(t')\,p(y|m,t'). \tag{9.13}$$

We give a sketch of the derivation, and refer to [104] for a proof using the rules of do-calculus. Since M mediates the causal effect of T on Y, we have that

$$p\big(y|\,\mathrm{do}(t)\big) = \sum_m p\big(m|\,\mathrm{do}(t)\big)\,p\big(y|\,\mathrm{do}(m)\big). \tag{9.14}$$

Since there are no back-door paths from T to M, we have $p(m|\,\mathrm{do}(t)) = p(m|t)$.

Moreover, $\{T\}$ is a valid adjustment set for the effect of M on Y by Proposition 9.3, so

$$p\big(y|\,\mathrm{do}(m)\big) = \sum_{t'} p(t')\,p(y|m,t'). \tag{9.15}$$

Substituting into (9.14) then yields expression (9.13).

We point out that the setting presented here is only the simplest form of front-door adjustment which is sufficient to convey the main idea. It can be amended to include observed covariates $\mathbf{X}$ as well, as long as the conditions on the mediator remain satisfied.

Instrumental variables (IVs). The second setting for causal reasoning with hidden confounders is based on the idea of instrumental variables [2, 29, 160], see Figure 10b. The IV approach relies on the existence of a special observed variable I, called instrument.

Definition 9.6 (IV). A variable I is a valid instrument for estimating the effect of treatment T on outcome Y confounded by a hidden variable H if all of the following three conditions hold: (i) $I \perp\!\!\!\perp H$; (ii) $I \not\perp\!\!\!\perp T$; and (iii) $I \perp\!\!\!\perp Y \mid T$.

Condition (i) states that the instrument is independent of any hidden confounders H. Since this assumption cannot be tested, background knowledge is necessary to justify the use of a variable as IV in practice. Conditions (ii) and (iii) state that the instrument is correlated with treatment, and only affects the outcome through T, and are referred to as relevance and exclusion restriction, respectively.

Given a valid IV, we apply a two-stage procedure: first obtain an estimate $\hat{T}$ of the treatment variable T that is independent of H by predicting T from I. Having thus created an unconfounded version of the treatment, a regression of Y on $\hat{T}$ then reveals the correct causal effect. We demonstrate this idea for a simple linear model with continuous treatment variable where the causal effect can be obtained by two-stage least squares (2SLS).

Example 9.7 (Linear IV with 2SLS). Consider the linear SCM defined by

$$T := aI + bH + U_T, \quad Y := cH + dT + U_Y,$$

with U_T, U_Y independent noise terms. Then, since $I \perp\!\!\!\perp H$, linear regression of T on I recovers the coefficient a via $\hat{T} = aI$. Substituting for T in the structural equation for Y gives

$$Y := daI + (c + bd)H + U_Y + dU_T.$$

A second linear regression of Y on $\hat{T} = aI$ recovers the causal effect d because $(I \perp\!\!\!\perp H) \implies (\hat{T} \perp\!\!\!\perp H)$, whereas a naive regression of Y on T would give a different result, as $T \not\perp\!\!\!\perp H$.

Instrumental variables have been studied extensively and more sophisticated versions than the simple example above exist, allowing for nonlinear interactions and observed covariates.

Having discussed some special settings to deal with hidden confounding, we briefly present a technique to deal with violations of the overlap assumption.

Regression discontinuity design. In a *regression discontinuity design* (RDD), the treatment assignment mechanism behaves like a threshold function, i.e., the propensity score is discontinuous [60]. In the simplest setting, the assignment of treatment or control is determined by whether an *observed score* S is above a threshold s_0, $T := \mathbb{I}\{S \geq s_0\}$. This score in turn depends on other covariates which may or may not be observed. For example, patients may be assigned a risk score, and treatment is only prescribed if this score surpasses a given threshold. Since the score may be assigned by another institution, not all relevant covariates H are usually observed. However, it is assumed that the treatment decision only depends on the score, e.g., because doctors comply with the official rules. The causal graph for such a simple RDD setting is shown in Figure 10c. While the score S constitutes a valid adjustment set in principle, the problem with RDDs is the lack of overlap: patients with low scores are always assigned $T = 0$ and patients with high scores are always assigned $T = 1$. Because of this, covariate adjustment, matching, or weighting approaches do not apply. The general idea of an RDD is to overcome this challenge by comparing observations with score in a small neighborhood of the decision cut-off value s_0, motivated by the consideration that patients with close scores but on opposite sides of s_0 differ only in whether they received the treatment or not. For example, if the treatment cut-off value is 0.5 for a score in [0,1], then patients with scores of 0.49 and 0.51 are comparable and can be treated as samples from an RCT. An RDD (in its simplest form) thus focuses on differences in the regression function $\mathbb{E}[Y|S = s, T = t(s)] = f(s)$ for $s \in [s_0 - \varepsilon, s_0 + \varepsilon]$, where $\varepsilon > 0$ is small.

Half-sibling regression and exoplanet detection. We conclude this section with a real-world application performing causal reasoning in a confounded additive noise model. Launched in 2009, NASA's Kepler space telescope initially observed 150000 stars over four years, in search of exoplanet transits. These are events where a planet partially occludes its host star, causing a slight decrease in brightness, often orders of magnitude smaller than the influence of telescope errors. When looking at stellar light curves, we noticed that the noise structure was often shared across stars that were light years apart. Since that made direct interaction of the stars impossible, it was clear that the shared information was due to the telescope acting as a confounder. We thus devised a method that (a) regresses a given star of interest on a large set of other stars chosen such that their measurements contain no information about the star's astrophysical signal, and (b) removes that regression in order to cancel the telescope's influence.[13] The method is called "half-sibling" regression since target and predictors share a parent, namely the telescope. The method recovers the random variable representing the astrophysical signal almost surely (up to a constant offset), for an additive noise model (specifically, the observed light curve is a sum of the unknown astrophysical signal and an unknown function of the telescope noise), subject to the assumption that the telescope's effect on the star is in principle predictable from the other stars [132].

13 For events that are localized in time (such as exoplanet transits), we further argued that the same applies for suitably chosen past and future values of the star itself, which can thus also be used as predictors.

In 2013, the Kepler spacecraft suffered a technical failure, which left it with only two functioning reaction wheels, insufficient for the precise spatial orientation required by the original Kepler mission. NASA decided to use the remaining fuel to make further observations, however, the systematic error was significantly larger than before—a godsend for our method designed to remove exactly these errors. We augmented it with models of exoplanet transits and an efficient way to search light curves, leading to the discovery of 36 planet candidates [32], of which 21 were subsequently validated as bona fide exoplanets [92]. Four years later, astronomers found traces of water in the atmosphere of the exoplanet K2-18b—the first such discovery for an exoplanet in the habitable zone, i.e., allowing for liquid water [10, 151]. The planet turned out to be one that had been first detected in our work [32] (exoplanet candidate EPIC 201912552).

10. CURRENT RESEARCH AND OPEN PROBLEMS

Conservation of information. We have previously argued that the mechanization of information processing currently plays a similar role to the mechanization of energy processing in earlier industrial revolutions [125]. Our present understanding of information is rather incomplete, as was the understanding of energy during the course of the first two industrial revolutions. The profound modern understanding of energy came with Emmy Noether and the insight that energy conservation is due to a symmetry (or covariance) of the fundamental laws of physics: they look the same no matter how we shift time. One might argue that information, suitably conceptualized, should also be a conserved quantity, and that this might also be a consequence of symmetries. The notions of invariance/independence discussed above may be able to play a role in this respect.

Mass seemingly played two fundamentally different roles (inertia and gravitation) until Einstein furnished a deeper connection in general relativity. It is noteworthy that causality introduces a layer of complexity underlying the symmetric notion of statistical mutual information. Discussing source coding and channel coding, Shannon [138] remarked: *This duality can be pursued further and is related to a duality between past and future and the notions of control and knowledge. Thus we may have knowledge of the past but cannot control it; we may control the future but have no knowledge of it.*

What is an object? Following the i.i.d. pattern recognition paradigm, machine learning learns objects by extracting patterns from many observations. A complementary view may consider objects as modules that can be separately manipulated or intervened upon [149]. The idea that objects are defined by their behavior under transformation has been influential in fields ranging from psychology to mathematics [74, 88].

Causal representation learning. In hindsight, it appears somewhat naive that first attempts to build AI tried to realize intelligence by programs written by humans, since existing examples of intelligent systems appear much too complex for that. However, there is a second problem, which is just as significant: classic AI assumed that the symbols which were the basis of algorithms were provided a priori by humans. When building a chess program, it is

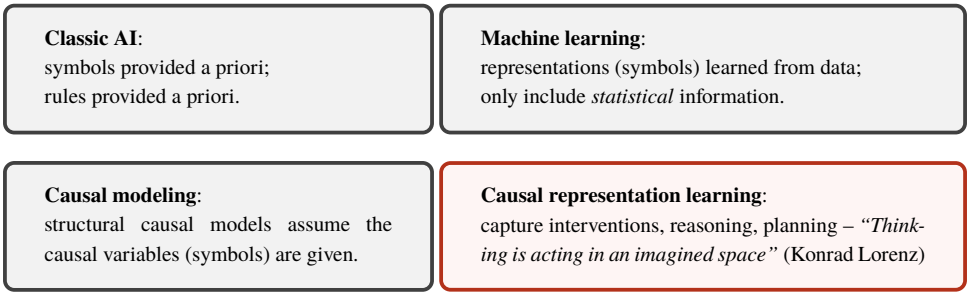

FIGURE 11
Causal representation learning aims to automatically learn representations that contain not just statistical information, but support interventions, reasoning, and planning. The long-term goal of this field is to learn causal world models supporting AI, or causal digital twins of complex systems.

clear that the algorithms operate on chess board positions and chess pieces; however, if we want to solve a real-world problem in an unstructured environment (e.g., recognize spoken language), it is not clear what constitutes the basic symbols to be processed.

Traditional causal discovery and reasoning assumed that the elementary units are random variables connected by a causal graph. Real-world observations, however, are usually not structured into such units to begin with. For instance, objects in images that permit causal reasoning first need to be discovered [84, 85, 149, 157]. The emerging field of *causal representation learning* strives to learn these variables from data, much like machine learning went beyond symbolic AI in not requiring that the symbols that algorithms manipulate be given a priori (see Figure 11).

Defining objects or variables, and structural models connecting them, can sometimes be achieved by coarse-graining of microscopic models, including microscopic SCMs [123], ordinary differential equations [122], and temporally aggregated time series [37]. While most causal models in economics, medicine, or psychology use variables that are abstractions of more elementary concepts, it is challenging to state general conditions under which coarse-grained variables admit causal models with well-defined interventions [15, 16, 123]. The task of identifying suitable units that admit causal models aligns with the general goal of modern machine learning to learn meaningful representations for data, where meaningful can mean *robust*, *transferable*, *interpretable*, *explainable*, or *fair* [70–72, 77, 155, 162]. To combine structural causal modeling (Definition 4.4) and representation learning, we may try to devise machine learning models whose inputs may be high-dimensional and unstructured, but whose inner workings are (partly) governed by an SCM.

Suppose that our high-dimensional, low-level observations $\mathbf{X} = (X_1, \ldots, X_d)$ are explained by a small number of unobserved, or *latent*, variables $\mathbf{S} = (S_1, \ldots, S_n)$ where $n \ll d$, in that $\mathbf{X}$ is generated by applying an injective map $g : \mathbb{R}^n \to \mathbb{R}^d$ to $\mathbf{S}$ (see Figure 12c),

$$\mathbf{X} = g(\mathbf{S}). \tag{10.1}$$

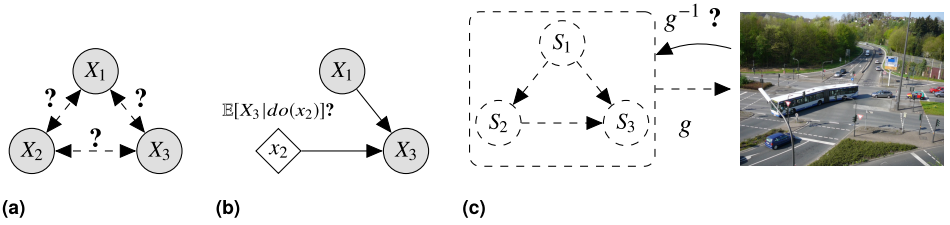

FIGURE 12

Overview of different causal learning tasks: (a) *causal discovery* (Section 7) aims to learn the causal graph (or SCM) connecting a set of *observed* variables; (b) *causal reasoning* (Section 9) aims to answer interventional or counterfactual queries based on a (partial) causal model over observed variables X_i; (c) *causal representation learning* (Section 10) aims to infer a causal model consisting of a small number of high-level, abstract causal variables S_i and their relations from potentially high-dimensional, low-level observations $\mathbf{X} = g(\mathbf{S})$.

A common assumption regarding (10.1) is that the latent S_i are jointly independent, e.g., for independent component analysis (ICA) [57] (where g is referred to as a *mixing*) or disentangled representation learning [9] (where g is called a *decoder*). Presently, however, we instead want think of the latent S_i as *causal variables* that support interventions and reasoning.

The S_i may thus well be dependent, and possess a causal factorization (4.1),

$$p(S_1, \ldots, S_n) = \prod_{i=1}^{n} p(S_i|\mathbf{PA}_i), \tag{10.2}$$

induced by an underlying (acyclic) SCM $\mathcal{M} = (\mathbf{F}, p_{\mathbf{U}})$ with jointly independent U_i and

$$\mathbf{F} = \{S_i := f_i(\mathbf{PA}_i, U_i)\}_{i=1}^{n}. \tag{10.3}$$

Our goal is to learn a latent causal model consisting of (i) the causal representation $\mathbf{S} = g^{-1}(\mathbf{X})$, along with (ii) the corresponding causal graph and (iii) the mechanisms $p(S_i|\mathbf{PA}_i)$ or f_i. This is a challenging task, since none of them are directly observed or known a priori; instead we typically only have access to observations of $\mathbf{X}$. In fact, there is no hope in an i.i.d. setting since already the simpler case with independent S_i (and $n = d$) is not identifiable in general (i.e., for arbitrary nonlinear g in (10.1)): even independence does not sufficiently constrain the problem to uniquely recover, or identify, the true S_i's up to any simple class of ambiguities such as permutations and elementwise invertible transformations of the S_i [58].

To link causal representation learning to the well-studied ICA setting with independent latents in (10.1), we can consider the so-called *reduced form* of an (acyclic) SCM: by recursive substitution of the structural assignments (10.3) in topological order of the causal graph, we can write the latent causal variables $\mathbf{S}$ as function of the noise variables only

$$\mathbf{S} = f_{\mathrm{RF}}(\mathbf{U}). \tag{10.4}$$

Due to acyclicity, this mapping $f_{\mathrm{RF}} : \mathbb{R}^n \to \mathbb{R}^n$ has a lower triangular Jacobian (possibly after reordering the S_i without loss of generality). However, (10.4) is strictly less informative

than (10.3): while they entail the same distribution (10.2), the former no longer naturally supports interventions on the S_i but only changes to the noise distribution $p_{\mathbf{U}}$ (an example of a so-called *soft* intervention [30]). At the same time, the reduced form (10.4) allows us to rewrite (10.1) as

$$\mathbf{X} = g \circ f_{\text{RF}}(\mathbf{U}). \tag{10.5}$$

Through this lens, the task of learning the reduced form (10.4) could be seen as structured form of nonlinear ICA (i.e., (10.1) with independent latents) where we additionally want to learn an intermediate representation through f_{RF}. However, as discussed, we cannot even solve the problem with independent latents (i.e., identify $g \circ f_{\text{RF}}$ in (10.5)) [58], let alone separate the SCM and mixing functions to recover the intermediate causal representation.

It is not surprising that is is not possible to solve the strictly harder causal representation learning problem in an i.i.d. setting and that additional causal learning signals are needed. This gives rise to the following questions: How can we devise causal training algorithms to learn the S_i? And, what types of additional data, assumptions, and constraints do they require beyond the i.i.d. setting? Two general ideas are to (i) build on the ICM Principle 5.1 and enforce some form of (algorithmic) independence between the learned causal mechanisms $p(S_i|\mathbf{PA}_i)$ or f_i, and (ii) use heterogeneous (*non-i.i.d.*) data, e.g., from multiple views or different environments, arising from interventions in the underlying latent SCM (10.3). We briefly discuss some more concrete ideas based on recent work.

Generative approach: Causal autoencoders. One approach is to try to learn the generative causal model (10.1) and (10.3), or its reduced form (10.4), using an *autoencoder* approach [73]. An autoencoder consists of an *encoder* function $q : \mathbb{R}^d \to \mathbb{R}^n$ which maps $\mathbf{X}$ to a latent "bottleneck" representation (e.g., comprising the unexplained noise variables $\mathbf{U}$), and a *decoder* function $\hat{g} : \mathbb{R}^n \to \mathbb{R}^d$ mapping back to the observations. For example, the decoder may directly implement the composition $\hat{g} = g \circ f_{\text{RF}}$ from (10.4). Alternatively, it could consist of multiple modules, implementing (10.1) and (10.3) separately. A standard procedure to train such an autoencoder architecture is to minimize the reconstruction error, i.e., to satisfy $\hat{g} \circ q \approx \text{id}$ on a training set of observations of $\mathbf{X}$. As discussed, this alone is insufficient, so to make it causal we can impose additional constraints on the structure of the decoder [80] and try to make the causal mechanisms independent by ensuring that they are invariant across problems and can be independently intervened upon. For example, if we intervene on the causal variables S_i or noise distribution $p_{\mathbf{U}}$ in our model of (10.3) or (10.4), respectively, this should still produce "valid" observations, as assessed, e.g., by the discriminator of a generative adversarial network [38]. While we ideally want to manipulate the causal variables, another way to intervene is to replace noise variables with the corresponding values computed from other input images, a procedure that has been referred to as hybridization [11]. Alternatively, if we have access to multiple environments, i.e., datasets collected under different conditions, we could rely on the Sparse Mechanism Shift Principle 5.2 by requiring that changes can be explained by shifts in only a few of the $p(S_i|\mathbf{PA}_i)$.

Discriminative approach: Self-supervised causal representation learning. A different machine learning approach for unsupervised representation learning, that is not based on

generative modeling but is discriminative in nature, is *self-supervised learning with data augmentation*. Here, the main idea is to apply some hand-crafted transformations to the observation to generate augmented views that are thought to share the main semantic characteristics with the original observation (e.g., random crops or blurs for images). One then directly learns a representation by maximizing the similarity across encodings of views related to each other by augmentations, while enforcing diversity across those of unrelated views. In a recent work [158], we set out to better understand this approach theoretically, as well as to investigate its potential for learning causal representations. Starting from (10.1), we postulate a latent causal model of the form $\mathbf{S}_c \to \mathbf{S}_s$, where $\mathbf{S}_c$ is a (potentially multivariate) *content* variable, defined as the high-level semantic part of the representation $\mathbf{S} = (\mathbf{S}_c, \mathbf{S}_s)$ that is assumed invariant across views; and $\mathbf{S}_s$ is a (potentially multivariate) *style* variable, defined as the remaining part of the representation that may change. Within this setting, data augmentations have a natural interpretation as counterfactuals under a hypothetical intervention on the style variables, given the original view. It can be shown that in this case, subject to some technical assumptions, common contrastive self-supervised learning algorithms [19, 45, 152] as well as appropriately constrained generative models isolate, or recover, the true content variables $\mathbf{S}_c$ up to an invertible transformation. By extending this approach to use multiple augmented views of the same observation, and linking these to different counterfactuals in the underlying latent SCM, it may be possible to recover a more-fine-grained causal representation.

Independent mechanism analysis. We also explored [40] to what extent the ICM Principle 5.1 may be useful for unsupervised representation learning tasks such as (10.1), particularly for imposing additional constraints on the mixing function g. It turns out that independence between $p(\mathbf{S})$ and the mixing g—measured, e.g., as discussed in Section 5 in the context of Figure 6 and [68]—does not impose nontrivial constraints when $\mathbf{S}$ is not observed, even when the S_i are assumed independent as in ICA. However, by thinking of each S_i as independently *influencing* the observed distribution, we postulate another type of independence between the partial derivatives $\frac{\partial g}{\partial S_i}$ of the mixing g which has a geometric interpretation as an orthogonality condition on the columns of the Jacobian of g. The resulting *independent mechanism analysis* (IMA) approach rules out some of the common examples of nonidentifiability of nonlinear ICA [58, 83] mentioned above. Since IMA does not require independent sources, it may also be a useful constraint for causal representation learning algorithms.

Learning transferable mechanisms and multitask learning. Machine learning excels in i.i.d. settings, and through the use of high capacity learning algorithms we can achieve outstanding performance on many problems, provided we have i.i.d. data for each individual problem (Section 2). However, natural intelligence excels at generalizing across tasks and settings. Suppose we want to build a system that can solve multiple tasks in multiple environments. If we view learning as data compression, it would make sense for that system to utilize components that apply across tasks and environments, and thus need to be stored only once [125].

Indeed, an artificial or natural agent in a complex world is faced with limited resources. This concerns training data, i.e., we only have limited data for each individual task/domain, and thus need to find ways of pooling/reusing data, in stark contrast to the current industry practice of large-scale labeling work done by humans. It also concerns computational resources: animals have constraints on the resources (e.g., space, energy) used by their brains, and evolutionary neuroscience knows examples where brain regions get repurposed. Similar constraints apply as machine learning systems get embedded in physical devices that may be small and battery-powered. Versatile AI models that robustly solve a range of problems in the real world will thus likely need to reuse components, which requires that the components are robust across tasks and environments [127,133]. This calls for a structure whose modules are maximally reusable. An elegant way to do this would be to employ a modular structure that mirrors modularity that exists in the world. In other words, if the are mechanisms at play in the world play similar roles across a range of environments, tasks, and settings, then it would be prudent for a model to employ corresponding computational modules [39]. For instance, if variations of natural lighting (the position of the sun, clouds, etc.) imply that the visual environment can appear in brightness conditions spanning several orders of magnitude, then visual processing algorithms in our nervous system should employ methods that can factor out these variations, rather than building separate sets of object recognizers for every lighting condition. If our brain were to model the lighting changes by a gain control mechanism, say, then this mechanism in itself need not have anything to do with the physical mechanisms bringing about brightness differences. It would, however, play a role in a modular structure that corresponds to the role the physical mechanisms play in the world's modular structure—in other words, it would *represent* the physical mechanism. Searching for the most versatile, yet compact, models would then automatically produce a bias towards models that exhibit certain forms of structural isomorphy to a world that we cannot directly recognize.

A sensible inductive bias to learn such models is to look for independent causal mechanisms [82], and competitive training can play a role in this: for a pattern recognition task, learning causal models that contain independent mechanisms helps in transferring modules across substantially different domains [99].

Interventional world models, surrogate models, digital twins, and reasoning. Modern representation learning excels at learning representations of data that preserve relevant statistical properties [9,79]. It does so, however, without taking into account causal properties of the variables, i.e., it does not care about the interventional properties of the variables it analyzes or reconstructs. Going forward, causality will play a major role in taking representation learning to the next level, moving beyond the representation of statistical dependence structures towards models that support intervention, planning, and reasoning. This would realize Konrad Lorenz' notion of *thinking* as *acting in an imagined space*. It would also provide a means to learn causal *digital twins* that go beyond reproducing statistical dependences captured by *surrogate models* trained using machine learning.

The idea of surrogate modeling is that we may have a complex phenomenon for which we have access to computationally expensive simulation data. If the mappings involved (e.g., from parameter settings to target quantities) can be fitted from data, we can employ machine learning, which will often speed them up by orders of magnitude. Such a speed-up can qualitatively change the usability of a model: for instance, we have recently built a system to map gravitational wave measurements to a probability distribution of physical parameters of a black hole merger event, including sky position [27]. The fact that this model only requires seconds to evaluate makes it possible to immediately start electromagnetic follow-up observations using telescopes as soon as a gravitational wave event has been detected, enabling analysis of transient events.

Going forward, we anticipate that surrogate modeling will benefit from respecting the causal factorization (4.1) decomposing the overall dependence structure into mechanisms (i.e., causal conditionals). We can then build an overall model of a system by modeling the mechanisms independently, each of them using the optimal method. Some of the conditionals we may know analytically, some we may be able to transfer from related problems, if they are invariant. For some, we may have access to real data to estimate them, and for others, we may need to resort to simulations, possibly fitted using surrogate models.

If the model is required to fully capture the effects of all possible interventions, then all components should be fitted as described in the causal directions (i.e., we fit the causal mechanisms). Such a model then allows employing all the causal reasoning machinery described in Sections 4 and 9 (e.g., computing interventional and, in the case of SCMs, counterfactual distributions). If, on the other hand, a model only needs to capture *some* of the possible interventions, and is used in a purely predictive/observational mode for other variables, then we can get away with also using and fitting some noncausal modules, i.e., using a decomposition which lies in-between (4.1) and (4.2).

We believe that this overall framework will be a principled and powerful approach to build such (causal) digital twins or causal surrogate models by combining a range of methods and bringing them to bear according to their strengths.

Concluding remarks. Most of the discussed fields are still in their infancy, and the above account is biased by personal taste and knowledge. With the current hype around machine learning, there is much to say in favor of some humility towards what machine learning can do, and thus towards the current state of AI—the hard problems have not been solved yet, making basic research in this field all the more exciting.

ACKNOWLEDGMENTS

Many thanks to all past and present members of the Tübingen causality team, and to Cian Eastwood and Elias Bareinboim for feedback on the manuscript.

REFERENCES

[1] J. Aldrich, Autonomy. *Oxf. Econ. Pap.* **41** (1989), 15–34.

[2] J. D. Angrist, G. W. Imbens, and D. B. Rubin, Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* **91** (1996), no. 434, 444–455.

[3] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, Invariant risk minimization. 2019, arXiv:1907.02893.

[4] E. Bareinboim and J. Pearl, Transportability from multiple environments with limited experiments: completeness results. In *Advances in neural information processing systems 27*, pp. 280–288, Curran Associates, Inc., 2014.

[5] E. Bareinboim and J. Pearl, Causal inference and the data-fusion problem. *Proc. Natl. Acad. Sci.* **113** (2016), no. 27, 7345–7352.

[6] S. Bauer, B. Schölkopf, and J. Peters, The arrow of time in multivariate time series. In *Proceedings of the 33nd international conference on machine learning 48*, pp. 2043–2051, PMLR, 2016.

[7] M. Belkin, D. Hsu, S. Ma, and S. Mandal, Reconciling modern machine learning practice and the bias-variance trade-off. 2018, arXiv:1812.11118.

[8] S. Ben-David, T. Lu, T. Luu, and D. Pál, Impossibility theorems for domain adaptation. In *Proceedings of the international conference on artificial intelligence and statistics 13 (AISTATS)*, pp. 129–136, PMLR, 2010.

[9] Y. Bengio, A. Courville, and P. Vincent, Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35** (2013), no. 8, 1798–1828.

[10] B. Benneke, I. Wong, C. Piaulet, H. A. Knutson, I. J. M. Crossfield, J. Lothringer, C. V. Morley, P. Gao, T. P. Greene, C. Dressing, D. Dragomir, A. W. Howard, P. R. McCullough, E. M. R. K. J. J. Fortney, and J. Fraine, Water vapor on the habitable-zone exoplanet K2-18b. 2019, arXiv:1909.04642.

[11] M. Besserve, A. Mehrjou, R. Sun, and B. Schölkopf, Counterfactuals uncover the modular structure of deep generative models. In *International conference on learning representations*, OpenReview.net, 2020.

[12] M. Besserve, N. Shajarisales, B. Schölkopf, and D. Janzing, Group invariance principles for causal generative models. In *Proceedings of the 21st international conference on artificial intelligence and statistics (AISTATS)*, pp. 557–565, PMLR, 2018.

[13] S. Bongers, P. Forré, J. Peters, and J. M. Mooij, Foundations of structural causal models with cycles and latent variables. *Ann. Statist.* **49** (2021), no. 5, 2885–2915.

[14] D. Buchsbaum, S. Bridgers, D. Skolnick Weisberg, and A. Gopnik, The power of possibility: Causal learning, counterfactual reasoning, and pretend play. *Philos. Trans. R. Soc. B, Biol. Sci.* **367** (2012), no. 1599, 2202–2212.

[15] K. Chalupka, F. Eberhardt, and P. Perona, Multi-level cause-effect systems. In *Artificial intelligence and statistics*, pp. 361–369, PMLR, 2016.

[16] K. Chalupka, F. Eberhardt, and P. Perona, Causal feature learning: an overview. *Behaviormetrika* **44** (2017), no. 1, 137–164.

[17] O. Chapelle, B. Schölkopf, and A. Zien (eds.), *Semi-supervised learning*. MIT Press, Cambridge, MA, USA, 2006.

[18] C. R. Charig, D. R. Webb, S. R. Payne, and J. E. Wickham, Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *Br. Med. J. (Clin. Res. Ed.)* **292** (1986), no. 6524, 879–882.

[19] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, A simple framework for contrastive learning of visual representations. 2020, arXiv:2002.05709.

[20] D. M. Chickering, Learning Bayesian networks is NP-complete. In *Learning from data*, pp. 121–130, Springer, 1996.

[21] D. M. Chickering, Optimal structure identification with greedy search. *J. Mach. Learn. Res.* **3** (2002), 507–554.

[22] G. F. Cooper and E. Herskovits, A Bayesian method for the induction of probabilistic networks from data. *Mach. Learn.* **9** (1992), no. 4, 309–347.

[23] D. R. Cox, *Planning of experiments*, Wiley, 1958.

[24] P. Daniušis, D. Janzing, J. M. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf, Inferring deterministic causal relations. In *Proceedings of the 26th annual conference on uncertainty in artificial intelligence (UAI)*, pp. 143–150, AUAI Press, 2010.

[25] A. P. Dawid, Conditional independence in statistical theory. *J. R. Stat. Soc. Ser. B.* **41** (1979), no. 1, 1–31.

[26] A. P. Dawid, Causal inference without counterfactuals. *J. Amer. Statist. Assoc.* **95** (2000), no. 450, 407–424.

[27] M. Dax, S. R. Green, J. Gair, J. H. Macke, A. Buonanno, and B. Schölkopf, Real-time gravitational-wave science with neural posterior estimation. *Phys. Rev. Lett.* **127** (2021), no. 24.

[28] L. Devroye, L. Györfi, and G. Lugosi, *A probabilistic theory of pattern recognition*. Appl. Math. 31, Springer, New York, NY, 1996.

[29] V. Didelez, S. Meng, and N. A. Sheehan, Assumptions of IV methods for observational epidemiology. *Statist. Sci.* **25** (2010), 22–40.

[30] F. Eberhardt and R. Scheines, Interventions and causal inference. *Philos. Sci.* **74** (2007), no. 5, 981–995.

[31] R. A. Fisher, *The design of experiments 2*. Oliver & Boyd, Edinburgh & London, 1937.

[32] D. Foreman-Mackey, B. T. Montet, D. W. Hogg, T. D. Morton, D. Wang, and B. Schölkopf, A systematic search for transiting planets in the K2 data. *Astrophys. J.* **806** (2015), no. 2.

[33] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf, Kernel measures of conditional dependence. In *Advances in neural information processing systems*, pp. 489–496, Curran Associates, Inc., 2008.

[34] D. Geiger and D. Heckerman, Learning Gaussian networks. In *Proceedings of the tenth international conference on uncertainty in artificial intelligence*, pp. 235–243, AUAI Press, 1994.

[35] D. Geiger and J. Pearl, Logical and algorithmic properties of independence and their application to Bayesian networks. *Ann. Math. Artif. Intell.* **2** (1990), 165–178.

[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*, pp. 2839–2848, PMLR, 2016.

[37] M. Gong, K. Zhang, B. Schölkopf, C. Glymour, and D. Tao, Causal discovery from temporally aggregated time series. In *Proceedings of the thirty-third conference on uncertainty in artificial intelligence*, pp. 1066–1075, AUAI Press, 2017.

[38] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial nets. In *Advances in neural information processing systems 27*, pp. 2672–2680, Curran Associates, Inc., 2014.

[39] A. Goyal, A. Lamb, J. Hoffmann, S. Sodhani, S. Levine, Y. Bengio, and B. Schölkopf, Recurrent independent mechanisms. In *International conference on learning representations*, OpenReview.net, 2021.

[40] L. Gresele, J. von Kügelgen, V. Stimper, B. Schölkopf, and M. Besserve, Independent mechanism analysis, a new concept? In *Advances in neural information processing systems 34*, Curran Associates, Inc., 2021.

[41] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola, A kernel method for the two-sample-problem. In *Advances in neural information processing systems 19*, pp. 513–520, Curran Associates, Inc., 2007.

[42] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, Measuring statistical dependence with Hilbert–Schmidt norms. In *Algorithmic learning theory*, pp. 63–78, Springer, 2005.

[43] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola, A kernel statistical test of independence. In *Advances in neural information processing systems 20*, pp. 585–592, Curran Associates, Inc., 2008.

[44] A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf, Kernel methods for measuring independence. *J. Mach. Learn. Res.* **6** (2005), 2075–2129.

[45] U. M. Gutmann and A. Hyvärinen, Noise-contrastive estimation: a new estimation principle for unnormalized statistical models. In *International conference on artificial intelligence and statistics*, pp. 297–304, PMLR, 2010.

[46] T. Haavelmo, The probability approach in econometrics. *Econometrica* (1944), iii–115.

[47] D. Heckerman, D. Geiger, and D. M. Chickering, Learning Bayesian networks: the combination of knowledge and statistical data. *Mach. Learn.* **20** (1995), no. 3, 197–243.

[48] D. Heckerman, C. Meek, and G. Cooper, A Bayesian approach to causal discovery. In *Innovations in machine learning*, pp. 1–28, Springer, 2006.

[49] M. A. Hernán, D. Clayton, and N. Keiding, The Simpson's paradox unraveled. *Int. J. Epidemiol.* **40** (2011), no. 3, 780–785.

[50] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, Beta-VAE: learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, OpenReview.net, 2017.

[51] P. W. Holland, Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** (1986), no. 396, 945–960.

[52] K. D. Hoover, *Causality in macroeconomics*. Cambridge University Press, 2001.

[53] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems 21*, pp. 689–696, Curran Associates, Inc., 2009.

[54] B. Huang, K. Zhang, J. Zhang, R. Sanchez-Romero, C. Glymour, and B. Schölkopf, Behind distribution shift: mining driving forces of changes and causal arrows. In *IEEE 17th international conference on data mining (ICDM 2017)*, pp. 913–918, IEEE, 2017.

[55] B. Huang, K. Zhang, J. Zhang, J. D. Ramsey, R. Sanchez-Romero, C. Glymour, and B. Schölkopf, Causal discovery from heterogeneous/nonstationary data. *J. Mach. Learn. Res.* **21** (2020), no. 89, 1–53.

[56] Y. Huang and M. Valtorta, Pearl's calculus of intervention is complete. In *Proceedings of the twenty-second conference on uncertainty in artificial intelligence*, pp. 217–224, AUAI Press, 2006.

[57] A. Hyvärinen and E. Oja, Independent component analysis: algorithms and applications. *Neural Netw.* **13** (2000), no. 4–5, 411–430.

[58] A. Hyvärinen and P. Pajunen, Nonlinear independent component analysis: existence and uniqueness results. *Neural Netw.* **12** (1999), no. 3, 429–439.

[59] A. Hyvarinen, H. Sasaki, and R. Turner, Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *The 22nd international conference on artificial intelligence and statistics*, pp. 859–868, PMLR, 2019.

[60] G. W. Imbens and T. Lemieux, Regression discontinuity designs: a guide to practice. *J. Econometrics* **142** (2008), no. 2, 615–635.

[61] G. W. Imbens and D. B. Rubin, *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[62] D. Janzing, R. Chaves, and B. Schölkopf, Algorithmic independence of initial condition and dynamical law in thermodynamics and causal inference. *New J. Phys.* **18** (2016), no. 093052, 1–13.

[63] D. Janzing, P. Hoyer, and B. Schölkopf, Telling cause from effect based on high-dimensional observations. In *Proceedings of the 27th international conference on machine learning*, edited by J. Fürnkranz and T. Joachims, pp. 479–486, PMLR, 2010.

[64] D. Janzing, J. Peters, J. M. Mooij, and B. Schölkopf, Identifying confounders using additive noise models. In *Proceedings of the 25th annual conference on uncertainty in artificial intelligence (UAI)*, pp. 249–257, AUAI Press, 2009.

[65] D. Janzing and B. Schölkopf, Causal inference using the algorithmic Markov condition. *IEEE Trans. Inf. Theory* **56** (2010), no. 10, 5168–5194.

[66] D. Janzing and B. Schölkopf, Semi-supervised interpolation in an anticausal learning scenario. *J. Mach. Learn. Res.* **16** (2015), 1923–1948.

[67] D. Janzing and B. Schölkopf, Detecting non-causal artifacts in multivariate linear regression models. In *Proceedings of the 35th international conference on machine learning (ICML)*, pp. 2250–2258, PMLR, 2018.

[68] D. Janzing, J. M. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, and B. Schölkopf, Information-geometric approach to inferring causal directions. *Artificial Intelligence* **182–183** (2012), 1–31.

[69] 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)*, pp. 9499–9513, Association for Computational Linguistics, 2021.

[70] A.-H. Karimi, B. Schölkopf, and I. Valera, Algorithmic recourse: from counterfactual explanations to interventions. In *Conference on fairness, accountability, and transparency*, pp. 353–362, ACM, 2021.

[71] A.-H. Karimi, 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*, pp. 265–277, Curran Associates, Inc., 2020.

[72] N. Kilbertus, M. Rojas Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, Avoiding discrimination through causal reasoning. In *Advances in neural information processing systems 30*, pp. 656–666, Curran Associates, Inc., 2017.

[73] D. P. Kingma and M. Welling, Auto-encoding variational Bayes. 2013, arXiv:1312.6114.

[74] F. Klein, *Vergleichende Betrachtungen über neuere geometrische Forschungen*. Verlag von Andreas Deichert, Erlangen, 1872.

[75] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT Press, 2009.

[76] S. Kpotufe, E. Sgouritsa, D. Janzing, and B. Schölkopf, Consistency of causal inference under the additive noise model. In *Proceedings of the 31th international conference on machine learning*, pp. 478–486, PMLR, 2014.

[77] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, Counterfactual fairness. In *Advances in neural information processing systems 30*, pp. 4066–4076, Curran Associates, Inc., 2017.

[78] S. L. Lauritzen, *Graphical models. 17*. Clarendon Press, 1996.

[79] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning. *Nature* **521** (2015), no. 7553, 436–444.

[80] F. Leeb, Y. Annadani, S. Bauer, and B. Schölkopf, Structural autoencoders improve representations for generation and transfer. 2020, arXiv:2006.07796.

[81] G. W. Leibniz, *Discours de métaphysique*, 1686 (cited after Chaitin, 2010).

[82] F. Locatello, D. Vincent, I. Tolstikhin, G. Rätsch, S. Gelly, and B. Schölkopf, Competitive training of mixtures of independent deep generative models. 2018, arXiv:1804.11130.

[83] F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem, Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th international conference on machine learning*, PMLR, 2019.

[84] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf, Object-centric learning with slot attention. In *Advances in neural information processing systems*, pp. 11525–11538, Curran Associates, Inc., 2020.

[85] D. Lopez-Paz, R. Nishihara, S. Chintala, B. Schölkopf, and L. Bottou, Discovering causal signals in images. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 58–66, IEEE Computer Society, 2017.

[86] J. Loschmidt, Über den Zustand des Wärmegleichgewichtes eines Systems von Körpern mit Rücksicht auf die Schwerkraft. *Sitzungsber. Akad. Wiss. Wien, Math.-Naturwiss. Kl.* **73** (1876), 128–142.

[87] C. Lu, Y. Wu, J. M. Hernández-Lobato, and B. Schölkopf, Nonlinear invariant risk minimization: a causal approach. 2021, arXiv:2102.12353.

[88] S. MacLane, *Categories for the working mathematician*. Springer, New York, 1971.

[89] R. Matthews, Storks deliver babies ($p = 0.008$). *Teach. Stat.* **22** (2000), no. 2, 36–38.

[90] C. Meek, Causal inference and causal explanation with background knowledge. In *Proceedings of the eleventh conference on uncertainty in artificial intelligence*, pp. 403–410, Morgan Kaufmann Publishers Inc., 1995.

[91] F. H. Messerli, Chocolate consumption, cognitive function, and Nobel laureates. *N. Engl. J. Med.* **367** (2012), no. 16, 1562–1564.

[92] B. T. Montet, T. D. Morton, D. Foreman-Mackey, J. A. Johnson, D. W. Hogg, B. P. Bowler, D. W. Latham, A. Bieryla, and A. W. Mann, Stellar and planetary properties of K2 campaign 1 candidates and validation of 17 planets, including a planet receiving earth-like insolation. *Astrophys. J.* **809** (2015), no. 1, 25.

[93] R. P. Monti, K. Zhang, and A. Hyvärinen, Causal discovery with general non-linear relationships using non-linear ICA. In *Uncertainty in artificial intelligence*, pp. 186–195, PMLR, 2020.

[94] J. M. Mooij, D. Janzing, T. Heskes, and B. Schölkopf, On causal discovery with cyclic additive noise models. In *Advances in neural information processing systems 24 (NIPS)*, pp. 639–647, Curran Associates, Inc., 2011.

[95] J. M. Mooij, D. Janzing, J. Peters, and B. Schölkopf, Regression by dependence minimization and its application to causal inference. In *Proceedings of the 26th international conference on machine learning (ICML)*, pp. 745–752, PMLR, 2009.

[96] J. M. Mooij, D. Janzing, and B. Schölkopf, From ordinary differential equations to structural causal models: the deterministic case. In *Proceedings of the 29th annual conference on uncertainty in artificial intelligence (UAI)*, pp. 440–448, AUAI Press, 2013.

[97] J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf, Distinguishing cause from effect using observational data: methods and benchmarks. *J. Mach. Learn. Res.* **17** (2016), no. 32, 1–102.

[98] J. S. Neyman, On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Ann. Agric. Sci.* **10** (1923), 1–51. (Translated and edited by D. M. Dabrowska and T. P. Speed, *Statist. Sci.* **5** (1990), 465–480).

[99] G. Parascandolo, N. Kilbertus, M. Rojas-Carulla, and B. Schölkopf, Learning independent causal mechanisms. In *Proceedings of the 35th international conference on machine learning (PMLR) 80*, pp. 4036–4044, PMLR, 2018.

[100] J. Park and K. Muandet, A measure-theoretic approach to kernel conditional mean embeddings. In *Advances in neural information processing systems 33 (NEURIPS 2020)*, pp. 21247–21259, Curran Associates, Inc., 2020.

[101] J. Pearl, Bayesian networks: a model of self-activated memory for evidential reasoning. In *Proceedings of the 7th conference of the cognitive science society*, pp. 329–334, University of California (Los Angeles), Computer Science Department, 1985.

[102] J. Pearl, Causal diagrams for empirical research. *Biometrika* 82 (1995), no. 4, 669–-688.

[103] J. Pearl, Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, pp. 411–420, AUAI Press, 2001.

[104] J. Pearl, *Causality: models, reasoning, and inference*. 2nd edn., Cambridge University Press, New York, NY, 2009.

[105] J. Pearl, Comment: understanding Simpson's paradox. *Amer. Statist.* **68** (2014), no. 1, 8–13.

[106] J. Pearl and E. Bareinboim, External validity: From do-calculus to transportability across populations. *Statist. Sci.* **29** (2014), no. 4, 579–595.

[107] J. Pearl and D. Mackenzie, *The book of why: the new science of cause and effect*. Basic Books, 2018.

[108] J. Pearl and A. Paz, Confounding equivalence in causal inference. *J. Causal Inference* 2 (2014), no. 1, 75-–93.

[109] J. Pearl and T. Verma, A theory of inferred causation. In *Principles of knowledge representation and reasoning: proceedings of the second international conference 2*, p. 441, Morgan Kaufmann, 1991.

[110] J. Peters, P. Bühlmann, and N. Meinshausen, Causal inference by using invariant prediction: identification and confidence intervals. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** (2016), no. 5, 947–1012.

[111] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference – foundations and learning algorithms*. MIT Press, Cambridge, MA, USA, 2017.

[112] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf, Identifiability of causal graphs using functional models. In *Proceedings of the 27th annual conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 589–598, AUIA Press, 2011.

[113] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf, Causal discovery with continuous additive noise models. *J. Mach. Learn. Res.* **15** (2014), 2009–2053.

[114] N. Pfister, P. Bühlmann, B. Schölkopf, and J. Peters, Kernel-based tests for joint independence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** (2018), no. 1, 5–31.

[115] K. Popper, *The logic of scientific discovery*. 1959.

[116] H. Reichenbach, *The direction of time*. University of California Press, Berkeley, CA, 1956.

[117] J. Robins, A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math. Model.* **7** (1986), no. 9–12, 1393–1512.

[118] J. M. Robins, M. A. Hernan, and B. Brumback, Marginal structural models and causal inference in epidemiology. *Epidemiology* **11** (2000), no. 5, 550–560.

[119] R. W. Robinson, Counting labeled acyclic digraphs, new directions in the theory of graphs. In *Proc. third Ann Arbor conf., Univ. Michigan, Ann Arbor, MI, 1971*, pp. 239–273, Academic Press, 1973.

[120] M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters, Invariant models for causal transfer learning. *J. Mach. Learn. Res.* **19** (2018), no. 36, 1–34.

[121] P. R. Rosenbaum and D. B. Rubin, The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** (1983), no. 1, 41–55.

[122] P. K. Rubenstein, S. Bongers, B. Schölkopf, and J. M. Mooij, From deterministic ODEs to dynamic structural causal models. In *Proceedings of the 34th conference on uncertainty in artificial intelligence (UAI)*, pp. 114–123, AUAI Press, 2018.

[123] P. K. Rubenstein, S. Weichwald, S. Bongers, J. M. Mooij, D. Janzing, M. Grosse-Wentrup, and B. Schölkopf, Causal consistency of structural equation models. In *Proceedings of the thirty-third conference on uncertainty in artificial intelligence*, pp. 808–817, AUAI Press, 2017.

[124] D. B. Rubin, Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** (1974), no. 5, 688.

[125] B. Schölkopf, Causality for machine learning. 2019, arXiv:1911.10500. To appear in: R. Dechter, J. Halpern, and H. Geffner, *Probabilistic and causal inference: the works of Judea Pearl*. ACM books, 2019.

[126] B. Schölkopf, R. Herbrich, and A. J. Smola, A generalized representer theorem. In *Annual conference on computational learning theory*, edited by D. Helmbold and R. Williamson, pp. 416–426, Lecture Notes in Comput. Sci. 2111, Springer, Berlin, 2001.

[127] B. Schölkopf, D. Janzing, and D. Lopez-Paz, Causal and statistical learning. *Oberwolfach Rep.* **13** (2016), no. 3, 1896–1899.

[128] 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)*, pp. 1255–1262, PMLR, 2012.

[129] B. Schölkopf, K. Muandet, K. Fukumizu, S. Harmeling, and J. Peters, Computing functions of random variables via reproducing kernel Hilbert space representations. *Stat. Comput.* **25** (2015), no. 4, 755–766.

[130] B. Schölkopf and A. J. Smola, *Learning with kernels*. MIT Press, Cambridge, MA, 2002.

[131] B. Schölkopf, B. K. Sriperumbudur, A. Gretton, and K. Fukumizu, RKHS representation of measures applied to homogeneity, independence, and Fourier optics. *Oberwolfach Rep.* **30** (2008), 42–44.

[132] B. Schölkopf, D. Hogg, D. Wang, D. Foreman-Mackey, D. Janzing, C.-J. Simon-Gabriel, and J. Peters, Modeling confounding by half-sibling regression. *Proc. Natl. Acad. Sci.* **113** (2016), no. 27, 7391–7398.

[133] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, Toward causal representation learning. *Proc. IEEE* **109** (2021), no. 5, 612–634.

[134] G. Schwarz, et al., Estimating the dimension of a model. *Ann. Statist.* **6** (1978), no. 2, 461–464.

[135] R. D. Shah and J. Peters, The hardness of conditional independence testing and the generalised covariance measure. *Ann. Statist.* **48** (2020), no. 3, 1514–1538.

[136] N. Shajarisales, D. Janzing, B. Schölkopf, and M. Besserve, Telling cause from effect in deterministic linear dynamical systems. In *Proceedings of the 32nd international conference on machine learning (ICML)*, pp. 285–294, PMLR, 2015.

[137] U. Shalit, F. D. Johansson, and D. Sontag, Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pp. 3076–3085, PMLR, 2017.

[138] C. E. Shannon, Coding theorems for a discrete source with a fidelity criterion. In *IRE international convention records 7*, pp. 142–163, Wiley-IEEE Press, 1959.

[139] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. J. Kerminen, A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.* **7** (2006), 2003–2030.

[140] I. Shpitser and J. Pearl, Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the 21st national conference on artificial intelligence*, pp. 1219–1226, AAAI Press, 2006.

[141] I. Shpitser, T. VanderWeele, and J. M. Robins, On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the twenty-sixth conference on uncertainty in artificial intelligence*, pp. 527–536, AUAI Press, 2010.

[142] E. H. Simpson, The interpretation of interaction in contingency tables. *J. Roy. Statist. Soc. Ser. B* **13** (1951), no. 2, 238–241.

[143] A. J. Smola, A. Gretton, L. Song, and B. Schölkopf, A Hilbert space embedding for distributions. In *Algorithmic learning theory: 18th international conference*, pp. 13–31, Springer, 2007.

[144] P. Spirtes, C. Glymour, and R. Scheines, *Causation, prediction, and search*. 2nd edn., MIT Press, Cambridge, MA, 2000.

[145] W. Spohn, *Grundlagen der Entscheidungstheorie*. Scriptor, 1978.

[146] I. Steinwart and A. Christmann, *Support vector machines*. Springer, New York, NY, 2008.

[147] R. Suter, D. Miladinovic, B. Schölkopf, and S. Bauer, Robustly disentangled causal mechanisms: validating deep representations for interventional robustness. In *International conference on machine learning*, pp. 6056–6065, PMLR, 2019.

[148] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, Intriguing properties of neural networks. 2013, arXiv:1312.6199.

[149] M. Tangemann, S. Schneider, J. von Kügelgen, F. Locatello, P. Gehler, T. Brox, M. Kümmerer, M. Bethge, and B. Schölkopf, Unsupervised object learning via common fate. 2021, arXiv:2110.06562.

[150] J. Tian and J. Pearl, Causal discovery from changes. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, pp. 512–521, Morgan Kaufmann Publishers Inc., 2001.

[151] A. Tsiaras, I. Waldmann, G. Tinetti, J. Tennyson, and S. Yurchenko, Water vapour in the atmosphere of the habitable-zone eight-earth-mass planet K2-18b. *Nat. Astron.* **3** (2019), 1086–1091.

[152] A. van den Oord, Y. Li, and O. Vinyals, Representation learning with contrastive predictive coding. 2018, arXiv:1807.03748.

[153] V. N. Vapnik, *Statistical learning theory*. Wiley, New York, NY, 1998.

[154] J. von Kügelgen, L. Gresele, and B. Schölkopf, Simpson's paradox in Covid-19 case fatality rates: a mediation analysis of age-related causal effects. *IEEE Trans. Artif. Intell.* **2** (2021), no. 1, 18–27.

[155] J. von Kügelgen, A.-H. Karimi, U. Bhatt, I. Valera, A. Weller, and B. Schölkopf, On the fairness of causal algorithmic recourse. In *36th AAAI conference on artificial intelligence*, AAAI Press, 2022.

[156] J. von Kügelgen, A. Mey, M. Loog, and B. Schölkopf, Semi-supervised learning, causality and the conditional cluster assumption. In *Conference on uncertainty in artificial intelligence*, pp. 1–10, AUAI Press, 2020.

[157] J. von Kügelgen, I. Ustyuzhaninov, P. Gehler, M. Bethge, and B. Schölkopf, Towards causal generative scene models via competition of experts. In *ICLR 2020 workshop on causal learning for decision making*, OpenReview.net, 2020.

[158] J. von 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 34*, Curran Associates, Inc., 2021.

[159] J. Woodward. *Causation and manipulability*, Stanford Encyclopedia of Philosophy, 2001.

[160] P. G. Wright, *Tariff on animal and vegetable oils*. Macmillan Company, New York, 1928.

[161] S. Wright, Correlation and causation. *J. Agric. Res.* **20** (1921), 557–580.

[162] J. Zhang and E. Bareinboim, Fairness in decision-making – the causal explanation formula. In *Proceedings of the thirty-second AAAI conference on artificial intelligence*, pp. 2037–2045, AAAI Press, 2018.

[163] K. Zhang, M. Gong, and B. Schölkopf, Multi-source domain adaptation: a causal view. In *Proceedings of the 29th AAAI conference on artificial intelligence*, pp. 3150–3157, AAAI Press, 2015.

[164] K. Zhang and A. Hyvärinen, On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th annual conference on uncertainty in artificial intelligence (UAI)*, pp. 647–655, AUAI Press, 2009.

[165] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf, Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the 27th annual conference on uncertainty in artificial intelligence (UAI)*, pp. 804–813, AUAI Press, 2011.

[166] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang, Domain adaptation under target and conditional shift. In *Proceedings of the 30th international conference on machine learning*, pp. 819–827, PMLR, 2013.

BERNHARD SCHÖLKOPF

Max Planck Institute for Intelligent Systems, Tübingen, Germany, bs@tuebingen.mpg.de

JULIUS VON KÜGELGEN

Max Planck Institute for Intelligent Systems, Tübingen, Germany; and University of Cambridge, Cambridge, United Kingdom, jvk@tuebingen.mpg.de