INVITED PAPER

Toward Causal Representation Learning

This article reviews fundamental concepts of causal inference and relates them to crucial open problems of machine learning, including transfer learning and generalization, thereby assaying how causality can contribute to modern machine learning research.

By Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio

ABSTRACT | The two fields of machine learning and graphical causality arose and are developed separately. However, there is, now, cross-pollination and increasing interest in both fields to benefit from the advances of the other. In this article, we review fundamental concepts of causal inference and relate them to crucial open problems of machine learning, including transfer and generalization, thereby assaying how causality can contribute to modern machine learning research. This also applies in the opposite direction: we note that most work in causality starts from the premise that the causal variables are given. A central problem for AI and causality is, thus, causal representation learning, that is, the discovery of high-level causal variables from low-level observations. Finally, we delineate some implications of causality for machine learning and propose key research areas at the intersection of both communities.

KEYWORDS | Artificial intelligence; causality; deep learning; representation learning.

I. INTRODUCTION

If we compare what machine learning can do to what animals accomplish, we observe that the former is rather limited at some crucial feats where natural intelligence excels. These include transfer to new problems and any form of generalization that is not from one data point to the next (sampled from the same distribution), but rather from one problem to the next—both have been termed *generalization*, but the latter is a much harder form thereof, sometimes referred to as *horizontal*, *strong*, or *out-of-distribution* generalization. This shortcoming is not too surprising, given that machine learning often disregards information that animals use heavily: interventions in the world, domain shifts, and temporal structure—by and large, we consider these factors a nuisance and try to engineer them away. In accordance with this, the majority of current successes of machine learning boil down to large-scale pattern recognition on suitably collected *independent and identically distributed (i.i.d.)* data.

To illustrate the implications of this choice and its relation to causal models, we start by highlighting key research challenges.

A. Issue 1—Robustness

With the widespread adoption of deep learning approaches in computer vision [103], [140], natural language processing [55], and speech recognition [86], a substantial body of literature explored the robustness of the prediction of state-of-the-art deep neural network architectures. The underlying motivation originates from the fact that, in the real world, there is often little control over the distribution from which the data come from. In computer vision [76], [228], changes in the test distribution may, for instance, come from aberrations, such as camera blur, noise, or compression quality [107], [129], [170], [206], or from shifts, rotations, or viewpoints [7],

Manuscript received August 14, 2020; revised December 29, 2020; accepted February 8, 2021. Date of publication February 26, 2021; date of current version April 30, 2021. *(Bernhard Schölkopf and Francesco Locatello contributed equally to this work. Stefan Bauer and Nan Rosemary Ke contributed equally to this work.) (Corresponding author: Francesco Locatello.)*

Bernhard Schölkopf and **Stefan Bauer** are with the Max Planck Institute for Intelligent Systems, 72076 Tübingen, Germany (e-mail: bs@tuebingen.mpg.de; stefan.bauer@tuebingen.mpg.de).

Francesco Locatello was with Google Research Amsterdam 1082 MD, The Netherlands. He is now with the Computer Science Department, ETH Zürich, 8092 Zürich, Switzerland, and also with the Max Planck Institute for Intelligent Systems, 72076 Tübingen, Germany (e-mail: francesco.locatello@gmail.com).

Nan Rosemary Ke and **Anirudh Goyal** are with Mila, Montreal, QC H2S 3H1, Canada, and also with the Department of Computer Science and Operational Research, University of Montreal, Montreal, QC H3T 1J4, Canada (e-mail: rosemary.nan.ke@gmail.com; anirudhgoyal9119@gmail.com).

Nal Kalchbrenner is with Google Research Amsterdam 1082 MD, The Netherlands (e-mail: nalk@google.com).

Yoshua Bengio is with Mila, Montreal, QC H2S 3H1, Canada, with the Department of Computer Science and Operational Research, University of Montreal, Montreal, QC H3T 1J4, Canada, and also with CIFAR, Toronto, ON M5G 1M1, Canada (e-mail: yoshua.bengio@mila.quebec).

Digital Object Identifier 10.1109/JPROC.2021.3058954

[12], [64], [282]. Motivated by this, new benchmarks were proposed to specifically test a generalization of classification and detection methods with respect to simple algorithmically generated interventions, such as spatial shifts, blur, changes in brightness or contrast [107], [170], time consistency [95], [227], control over background and rotation [12], as well as images collected in multiple environments [20]. Studying the failure modes of deep neural networks from simple interventions has the potential to lead to insights into the inductive biases of state-of-the-art architectures. So far, there has been no definitive consensus on how to solve these problems, although progress has been made using data augmentation, pretraining, self-supervision, and architectures with suitable inductive biases with respect to a perturbation of interest [60], [64], [137], [170], [206], [233]. It has been argued [188] that such fixes may not be sufficient, and generalizing well outside the i.i.d. setting requires learning not mere statistical associations between variables, but an underlying *causal model*. The latter contains the mechanisms giving rise to the observed statistical dependences and allows to model distribution shifts through the notion of interventions [35], [180], [183], [188], [220], [237].

B. Issue 2—Learning Reusable Mechanisms

Infants' understanding of physics relies upon objects that can be tracked over time and behave consistently [53], [236]. Such a representation allows children to quickly learn new tasks as their knowledge and intuitive understanding of physics can be reused [17], [53], [144], [250]. Similarly, intelligent agents that robustly solve real-world tasks need to reuse and repurpose their knowledge and skills in novel scenarios. Machine learning models that incorporate or learn structural knowledge of an environment have been shown to be more efficient and generalize better [9], [11], [15], [16], [27], [58], [77], [84], [85], [141], [157], [177], [181], [197], [211], [212], [244], [258], [272], [274]. In a modular representation of the world where the modules correspond to physical causal mechanisms, many modules can be expected to behave similarly across different tasks and environments. An agent facing a new environment or task may thus only need to adapt a few modules in its internal representation of the world [85], [219]. When learning a causal model, one should, thus, require fewer examples to adapt as most knowledge, that is, modules, can be reused without further training.

C. Causality Perspective

Causation is a subtle concept that cannot be fully described using the language of Boolean logic [151] or that of probabilistic inference; it requires the additional notion of *intervention* [183], [237]. The manipulative definition of causation [118], [183], [237] focuses on the fact that conditional probabilities ("seeing people with open umbrellas suggests that it is raining") cannot reliably predict the outcome of active intervention ("closing umbrellas does not stop the rain"). Causal relations can also be viewed as the components of reasoning chains [151] that provide predictions for situations that are very far from the observed distribution and may even remain purely hypothetical [163], [183] or require conscious deliberation [128]. In that sense, discovering causal relations means acquiring robust knowledge that holds beyond the support of observed data distribution and a set of training tasks, and it extends to situations involving forms of reasoning.

Our contributions: In this article, we argue that causality, with its focus on representing structural knowledge about the data generating process that allows interventions and changes, can contribute toward understanding and resolving some limitations of current machine learning methods. This would take the field a step closer to a form of artificial intelligence that involves *thinking* in the sense of Konrad Lorenz, that is, acting in an imagined space [163]. Despite its success, statistical learning provides a rather superficial description of reality that only holds when the experimental conditions are fixed. Instead, the field of *causal learning* seeks to model the effect of interventions and distribution changes with a combination of data-driven learning and assumptions not already included in the statistical description of a system. This work reviews and synthesizes key contributions that have been made to this end.[1]

1) We describe different levels of modeling in physical systems in Section II and present the differences between causal and statistical models in Section III. We do so not only in terms of modeling abilities, but also discuss the assumptions and challenges involved.
2) We expand on the independent causal mechanism (ICM) principle as a key component that enables the estimation of causal relations from data in Section IV. In particular, we state the sparse mechanism shift (SMS) hypothesis as a consequence of the ICM principle and discuss its implications for learning causal models.
3) We review existing approaches to learn causal relations from appropriate descriptors (or features) in Section V. We cover both classical approaches and modern reinterpretations based on deep neural networks, with a focus on the underlying principles that enable causal discovery.
4) We discuss how useful models of reality may be learned from data in the form of causal representations and discuss several current problems of machine learning from a causal point of view in Section VI.
5) We assay the implications of causality for practical machine learning in Section VII. Using causal language, we revisit robustness and generalization, as well as existing common practices, such as semi-supervised learning (SSL), self-supervised learning,

[1] The present paper expands [221], leading to partial text overlap.

Table 1 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, but It 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 distr. shift/intervention	Answer counter-factual 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

data augmentation, and pretraining. We discuss examples at the intersection between causality and machine learning in scientific applications and speculate on the advantages of combining the strengths of both fields to build a more versatile AI.

II. LEVELS OF CAUSAL MODELING

The gold standard for modeling natural phenomena is a set of coupled differential equations modeling physical mechanisms responsible for time evolution. This allows us to predict the future behavior of a physical system, reason about the effect of interventions, and predict *statistical* dependencies between variables that are generated by coupled time evolution. It also offers physical insights, explaining the functioning of the system, and lets us read off its causal structure. To this end, consider the coupled set of differential equations:

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d \tag{1}$$

with initial value $\mathbf{x}(t_0) = \mathbf{x}_0$. The Picard–Lindelöf theorem states that, at least locally, if f is Lipschitz, there exists a unique solution $\mathbf{x}(t)$. This implies, in particular, that the immediate future of $\mathbf{x}$ is implied by its past values.

If we formally write this in terms of infinitesimal differentials dt and $d\mathbf{x} = \mathbf{x}(t+dt) - \mathbf{x}(t)$, we get

$$\mathbf{x}(t+dt) = \mathbf{x}(t) + dt \cdot f(\mathbf{x}(t)). \tag{2}$$

From this, we can ascertain which entries of the vector $\mathbf{x}(t)$ mathematically determine the future of others $\mathbf{x}(t+dt)$. This tells us that if we have a physical system whose physical mechanisms are correctly described using such an ordinary differential equation (1), solved for $(d\mathbf{x}/dt)$ (i.e., the derivative only appears on the left-hand side), then its causal structure can be directly read off.[2]

While a differential equation is a rather comprehensive description of a system, a statistical model can be viewed as a much more superficial one. It often does not refer to dynamic processes; instead, it tells us how some of the variables allow the prediction of others as long as experimental conditions do not change. For example, if we drive a differential equation system with certain types of noise, or we average over time, then it may be the case that statistical dependencies between components of $\mathbf{x}$ emerge and those can then be exploited by machine learning. Such a model does not allow us to predict the effect of interventions; however, its strength is that it can often be learned from observational data, while a differential equation usually requires an intelligent human to come up with it. Causal modeling lies in between these two extremes. Like models in physics, it aims to provide the understanding and predict the effect of interventions. However, causal discovery and learning try to arrive at such models in a data-driven way, replacing expert knowledge with weak and generic assumptions. The overall situation is summarized in Table 1, adapted from [188]. In the following, we address some of the tasks listed in Table 1 in more detail.

A. Predicting in the i.i.d. Setting

Statistical models are a superficial description of reality as they are only required to model associations. For a given set of input examples X and target labels Y, we may be interested in approximating $P(Y|X)$ to answer questions, such as "what is the probability that this particular image contains a dog?" or "what is the probability of heart failure given certain diagnostic measurements (e.g., blood pressure) carried out on a patient?" Subject to suitable assumptions, these questions can be provably answered by observing a sufficiently large amount of i.i.d. data from $P(X,Y)$ [257]. Despite the impressive advances of machine learning, causality offers an underexplored complement: accurate predictions may not be sufficient to inform decision-making. For example, the frequency of storks is a reasonable predictor for human birth rates in Europe [168]. However, as there is no direct causal link between these two variables, a change to the stork population would not affect the birth rates, even though a statistical model may predict so. The predictions of a statistical model are only accurate within identical experimental

[2] Note that this requires that the differential equation system describes the causal physical mechanisms. If, in contrast, we considered a set of differential equations that phenomenologically correctly describe the time evolution of a system without capturing the underlying mechanisms (e.g., due to unobserved confounding or a form of course graining that does not preserve the causal structure [208]), then (2) may not be causally meaningful [186], [217].

conditions. Performing an intervention changes the data distribution, which may lead to (arbitrarily) inaccurate predictions [183], [188], [220], [237].

B. Predicting Under Distribution Shifts

Interventional questions are more challenging than predictions as they involve actions that take us out of the usual i.i.d. setting of statistical learning. Interventions may affect both the value of a subset of causal variables and their relations. For example, "is increasing the number of storks in a country going to boost its human birth rate?" and "would fewer people smoke if cigarettes were more socially stigmatized?" As interventions change the joint distribution of the variables of interest, classical statistical learning guarantees [257] no longer apply. On the other hand, learning about interventions may allow training predictive models that are robust against the changes in distribution that naturally happen in the real world. Here, interventions do not need to be deliberate actions to achieve a goal. Statistical relations may change dynamically over time (e.g., people's preferences and tastes), or there may simply be a mismatch between a carefully controlled training distribution and the test distribution of a model deployed in production. The robustness of deep neural networks has recently been scrutinized and become an active research topic related to causal inference. We argue that predicting under distribution shift should not be reduced to just the accuracy on a test set. If we wish to incorporate learning algorithms into human decision-making, we need to trust that the predictions of the algorithm will remain valid if the experimental conditions are changed.

C. Answering Counterfactual Questions

Counterfactual problems involve reasoning about why things happened, imagining the consequences of different actions in hindsight, and determining which actions would have achieved the desired outcome. Answering counterfactual questions can be more difficult than answering interventional questions. However, this may be a key challenge for AI, as an intelligent agent may benefit from imagining the consequences of its actions and understanding in retrospect what led to certain outcomes, at least to some degree of approximation.[3] We have mentioned the example of statistical predictions of heart failure above. An interventional question would be "how does the probability of heart failure change if we convince a patient to exercise regularly?" A counterfactual one would be "would a given patient have suffered heart failure if they had started exercising a year earlier?" As we shall discuss in the following, counterfactuals, or approximations thereof, are especially critical in RL. They can enable agents to reflect on their decisions and formulate hypotheses that can be empirically verified in a process akin to the scientific method.

D. Nature of Data: Observational, Interventional, and (Un)structured

The data format plays a substantial role in which type of relation can be inferred. We can distinguish two axes of data modalities: observational versus interventional, and hand-engineered versus raw (unstructured) perceptual input.

1) Observational and Interventional Data: An extreme form of data which is often assumed but seldom strictly available is observational i.i.d. data, where each data point is independently sampled from the same distribution. Another extreme is interventional data with known interventions, where we observe data sets sampled from multiple distributions each of which is the result of a known intervention. In between, we have data with (domain) shifts or unknown interventions. This is observational in the sense that the data is only observed passively, but it is interventional in the sense that there are interventions/shifts, but unknown to us.

2) Hand-Engineered Data Versus Raw Data: Especially, in classical AI, data are often assumed to be structured into high level and semantically meaningful variables, which may partially (modulo some variables being unobserved) correspond to the causal variables of the underlying graph. *Raw data*, in contrast, are unstructured and do not expose any direct information about causality.

While statistical models are weaker than causal models, they can be efficiently learned from observational data alone on both hand-engineered features and raw perceptual input, such as images, videos, and speech. On the other hand, although methods for learning causal structure from observations exist [18], [37], [83], [113], [123], [139], [161], [174]–[176], [188]–[190], [229], [237], [246], [279], learning causal relations frequently requires collecting data from multiple environments or the ability to perform interventions [251]. In some cases, it is assumed that all common causes of measured variables are also observed (causal sufficiency).[4] Overall, a significant amount of prior knowledge is encoded in which variables are measured. Moving forward, one would hope to develop methods that replace expert data collection with suitable inductive biases and learning paradigms, such as metalearning and self-supervision. If we wish to learn a causal model that is useful for a particular set of tasks and environments, the appropriate granularity of the high-level

[3] Note that two types of questions occupy a continuum: to this end, consider a probability that is both conditional and interventional $P(A|B, do(C))$. If B is an empty set, we have a classical intervention; if B contained all (unobserved) noise terms, we have a counterfactual. If B is not identical to the noise terms, but, nevertheless, informative about them, we get something in between. For instance, reinforcement learning (RL) practitioners may call Q functions as providing counterfactuals even though they model P [return from $t|$ agent state at time t, do (action at time t)] and, therefore, closer to an intervention (which is why they can be estimated from data).

[4] There are also algorithms that do not require causal sufficiency [237].

variables depends on the tasks of interest and on the type of data that we have at our disposal, for example, which interventions can be performed and what is known about the domain.

III. CAUSAL MODELS AND INFERENCE

As discussed, reality can be modeled at different levels, from the physical one to statistical associations between epiphenomena. In this section, we expand on the difference between statistical and causal modeling and review a formal language to talk about interventions and distribution changes.

A. Methods Driven by i.i.d. Data

The machine learning community has produced impressive successes with machine learning applications to big data problems [54], [148], [171], [223], [232]. In these successes, there are several trends at work [215]: 1) we have massive amounts of data, often from simulations or large-scale human labeling; 2) we use high-capacity machine learning systems (i.e., complex function classes with many adjustable parameters); 3) we employ high-performance computing systems; and (often ignored, but crucial when it comes to causality) 4) the problems are i.i.d. The latter can be guaranteed by the construction of a task, including training and test set (e.g., image recognition using benchmark data sets). Alternatively, problems can be made approximately i.i.d., for example, by carefully collecting the right training set for a given application problem, or by methods, such as "experience replay" [171] where an RL agent stores observations in order to later permute them for the purpose of retraining.

For i.i.d. data, strong universal consistency results from statistical learning theory apply, guaranteeing convergence of a learning algorithm to the lowest achievable risk. Such algorithms do exist, for instance, nearest neighbor classifiers, support vector machines, and neural networks [67], [221], [239], [257]. Seen in this light, it is not surprising that we can indeed match or surpass human performance if given enough data. However, current machine learning methods often perform poorly when faced with problems that violate the i.i.d. assumption, yet seem trivial to humans. 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. Distribution shifts may also arise from simple corruptions that are common in real-world data collection pipelines [10], [107], [129], [170], [206]. An example of this is the impact of socioeconomic factors in clinics in Thailand on the accuracy of a detection system for diabetic retinopathy [19]. More dramatically, the phenomenon of "adversarial vulnerability" [249] highlights how even tiny but targeted violations of the i.i.d. assumption, generated by adding suitably chosen perturbations to images, imperceptible to humans, can lead to dangerous errors, such as confusion of traffic signs. Overall, it is fair to say that much of the current practice (of solving i.i.d. benchmark problems) and most theoretical results (about generalization in i.i.d. settings) fail to tackle the hard open challenge of generalization across problems.

To further understand how the i.i.d. assumption is problematic, let us consider a shopping example. Suppose that Alice is looking for a laptop rucksack on the Internet (i.e., a rucksack with a padded compartment for a laptop). The web shop's recommendation system suggests that she should buy a laptop to go along with the rucksack. This seems odd because she probably already has a laptop; otherwise, she would not be looking for the rucksack in the first place. In a way, the laptop is the cause, and the rucksack is an effect. Now, suppose that we are told whether a customer has bought a laptop. This reduces our uncertainty about whether she also bought a laptop rucksack, and vice versa—and it does so by the same amount (the *mutual information*), so the directionality of cause and effect is lost. However, the directionality is present in the physical mechanisms generating statistical dependence, for instance, the mechanism that makes a customer want to buy a rucksack once she owns a laptop.[5] Recommending an item to buy constitutes an intervention in a system, taking us outside the i.i.d. setting. We no longer work with the observational distribution but a distribution where certain variables or mechanisms have changed.

B. Reichenbach Principle: From Statistics to Causality

Reichenbach [198] clearly articulated the connection between causality and statistical dependence. He postulated the following:

> *Common cause principle:* If two observables X and Y are statistically dependent, then there exists a variable Z that causally influences both and explains all the dependence in the sense of making them independent when conditioned on Z.

As a special case, this variable can coincide with X or Y. Suppose that X is the frequency of storks and Y the human birth rate. If storks bring the babies, then the correct causal graph is $X \rightarrow Y$. If babies attract storks, it is $X \leftarrow Y$. If there is some other variable that causes both (such as economic development), we have $X \leftarrow Z \rightarrow Y$.

Without additional assumptions, we cannot distinguish these three cases using observational data. The class of observational distributions over X and Y that can be realized by these models is the same in all three cases. A causal model, thus, contains genuinely more information than a statistical one.

While causal structure discovery is hard if we have only two observables [190], the case of more observables is surprisingly easier, the reason being that, in that case, there are nontrivial conditional independence properties [52],

[5] Note that the physical mechanisms take place in time, and if available, time order may provide additional information about causality.

[75], [238] implied by causal structure. These generalize the Reichenbach principle and can be described by using the language of causal graphs or structural causal models (SCMs), merging probabilistic graphical models and the notion of interventions [183], [237]. They are best described using directed functional parent–child relationships rather than conditionals. While conceptually simple in hindsight, this constituted a major step in the understanding of causality.

C. Structural Causal Models

The SCM viewpoint considers a set of *observables* (or *variables*) $X_1, \ldots, X_n$ associated with the vertices of a directed acyclic graph (DAG). We assume that each observable is the result of an assignment

$$X_i := f_i(\mathbf{PA}_i, U_i) \quad (i = 1, \ldots, n) \tag{3}$$

using a deterministic function f_i depending on X_i's parents in the graph (denoted by $\mathbf{PA}_i$) and on an *unexplained* random variable U_i. Mathematically, the observables are, thus, random variables, too. Directed edges in the graph represent direct causation since the parents are connected to X_i by directed edges and, through (3), directly affect the assignment of X_i. The noise U_i ensures that the overall object (3) can represent a general conditional distribution $P(X_i|\mathbf{PA}_i)$, and the set of noises $U_1, \ldots, U_n$ is assumed to be *jointly independent*. If they were not, then, by the common cause principle, there should be another variable that causes their dependence, and thus, our model would not be *causally sufficient*.

If we specify the distributions of $U_1, \ldots, U_n$, recursive application of (3) allows us to compute the *entailed observational joint distribution* $P(X_1, \ldots, X_n)$. This distribution has structural properties inherited from the graph [147], [183]: it satisfies the *causal Markov condition* stating that conditioned on its parents, each X_j is independent of its nondescendants.

Intuitively, we can think of the independent noises as "information probes" that spread through the graph (much like independent elements of gossip can spread through a social network). Their information gets entangled, manifesting itself in a footprint of conditional dependencies, making it possible to infer aspects of the graph structure from observational data using independence testing. Like in the gossip analogy, the footprint may not be sufficiently characteristic to pin down a unique causal structure. In particular, it certainly is not if there are only two observables since any nontrivial conditional independence statement requires at least three variables. The two-variable problem can be addressed by making additional assumptions, as not only the graph topology leaves a footprint in the observational distribution, but the functions f_i do, too. This point is interesting for machine learning, where much attention is devoted to properties of function classes (e.g., priors or capacity measures), and we shall return to it below.

1) Causal Graphical Models: The graph structure along with the joint independence of the noises implies a canonical factorization of the joint distribution entailed by (3) into causal conditionals that we refer to as the *causal (or disentangled) factorization*

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \mathbf{PA}_i). \tag{4}$$

While many other *entangled factorizations* are possible, for example,

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_{i+1}, \ldots, X_n) \tag{5}$$

the factorization (4) yields practical computational advantages during inference, which is, in general, hard, even when it comes to nontrivial approximations [210]. But more interestingly, it is the only one that decomposes the joint distribution into conditionals corresponding to the structural assignments [see (3)]. We think of these as the *causal mechanisms* that are responsible for all statistical dependencies among the observables. Accordingly, in contrast to (5), the disentangled factorization represents the joint distribution as a product of causal mechanisms.

2) Latent Variables and Confounders: Variables in a causal graph may be unobserved, which can make causal inference particularly challenging. Unobserved variables may *confound* two observed variables so that they either appear statistically related while not being causally related (i.e., neither of the variables is an ancestor of the other), or their statistical relation is altered by the presence of the confounder (e.g., one variable is a causal ancestor for the other, but the confounder is a causal ancestor of both). Confounders may or may not be known or observed.

3) Interventions: The SCM language makes it straightforward to formalize *interventions* as operations that modify a subset of assignments (3), for example, changing U_i, setting f_i (and thus X_i) to a constant, or changing the functional form of f_i (and, thus, the dependence of X_i on its parents) [183], [237].

Several types of interventions may be possible [63], which can be categorized as follows.

1) *No intervention:* Only observational data are obtained from the causal model.
2) *Hard/perfect:* The function in the structural assignment [see (3)] of a variable (or, analogously, of multiple variables) is set to a constant (implying that the value of the variable is fixed), and then, the entailed distribution for the modified SCM is computed.
3) *Soft/imperfect:* The structural assignment (3) for a variable is modified by changing the function or the noise term (this corresponds to changing the conditional distribution given its parents).

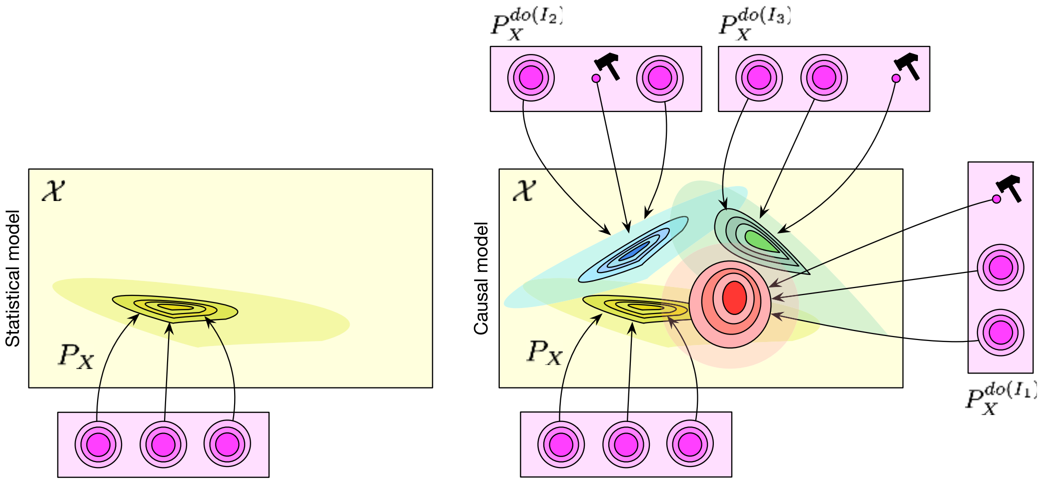

Fig. 1. ***Difference between statistical (left) and causal models (right) on a given set of three variables. While a statistical model specifies a single probability distribution, a causal model represents a set of distributions, one for each possible intervention (indicated with a 🔨).***

4) *Uncertain:* The learner is not sure which mechanism/variable is affected by the intervention.

One could argue that stating the structural assignments as in (3) is not yet sufficient to formulate a causal model. In addition, one should specify the set of possible interventions on the SCM. This may be done implicitly via the functional form of structural equations by allowing any intervention over the domain of the mechanisms. This becomes relevant when learning a causal model from data, as the SCM depends on the interventions. Pragmatically, we should aim at learning causal models that are useful for specific sets of tasks of interest [208], [266] on appropriate descriptors (in terms of which causal statements they support) that must either be provided or learned. We will return to the assumptions that allow learning causal models and features in Section IV.

D. Difference Between Statistical Models, Causal Graphical Models, and SCMs

An example of the difference between a statistical and a causal model is depicted in Fig. 1. A statistical model may be defined, for instance, through a graphical model, that is, a probability distribution along with a graph such that the former is Markovian with respect to the latter [in which case it can be factorized as (4)]. However, the edges in a (generic) graphical model do not need to be causal [98]. For instance, the two graphs $X_1 \to X_2 \to X_3$ and $X_1 \leftarrow X_2 \leftarrow X_3$ imply the same conditional independence(s) (X_1 and X_3 are independent given X_2). They are, thus, in the same Markov equivalence class, that is, if a distribution is Markovian with respect to one of the graphs, then it also is with respect to the other graph. Note that the above serves as an example that the Markov condition is not sufficient for causal discovery. Further assumptions are needed (see below and [183], [188], and [237]).

A graphical model becomes causal if the edges of its graph are causal (in which case the graph is referred to as a "causal graph") [see (3)]. This allows us to compute interventional distributions, as depicted in Fig. 1. When a variable is intervened upon, we disconnect it from its parents, fix its value, and perform ancestral sampling on its children.

An SCM is composed of: 1) a set of causal variables and 2) a set of structural equations with a distribution over the noise variables U_i (or a set of causal conditionals). While both causal graphical models and SCMs allow computing interventional distributions, only the SCMs allow computing counterfactuals. To compute counterfactuals, we need to fix the value of the noise variables. Moreover, there are many ways to represent a conditional as a structural assignment (by picking different combinations of functions and noise variables).

Causal learning and reasoning: The conceptual basis of statistical learning is a joint distribution $P(X_1, \ldots, X_n)$ (where, often, one of the X_i is a response variable denoted as Y), and we make assumptions about function classes used to approximate, say, a regression $\mathbb{E}[Y|X]$. *Causal learning* considers a richer class of assumptions and seeks to exploit the fact that the joint distribution possesses a causal factorization [see (4)]. It involves the causal conditionals $P(X_i \mid \mathbf{PA}_i)$ [e.g., represented by the functions f_i and the distribution of U_i in (3)], how these conditionals relate to each other, and interventions or changes that they admit. Once a causal model is available, either by external human knowledge or a learning process, *causal reasoning* allows drawing conclusions on the effect of interventions, counterfactuals, and potential outcomes. In contrast, statistical models only allow reasoning about the outcome of i.i.d. experiments.

IV. INDEPENDENT CAUSAL MECHANISMS

We now return to the disentangled factorization [see (4)] of the joint distribution $P(X_1, \ldots, X_n)$. This factorization according to the causal graph is always possible when U_i

is independent, but we will now consider an additional notion of independence relating the factors in (4) to one another.

Whenever we perceive an object, our brain assumes that the object and the mechanism by which the information contained in its light reaches our brain are *independent*. We can violate this by looking at the object from an accidental viewpoint, which can give rise to optical illusions [188]. The above independence assumption is useful because, in practice, it holds most of the time, and our brain, thus, relies on objects being independent of our vantage point and the illumination. Likewise, there should not be accidental coincidences, such as 3-D structures lining up in 2-D, or shadow boundaries coinciding with texture boundaries. In vision research, this is called the generic viewpoint assumption.

If we move around the object, our vantage point changes, but we assume that the other variables of the overall generative process (e.g., lighting, object position, and structure) are unaffected by that. This is an *invariance* implied by the above independence, allowing us to infer 3-D information even without stereo vision ("structure from motion").

For another example, consider a data set that consists of altitude A and average annual temperature T of weather stations [188]. A and T are correlated, which we believe is due to the fact that altitude has a causal effect on temperature. Suppose that we had two such data sets: one for Austria and one for Switzerland. The two joint distributions $P(A,T)$ may be rather different since the marginal distributions $P(A)$ over altitudes will differ. The conditionals $P(T|A)$, however, may be (close to) invariant since they characterize the physical mechanisms that generate temperature from altitude. This similarity is lost upon us if we only look at the overall joint distribution, without information about the causal structure $A \rightarrow T$. The causal factorization $P(A)P(T|A)$ will contain a component $P(T|A)$ that generalizes across countries, while the entangled factorization $P(T)P(A|T)$ will exhibit no such robustness. 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 to generalizing from an observational distribution to certain *interventional* distributions.

One can express the above insights as follows [188], [220]:

> *ICM principle: The causal generative process of a 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 entails several notions important to causality, including separate intervenability of causal variables, modularity and autonomy of subsystems, and invariance [183], [188]. If we have only two variables, it reduces to independence between the cause distribution and the mechanism producing the effect distribution.

Applied to the causal factorization [see (4)], the principle tells us that the factors should be independent in the sense that the following holds.

1) Changing (or performing an intervention upon) one mechanism $P(X_i|\mathbf{PA}_i)$ does not change any of the other mechanisms $P(X_j|\mathbf{PA}_j)$ $(i \neq j)$ [220].
2) Knowing some other mechanisms $P(X_i|\mathbf{PA}_i)$ $(i \neq j)$ does not give us information about a mechanism $P(X_j|\mathbf{PA}_j)$ [124].

This notion of independence, thus, subsumes two aspects: the former pertaining to influence and the latter to information.

The notion of invariant, autonomous, and independent mechanisms has appeared in various guises throughout the history of causality research [72], [100], [111], [124], [183], [188], [240]. Early work on this was done by Haavelmo [100], stating the assumption that changing one of the structural assignments leaves the other ones invariant. Hoover [111] attributed to Herb Simon the *invariance criterion*: the true causal order is the one that is invariant under the right sort of intervention. Aldrich [4] discussed the historical development of these ideas in economics. He argued that the "most basic question one can ask about a relation should be: how autonomous is it?" [72, preface]. Pearl [183] discussed autonomy in detail, arguing that a causal mechanism remains invariant when other mechanisms are subjected to external influences. He pointed out that causal discovery methods may best work "in longitudinal studies conducted under slightly varying conditions, where accidental independencies are destroyed and only structural independencies are preserved." Overviews are provided by Aldrich [4], Hoover [111], Pearl [183], and Peters *et al.* [188, Section 2.2]. These seemingly different notions can be unified [124], [240].

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 the implication 1) of the ICM Principle, we state the following hypothesis:

> *SMS: Small distribution changes tend to manifest themselves in a sparse or local way in the causal/disentangled factorization [see (4)], that is, they should usually not affect all factors simultaneously.*

In contrast, if we consider a noncausal factorization, for example, (5), then many, if not all, terms will be affected simultaneously as we change one of the physical mechanisms responsible for a system's statistical dependencies. Such a factorization may, thus, be called *entangled*, a term that has gained popularity in machine learning [24], [110], [158], [247].

The SMS hypothesis was stated in [25], [114], [180], and [217] and in earlier form in [219], [220], and [281]. An intellectual ancestor is Simon's invariance criterion, that is, that the causal structure remains invariant across changing background conditions [235]. The hypothesis is also related to ideas of looking for features that vary slowly [70], [270]. It has recently been used for learning causal models [131], modular architectures [29], [85], and disentangled representations [159].

We have informally talked about the dependence of two mechanisms $P(X_i|\mathbf{PA}_i)$ and $P(X_j|\mathbf{PA}_j)$ when discussing the ICM principle and the disentangled factorization [see (4)]. Note that the dependence of two such mechanisms 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 mechanisms are independent. Also, the independence of the noise terms U_i does not translate into the independence of the X_i. Intuitively speaking, the independent noise terms U_i provide and parameterize the uncertainty contained in the fact that a mechanism $P(X_i|\mathbf{PA}_i)$ is nondeterministic[6] and, thus, ensure that each mechanism adds an independent element of uncertainty. In this sense, the ICM principle contains the independence of the unexplained noise terms in an SCM [see (3)] as a special case.

In the ICM principle, we have stated that independence of two mechanisms (formalized as conditional distributions) should mean that the two conditional distributions do not *inform* or *influence* each other. The latter can be thought of as requiring that independent interventions are possible. To better understand the former, we next discuss a formalization in terms of *algorithmic independence*. In a nutshell, we encode each mechanism as a bit string and require that joint compression of these strings does not save space relative to independent compressions.

To this end, first recall that we have, so far, discussed links between causal and statistical structures. Of the two, the more fundamental one is the causal structure since it captures the physical mechanisms that generate statistical dependencies 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 statistical information contained in a mechanism since deterministic functions in the generic case neither generate nor destroy information. This serves as a motivation to devise an alternative model of causal structures in terms of the Kolmogorov complexity [124]. 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, that is, two conditionals are considered independent if knowing (the shortest compression of) one does not help us achieve a shorter compression of the other.

The algorithmic information theory provides a natural framework for nonstatistical graphical models [120], [124]. Just like that the latter is obtained from SCMs by making the unexplained variables U_i random, we obtain algorithmic graphical models by making the U_i bit strings, jointly independent across nodes, and viewing X_i as the output of a fixed Turing machine running the program U_i on the 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 an additive decomposition of the joint Kolmogorov complexity in analogy to (4), and prove that they are implied by the SCM [124]. Interestingly, in this case, independence of noises and independence of mechanisms coincide since the independent programs play the role of the unexplained noise terms. This approach shows that causality is not intrinsically bound to statistics.

V. CAUSAL DISCOVERY AND MACHINE LEARNING

Let us turn to the problem of causal discovery from data. Subject to suitable assumptions, such as *faithfulness* [237], one can sometimes recover aspects of the underlying graph[7] from observational data by performing conditional independence tests. However, there are several problems with this approach. One is that our data sets are always finite in practice, and conditional independence testing is a notoriously difficult problem, especially if conditioning sets are continuous and multidimensional. Thus, while, in principle, the conditional independencies implied by the causal Markov condition hold irrespective of the complexity of the functions appearing in an SCM, for finite data sets, conditional independence testing is hard without additional assumptions [225]. Recent progress in (conditional) independence testing heavily relies on kernel function classes to represent probability distributions in reproducing kernel Hilbert spaces [43], [61], [74], [91], [92], [193], [280]. The other problem is that, in the case of only two variables, the ternary concept of conditional independence collapses and the Markov condition, thus, has no nontrivial implications.

It turns out that both problems can be addressed by making assumptions on function classes. This is typical for machine learning, where it is well known that finite-sample generalization without assumptions on function classes is impossible. Specifically, although there are universally consistent learning algorithms, that is, approaching minimal expected error in the infinite sample limit, there are always cases where this convergence is arbitrarily slow. Thus, for given sample size, it will depend on the problem being learned whether we achieve low expected error, and the statistical learning theory provides probabilistic guarantees

[6] In the sense that the mapping from $\mathbf{PA}_i$ to X_i is described by a nontrivial conditional distribution, rather than by a function.

[7] One can recover the causal structure up to a *Markov equivalence class*, where DAGs have the same undirected skeleton and "immoralities" ($X_i \rightarrow X_j \leftarrow X_k$).

in terms of measures of complexity of function classes [56], [257].

Returning to causality, we provide an intuition why assumptions on the functions in an SCM should be necessary to learn about them from data. Consider a toy SCM with only two observables $X \rightarrow Y$. In this case, (3) turns into

$$X = U \tag{6}$$
$$Y = f(X, V) \tag{7}$$

with $U \perp\!\!\!\perp V$. Now, think of V acting as a random selector variable choosing from among a set of functions $\mathcal{F} = \{f_v(x) \equiv f(x, v) \mid v \in \text{supp}(V)\}$. If $f(x, v)$ depends on v in a nonsmooth way, it should be hard to glean information about the SCM from a finite data set, given that V is not observed and its value randomly selects among arbitrarily different f_v.

This motivates restricting the complexity with which f depends on V. A natural restriction is to assume an additive noise model

$$X = U \tag{8}$$
$$Y = f(X) + V. \tag{9}$$

If f in (7) depends smoothly on V, and if V is relatively well concentrated, this can be motivated by a local Taylor expansion argument. It drastically reduces the effective size of the function class—without such assumptions, the latter could depend exponentially on the cardinality of the support of V. Restrictions of function classes not only make it easier to learn functions from data but it turns out that they 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) [18], [113], [139], [175], [190] (see also [246]). This is subject to certain genericity assumptions, and notable exceptions include the case where U and V are Gaussian and f is linear. It generalizes results of Shimizu *et al.* [229] for linear functions, and it can be generalized to include nonlinear rescalings [279], loops [174], confounders [123], and multivariable settings [189]. Empirically, there is a number of methods that can detect causal direction better than chance [176], some of the building on the above Kolmogorov complexity model [37], some on generative models [83], and some directly learning to classify bivariate distributions into causal versus anticausal [161].

While restrictions of function classes are one possibility to allow identifying the causal structure, other assumptions or scenarios are possible. So far, we have discussed that causal models are expected to generalize under certain distribution shifts since they explicitly model interventions. By the SMS hypothesis, much of the causal structure is assumed to remain invariant. Hence, distribution shifts, such as observing a system in different "environments/contexts," can significantly help to identify causal structure [188], [251]. These contexts can come from interventions [187], [191], [220], nonstationary time series [101], [116], [192], or multiple views [90], [114]. The contexts can likewise be interpreted as different tasks, which provides a connection to metalearning [23], [68], [213].

The work of Bengio *et al.* [25] ties the generalization in metalearning to invariance properties of causal models, using the idea that a causal model should adapt faster to interventions than purely predictive models. This was extended to multiple variables and unknown interventions in [131], proposing a framework for causal discovery using neural networks by turning the discrete graph search into a continuous optimization problem. While Bengio *et al.* [25] and Ke *et al.* [131] focused on learning a causal model using neural networks with an unsupervised loss, the work of Dasgupta *et al.* [51] explores learning a causal model using an RL agent. These approaches have in common that semantically meaningful abstract representations are given and do not need to be learned from high-dimensional and low-level (e.g., pixel) data.

VI. LEARNING CAUSAL VARIABLES

Traditional causal discovery and reasoning assume that the units are random variables connected by a causal graph. However, real-world observations are usually not structured into those units, to begin with, for example, objects in images [162]. Hence, 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 [34]). To this end, we could try to connect causal variables $S_1, \ldots, S_n$ to observations

$$X = G(S_1, \ldots, S_n) \tag{10}$$

where G is a nonlinear function. An example can be seen in Fig. 2, where high-dimensional observations are the result of a view on the state of a causal system that is then processed by a neural network to extract high-level variables that are useful on a variety of tasks. Although causal models in economics, medicine, or psychology often use variables that are abstractions of underlying quantities, it is challenging to state general conditions under which coarse-grained variables admit causal models with well-defined interventions [42], [208]. Defining objects or variables that can be causally related amounts to coarse-graining of more detailed models of the world, including microscopic structural equation models [208], ordinary differential equations [173], [207], and temporally aggregated time series [79]. The task of identifying suitable units that admit causal models is challenging for both human and machine intelligence. Still, it aligns with

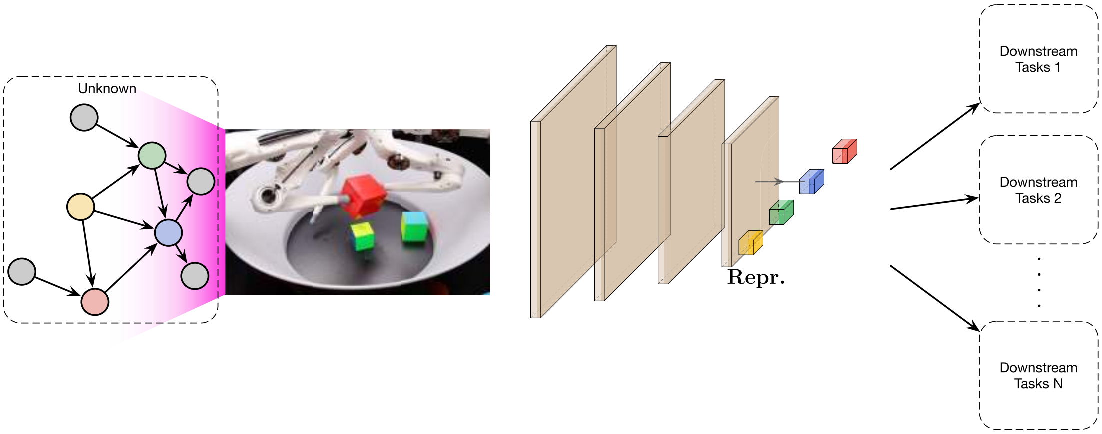

Fig. 2. ***Illustration of the causal representation learning problem setting. Perceptual data, such as images or other high-dimensional sensor measurements, can be thought of as entangled views of the state of an unknown causal system, as described in (10). With the exception of possible task labels, none of the variables describing the causal variables generating the system may be known. The goal of causal representation learning is to learn a representation (partially) exposing this unknown causal structure (e.g., which variables describe the system, and their relations). As full recovery may often be unreasonable, neural networks may map the low-level features to some high-level variables supporting causal statements relevant to a set of downstream tasks of interest. For example, if the task is to detect the manipulable objects in a scene, the representation may separate intrinsic object properties from their pose and appearance to achieve robustness to distribution shifts on the latter variables. Usually, we do not get labels for the high-level variables, but the properties of causal models can serve as useful inductive biases for learning (e.g., the SMS hypothesis).***

the general goal of modern machine learning to learn meaningful representations of data, where meaningful can include *robust*, *explainable*, or *fair* [130], [134], [142], [259], [275].

To combine structural causal modeling [see (3)] and representation learning, we should strive to embed an SCM into larger machine learning models whose inputs and outputs may be high-dimensional and unstructured, but whose inner workings are at least partly governed by an SCM (that can be parameterized with a neural network). The result may be a modular architecture, where the different modules can be individually fine-tuned and repurposed for new tasks [85], [180], and the SMS hypothesis can be used to enforce the appropriate structure. We visualize an example in Fig. 3 where changes are sparse for the appropriate causal variables (the position of the finger and the cube changed as a result of moving the finger) but dense in other representations, for example, in the pixel space (as finger and cube move, many pixels change their value). At the extreme, all pixels may change as a result of a sparse intervention, for example, if the camera view or the lighting changes.

We now discuss three problems of modern machine learning in the light of causal representation learning.

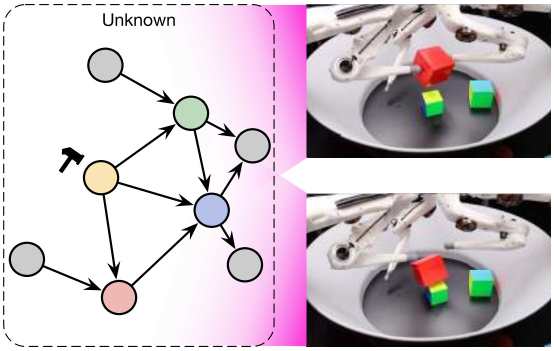

Fig. 3. ***Example of the SMS hypothesis where an intervention (which may or may not be intentional/observed) changes the position of one finger (↖), and as a consequence, the object falls. The change in pixel space is entangled (or distributed), in contrast to the change in the causal model.***

A. Problem 1—Learning Disentangled Representations

We have earlier discussed the ICM principle implying both the independence of the SCM noise terms in (3) and, thus, the feasibility of the disentangled representation

$$P(S_1, \ldots, S_n) = \prod_{i=1}^{n} P(S_i \mid \mathbf{PA}_i) \tag{11}$$

as well as the property that the conditionals $P(S_i \mid \mathbf{PA}_i)$ are independently manipulable and largely invariant across related problems. Suppose that we seek to reconstruct such a *disentangled representation using independent mechanisms* [see (11)] from data, but the causal variables S_i are not provided to us *a priori*. Rather, we are given (possibly high-dimensional) $X = (X_1, \ldots, X_d)$ (in the following, we think of X as an image with pixels $X_1, \ldots, X_d$), as in (10), from which we should construct causal variables $S_1, \ldots, S_n$ $(n \ll d)$ as well as mechanisms [see (3)]

$$S_i := f_i(\mathbf{PA}_i, U_i) \quad (i = 1, \ldots, n) \tag{12}$$

modeling the causal relationships among S_i. To this end, as a first step, we can use an *encoder* $q : \mathbb{R}^d \to \mathbb{R}^n$ taking X to a latent "bottleneck" representation comprising the unexplained noise variables $U = (U_1, \ldots, U_n)$. The next step is the mapping $f(U)$ determined by the structural assignments $f_1, \ldots, f_n$. Finally, we apply a *decoder* $p : \mathbb{R}^n \to \mathbb{R}^d$. For suitable n, the system can be trained using reconstruction error to satisfy $p \circ f \circ q \approx id$ on the observed images. If the causal graph is known, the topology of a neural network implementing f can be fixed accordingly; if not, the neural network decoder learns the composition $\tilde{p} = p \circ f$. In practice, one may not know f and, thus, only learn an autoencoder $\tilde{p} \circ q$, where the causal graph effectively becomes an unspecified part of the decoder $\tilde{p}$, possibly aided by a suitable choice of architecture [149].

Much of the existing work on disentanglement [62], [110], [135], [157]–[159], [202], [256] focuses on independent factors of variation. This can be viewed as the special case where the causal graph is trivial, that is, $\forall i : \mathbf{PA}_i = \emptyset$ in (12). In this case, the factors are functions of the independent exogenous noise variables and, thus, independent themselves.[8] However, the ICM principle is more general and contains statistical independence as a special case.

Note that the problem of *object-centric* representation learning [11], [40], [84], [87], [88], [138], [155], [160], [255], [262] can also be considered a special case of disentangled factorization as discussed here. Objects are constituents of scenes that in principle permit separate interventions. A disentangled representation of a scene containing objects should probably use objects as some of the building blocks of an overall causal factorization,[9] complemented by mechanisms, such as orientation, viewing direction, and lighting.

The problem of recovering the exogenous noise variables is ill-defined in the i.i.d. case as there are infinitely many equivalent solutions yielding the same observational distribution [117], [158], [188]. Additional assumptions or biases can help favoring certain solutions over others [158], [205]. Leeb *et al.* [149] propose a structured decoder that embeds an SCM and automatically learns a hierarchy of disentangled factors.

To make (12) causal, we can use the ICM principle, that is, we should make U_i statistically independent, and we should make the mechanisms independent. This could be done by ensuring that they are invariant across problems, exhibit sparse changes to actions or that they can be independently intervened upon [22], [30], [217]. Locatello *et al.* [159] showed that the SMS hypothesis stated above is theoretically sufficient when given suitable training data. Furthermore, the SMS hypothesis can be used as supervision signal, in practice, even if $\mathbf{PA}_i \neq \emptyset$ [252]. However, which factors of variation can be disentangled depend on which interventions can be observed [159], [230]. As discussed by Schölkopf *et al.* [219] and Shu *et al.* [230], different supervision signals may be used to identify subsets of factors. Similarly, when learning causal variables from data, which variables can be extracted and their granularity depends on which distribution shifts, explicit interventions, and other supervision signals are available.

B. Problem 2—Learning Transferable Mechanisms

An artificial or natural agent in a complex world is faced with limited resources. This concerns training data, that is, we only have limited data for each 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 size of their brains, and evolutionary neuroscience knows many examples where brain regions get repurposed. Similar constraints on size and energy apply as ML methods get embedded in (small) physical devices that may be battery-powered. Future AI models that robustly solve a range of problems in the real world will, thus, likely need to reuse components, which requires them to be robust across tasks and environments [219]. An elegant way to do this is to employ a modular structure that mirrors corresponding modularity in the world. In other words, if the world is indeed modular, in the sense that components/mechanisms of the world play roles across a range of environments, tasks, and settings, then it would be prudent for a model to employ corresponding modules [85]. For instance, if variations of natural lighting (the position of the sun, clouds, and so on) 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 face recognizers, say, for every lighting condition. If, for example, our nervous system were to compensate for the lighting changes by a gain control mechanism, then this mechanism in itself need not have anything to do with the physical mechanisms bringing about brightness differences. However, it would play a role in a modular structure that corresponds to the role that the physical mechanisms play in the world's modular structure. This could produce a bias toward models that exhibit certain forms of structural homomorphism to a world that we cannot directly recognize, which would be rather intriguing, given that ultimately our brains do nothing but turn neuronal signals into other neuronal

[8] For an example to see why this is often not desirable, note that the presence of fork and knife may be statistically dependent, yet we might want a disentangled representation to represent them as separate entities.

[9] Objects can be represented at different levels of granularity [208], that is, as a single entity or as a composition of other causal variables encoding parts, properties, and other factors of variation.

signals. A sensible inductive bias to learn such models is to look for ICMs [182], and competitive training can play a role in this. For pattern recognition tasks, Parascandolo *et al.* [180] and Goyal *et al.* [85] suggested that learning causal models that contain independent mechanisms may help in transferring modules across substantially different domains.

C. Problem 3—Learning Interventional World Models and Reasoning

Deep learning excels at learning representations of data that preserve relevant statistical properties [24], [148]. However, it does so without taking into account the causal properties of the variables, that is, it does not care about the interventional properties of the variables that it analyzes or reconstructs. Causal representation learning should move beyond the representation of statistical dependence structures toward models that support intervention, planning, and reasoning, realizing Konrad Lorenz' notion of *thinking* as *acting in an imagined space* [163]. This ultimately requires the ability to reflect back on one's actions and envision alternative scenarios, possibly necessitating (the illusion of) free will [184]. The biological function of self-consciousness may be related to the need for a variable representing oneself in one's Lorenzian *imagined space*, and free will may then be a means to communicate about actions taken by that variable, crucial for social and cultural learning, a topic that has not yet entered the stage of machine learning research although it is at the core of human intelligence [108].

VII. IMPLICATIONS FOR MACHINE LEARNING

All these discussions call for a learning paradigm that does not rest on the usual i.i.d. assumption. Instead, we wish to make a weaker assumption that the data on which the model will be applied comes from a possibly different distribution but involving (mostly) the same causal mechanisms [188]. This raises serious challenges: 1) in many cases, we need to infer abstract causal variables from the available low-level input features; 2) there is no consensus on which aspects of the data reveal causal relations; 3) the usual experimental protocol of training and test set may not be sufficient for inferring and evaluating causal relations on existing data sets, and we may need to create new benchmarks, for example, with access to environmental information and interventions; 4) even in the limited cases that we understand, we often lack scalable and numerically sound algorithms. Despite these challenges, we argue that this endeavor has concrete implications for machine learning and may shed light on desiderata and current practices alike.

A. Semisupervised Learning

Suppose that our underlying causal graph is $X \rightarrow Y$, and at the same time, we are trying to learn a mapping $X \rightarrow Y$. The causal factorization (4) for this case is

$$P(X,Y) = P(X)P(Y|X). \tag{13}$$

The ICM principle posits that the modules in a joint distribution's causal decomposition 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 SSL should be futile, in as far as it is using additional information about $P(X)$ (from unlabelled data) to improve our estimate of $P(Y|X=x)$.

In the opposite (*anticausal*) direction (i.e., the direction of prediction is opposite to the causal generative process), however, SSL may be possible. To see this, we refer to Daniušis *et al.* [50] who define a measure of dependence between input $P(X)$ and conditional $P(Y|X)$.[10] Assuming that this measure is zero in the causal direction (applying the ICM assumption described in Section IV to the two-variable case), they show that it is strictly positive in the anticausal direction. Applied to SSL in the anticausal direction, this implies that the distribution of the input (now: effect) variable should contain information about the conditional output (cause) given input, that is, the quantity that machine learning is usually concerned with.

The study [220] empirically corroborated these predictions, thus establishing an intriguing bridge between the *structure* of learning problems and certain *physical* properties (cause–effect direction) of real-world data generating processes. It also led to a range of follow-up work [32], [78], [97], [114], [115], [152], [153], [156], [167], [195], [204], [243], [263], [267], [277], [278], [281], complementing the studies of Bareinboim and Pearl [14], [185], and it inspired a thread of work in the statistics community exploiting invariance for causal discovery and other tasks [105], [106], [114], [187], [191].

On the SSL side, subsequent developments include further theoretical analyses [125], [188, Section 5.1.2] and a form of conditional SSL [261]. The view of SSL as exploiting dependencies between a marginal $P(X)$ and a noncausal conditional $P(Y|X)$ is consistent with the common assumptions employed to justify SSL [45]. The *cluster assumption* asserts that the labeling function [which is a property of $P(Y|X)$] should not change within clusters of $P(X)$. The *low-density separation assumption* posits that the area where $P(Y|X)$ takes the value of 0.5 should have small $P(X)$; the *semisupervised smoothness assumption*, applicable also to continuous outputs, states that if two points in a high-density region are close and so should be the corresponding output values. Note, moreover, that some of the theoretical results in the field use assumptions well-known from causal graphs (even if they do not mention causality): the *cotraining theorem* [33] makes a statement about learnability from unlabelled data and

[10] Other dependence measures have been proposed for high-dimensional linear settings and time series [28], [119], [121], [122], [126], [226].

relies on an assumption of predictors being conditionally independent given the label, which we would normally expect if the predictors are (only) caused by the label, that is, an anticausal setting. This is nicely consistent with the above findings.

B. Adversarial Vulnerability

One can hypothesize that the causal direction should also have an influence on whether classifiers are vulnerable to *adversarial attacks*. These attacks have recently become popular and consist of minute changes to inputs, invisible to a human observer yet changing a classifier's output [249]. This is related to causality in several ways. First, these attacks clearly constitute violations of the i.i.d. assumption that underlies statistical machine learning. If all we want to do is a prediction in an i.i.d. setting, then statistical learning is fine. In the adversarial setting, however, the modified test examples are not drawn from the same distribution as the training examples. The adversarial phenomenon also shows that the kind of robustness current classifiers exhibit is rather different from the one a human exhibits. If we knew both robustness measures, we could try to maximize one, while minimizing the other. Current methods can be viewed as crude approximations to this, effectively modeling the human's robustness as a mathematically simple set, say, an l_p ball of radius $\epsilon > 0$: they, often, try to find examples that lead to maximal changes in the classifier's output, subject to the constraint that they lie in an l_p ball in the pixel metric. As we think of a classifier as the approximation of a function, the large gradients exploited by these attacks are either property of this function or a defect of the approximation.

There are different ways of relating this to causal models. As described in [188, Section 1.4], different causal models can generate the same statistical pattern recognition model. In one of those, we might provide a writer with a sequence of class labels y, with the instruction to produce a set of corresponding images x. It is clear that intervening on y will impact x, but intervening on x will not impact y, so this is an anticausal learning problem. In another setting, we might ask the writer to decide for herself which digits to write and to record the labels alongside the digit (in this case, the classifier would try to predict one effect from another one, a situation that we might call a confounded one). In the last one, we might provide images to a person and ask the person to generate labels by classifying them.

Let us now assume that we are in the *causal* setting where the causal generative model factorizes into independent components, one of which is (essentially) the classification function. As discussed in Section III, when specifying a causal model, one needs to determine which interventions are allowed, and a structural assignment will then, by definition, be valid under every possible (allowed) intervention. One may, thus, expect that if the predictor approximates the causal mechanism that is inherently transferable and robust, adversarial examples should be harder to find [133], [216].[11] Recent work supports this view: it was shown that a possible defense against adversarial attacks is to solve the anticausal classification problem by modeling the causal generative direction, a method that, in vision, is referred to as *analysis by synthesis* [222]. A related defense method proceeds by reconstructing the input using an autoencoder before feeding it to a classifier [96].

C. Robustness and Strong Generalization

We can speculate that structures composed of autonomous modules, such as given by a causal factorization [see (4)], should be relatively robust to swapping out or modifying individual components. Robustness should also play a role when studying *strategic behavior*, that is, decisions or actions that take into account the actions of other agents (including AI agents). Consider a system that tries to predict the probability of successfully paying back a credit, based on a set of features. The set could include, for instance, the current debt of a person, as well as their address. To get a higher credit score, people could, thus, change their current debt (by paying it off), or they could change their address by moving to a more affluent neighborhood. The former probably has a positive causal impact on the probability of paying back; for the latter, this is less likely. Thus, we could build a scoring system that is more robust with respect to such strategic behavior by only using causal features as inputs [132].

To formalize this general intuition, one can consider a form of out-of-distribution generalization, which can be optimized by minimizing the empirical risk over a class of distributions induced by a causal model of the data [5], [169], [187], [204], [220]. To describe this notion, we start by recalling the usual empirical risk minimization setup. We have access to data from a distribution $P(X, Y)$ and train a predictor g in a hypothesis space $\mathcal{H}$ (e.g., a neural network with a certain architecture predicting Y from X) to minimize the empirical risk $\hat{R}$:

$$g^\star = \operatorname*{argmin}_{g \in \mathcal{H}} \hat{R}_{P(X,Y)}(g) \tag{14}$$

where

$$\hat{R}_{P(X,Y)}(g) = \hat{\mathbb{E}}_{P(X,Y)}\left[\text{loss}(Y, g(X))\right]. \tag{15}$$

Here, we denote by $\hat{\mathbb{E}}_{P(X,Y)}$ the empirical mean computed from a sample drawn from $P(X, Y)$. When we refer to "out-of-distribution generalization," we mean having a

[11] Adversarial attacks may still exploit the quality of the (parameterized) approximation of a structural equation.

small expected risk for a different distribution $P^\dagger(X,Y)$:

$$R^{OOD}_{P^\dagger(X,Y)}(g) = \mathbb{E}_{P^\dagger(X,Y)}\left[\text{loss}(Y, g(X))\right]. \tag{16}$$

It is clear that the gap between $\hat{R}_{P(X,Y)}(g)$ and $R^{OOD}_{P^\dagger(X,Y)}(g)$ will depend on how different the test distribution $P^\dagger$ is from the training distribution P. To quantify this difference, we call *environments* the collection of different circumstances that give rise to the distribution shifts, such as locations, times, and experimental conditions. Environments can be modeled in a causal factorization [see (4)] as they can be seen as interventions on one or several causal variables or mechanisms. As a motivating example, one environment may correspond to *where* a measurement is taken (e.g., a certain room), and from each environment, we obtain a collection of measurements (images of objects in the same room). It is nontrivial (and, in some cases, provably hard [21]) to learn statistical models that are stable across training environments and generalize to novel testing environments [2], [5], [167], [187], [204] drawn from the same environment distribution.

Using causal language, one could restrict $P^\dagger(X,Y)$ to be the result of a certain set of interventions, that is, $P^\dagger(X,Y) \in \mathbb{P}_{\mathcal{G}}$, where $\mathbb{P}_{\mathcal{G}}$ is a set of interventional distributions over a causal graph $\mathcal{G}$. The worst case out-of-distribution risk then becomes

$$R^{OOD}_{\mathbb{P}_{\mathcal{G}}}(g) = \max_{P^\dagger \in \mathbb{P}_{\mathcal{G}}} \mathbb{E}_{P^\dagger(X,Y)}\left[\text{loss}(Y, g(X))\right]. \tag{17}$$

To learn a robust predictor, we should have available a subset of environment distributions $\mathcal{E} \subset \mathbb{P}_{\mathcal{G}}$ and solve

$$g^\star = \operatorname*{argmin}_{g \in \mathcal{H}} \max_{P^\dagger \in \mathcal{E}} \hat{\mathbb{E}}_{P^\dagger(X,Y)}\left[\text{loss}(Y, g(X))\right]. \tag{18}$$

In practice, solving (18) requires specifying a causal model with an associated set of interventions. If the set of observed environments $\mathcal{E}$ does not coincide with the set of possible environments $\mathbb{P}_{\mathcal{G}}$, we have an additional estimation error that may be arbitrarily large in the worst case [5], [21].

D. Pretraining, Data Augmentation, and Self-Supervision

Learning predictive models solving the min–max optimization problem of (18) is challenging. We now interpret several common techniques in machine learning as means of approximating (18).

The first approach is enriching the distribution of the training set. This does not mean obtaining more examples from $P(X,Y)$ but training on a richer data set [54], [245], for example, through pretraining on a huge and diverse corpus [36], [46], [55], [60], [112], [137], [196], [253]. Since this strategy is based on standard empirical risk minimization, it can achieve stronger generalization in practice only if the new training distribution is sufficiently diverse to contain information about other distributions in $\mathbb{P}_{\mathcal{G}}$.

The second approach, often coupled with the previous one, is to rely on data augmentation to increase the diversity of the data by "augmenting" it through a certain type of artificially generated interventions [10], [140], [234]. For the visual domain, common augmentations include performing transformations, such as rotating the image, translating the image by a few pixels, or flipping the image horizontally. The high-level idea behind data augmentation is to encourage a system to learn underlying invariances or symmetries present in the augmented data distribution. For example, in a classification task, translating the image by a few pixels does not change the class label. One may view it as specifying a set of interventions $\mathcal{E}$ that the model should be robust to (e.g., random crops/interpolations/translation/rotations). Instead of computing the maximum over all distributions in $\mathcal{E}$, one can relax the problem by sampling from the interventional distributions and optimize an expectation over the different augmented images on a suitably chosen subset [39], using a search algorithm, such as RL [49] or an algorithm based on density matching [154].

The third approach is to rely on self-supervision to learn about $P(X)$. Certain pretraining methods [36], [46], [55], [112], [196], [253] have shown that it is possible to achieve good results using only very few class labels by first pretraining on a large unlabeled data set and then fine-tuning on few labeled examples. Similarly, pretraining on large unlabeled image data sets can improve performance by learning representations that can efficiently transfer to a downstream task, as demonstrated by Bachman *et al.* [8], Chen *et al.* [47], Grill *et al.* [93], He *et al.* [102], and Oord *et al.* [179]. These methods fall under the umbrella of self-supervised learning, a family of techniques for converting an unsupervised learning problem into a supervised one by using the so-called pretext tasks with artificially generated labels without human annotations. The basic idea behind using pretext tasks is to force the learner to learn representations that contain information about $P(X)$ that may be useful for (an unknown) downstream task. Much of the work on methods that use self-supervision relies on carefully constructing pretext tasks. A central challenge here is to extract features that are indeed informative about the data-generating distribution. Ideas from the ICM principle could help develop methods that can automate the process of constructing pretext tasks. Finally, one can explicitly optimize (18), for example, through adversarial training [80]. In that case, $\mathbb{P}_{\mathcal{G}}$ would contain a set of attacks that an adversary might perform, while, presently, we consider a set of natural interventions.

An interesting research direction is the combination of all these techniques, large-scale training, data augmentation, self-supervision, and robust fine-tuning on the available data from multiple, potentially simulated environments.

E. Reinforcement Learning

RL is closer to causality research than the machine learning mainstream in which it sometimes effectively directly estimates do-probabilities. For example, on-policy learning estimates do-probabilities for the interventions specified by the policy (note that these may not be hard interventions if the policy depends on other variables). However, as soon as off-policy learning is considered, in particular, in the batch (or observational) setting [146], issues of causality become subtle [82], [165]. An emerging line of work devoted to the intersection of RL and causality includes [1], [13], [22], [38], [51], [165], [276]. Causal learning applied to RL can be divided into two aspects: causal induction and causal inference. *Causal induction (discovery)* involves learning causal relations from data, for example, an RL agent learning a causal model of the environment. *Causal inference* learns to plan and act based on a causal model. Causal induction in an RL setting poses different challenges than the classic causal learning settings where the causal variables are often given. However, there is accumulating evidence supporting the usefulness of an appropriate structured representation of the environment [2], [27], [258].

1) World Models: Model-based RL [68], [248] is related to causality as it aims at modeling the effect of actions (interventions) on the current state of the world. Particularly relevant for causal leaning are generative world models that capture some of the causal relations underlying the environment and serve as Lorenzian imagined spaces (see INTRODUCTION above) to train RL agents [48], [99], [127], [178], [214], [231], [248], [268], [271]. Structured generative approaches further aim at decomposing an environment into multiple entities with causally correct relations among them, modulo the completeness of the variables, and confounding [15], [44], [59], [136], [264], [265]. However, many of the current approaches (regardless of structure), only build partial models of the environment [89]. Since they do not observe the environment at every time step, the environment may become an unobserved confounder affecting both the agent's actions and the reward. To address this issue, a model can use the backdoor criterion conditioning on its policy [200].

2) Generalization, Robustness, and Fast Transfer: While RL has already achieved impressive results, the sample complexity required to achieve consistently good performance is often prohibitively high. Furthermore, RL agents are often brittle (if data is limited) in the face of even tiny changes to the environment (either visual or mechanistic changes) unseen in the training phase. The question of generalization in RL is essential to the field's future both in theory and practice. One proposed solution toward the goal of designing machines that can extrapolate experience across environments and tasks is to learn invariances in a causal graph structure. A key requirement to learn invariances from data may be the possibility to perform and learn from interventions. Work in developmental psychology argues that there is a need to experiment in order to discover causal relationships [81]. This can be modeled as an RL environment, where the agent can discover causal factors through interventions and observing their effects. Furthermore, causal models may allow modeling the environment as a set of underlying ICMs such that, if there is a change in distribution, not all the mechanisms need to be relearned. However, there are still open questions about the right way to think about generalization in RL, the right way to formalize the problem, and the most relevant tasks.

3) Counterfactuals: Counterfactual reasoning has been found to improve the data efficiency of RL algorithms [38], [164] and improve performance [51], and it has been applied to communicate about past experiences in the multiagent setting [69], [241]. These findings are consistent with work in cognitive psychology [65], arguing that counterfactuals allow to reason about the usefulness of past actions and transfer these insights to corresponding behavioral intentions in future scenarios [145], [199], [203].

We argue that future work in RL should consider counterfactual reasoning as a critical component to enable acting in imagined spaces and formulating hypotheses that can be subsequently tested with suitably chosen interventions.

4) Off-Line RL: The success of deep learning methods in the case of supervised learning can be largely attributed to the availability of large data sets and methods that can scale to large amounts of data. In the case of RL, collecting large amounts of high-fidelity diverse data from scratch can be expensive and, hence, becomes a bottleneck. Off-line RL [73], [150] tries to address this concern by learning a policy from a *fixed* data set of trajectories, without requiring any experimental or interventional data (i.e., without any interaction with the environment). The effective use of observational data (or logged data) may make real-world RL more practical by incorporating diverse prior experiences. To succeed at it, an agent should be able to infer the consequence of different sets of actions compared to those seen during training (i.e., the actions in the logged data), which essentially makes it a counterfactual inference problem. The distribution mismatch between the current policy and the policy that was used to collect off-line data makes off-line RL challenging as this requires us to move well beyond the assumption of independently and identically distributed data. Incorporating invariances by factorizing knowledge in terms of ICMs can help make progress toward the off-line RL setting.

F. Scientific Applications

A fundamental question in the application of machine learning in natural sciences is to which extent we can complement our understanding of a physical system with machine learning. One interesting aspect is physics simulation with neural networks [94], which can substantially increase the efficiency of hand-engineered simulators [104], [143], [211], [265], [269]. Significant out-of-distribution generalization of learned physical simulators may not be necessary if experimental conditions are carefully controlled although the simulator has to be completely retrained if the conditions change.

On the other hand, the lack of systematic experimental conditions may become problematic in other applications, such as health care. One example is personalized medicine, where we may wish to build a model of a patient health state through a multitude of data sources, such as electronic health records and genetic information [66], [109]. However, if we train a clinical system on doctors' actions in controlled settings, the system will likely provide little additional insight compared to the doctors' knowledge and may fail in surprising ways when deployed [19]. While it may be useful to automate certain decisions, an understanding of causality may be necessary to recommend treatment options that are personalized and reliable [3], [6], [31], [164], [201], [224], [242], [273].

Causality also has significant potential in helping understand medical phenomena, for example, in the current COVID-19 pandemic, where causal mediation analysis helps disentangle different effects contributing toward case fatality rates when a textbook example of Simpson's paradox was observed [260].

Another example of a scientific application is in astronomy, where causal models were used to identify exoplanets under the confounding of the instrument. Exoplanets are often detected as they partially occlude their host star when they transit in front of it, causing a slight decrease in brightness. Shared patterns in measurement noise across stars light-years apart can be removed in order to reduce the instrument's influence on the measurement [218], which is critical especially in the context of partial technical failures as experienced in the Kepler exoplanet search mission. The application of [218] leads to the discovery of 36 planet candidates [71], of which 21 were subsequently validated as bona fide exoplanets [172]. 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, that is, allowing for liquid water [26], [254]. This planet turned out to be one that had first been detected in [71, exoplanet candidate EPIC 201912552].

G. Multitask Learning and Continual Learning

State-of-the-art AI is relatively *narrow*, that is, trained to perform specific tasks, as opposed to the *broad*, versatile intelligence allowing humans to adapt to a wide range of environments and develop a rich set of skills. The human ability to discover robust, invariant high-level concepts and abstractions and to identify causal relationships from observations appears to be one of the key factors allowing for a successful generalization from prior experiences to new, often quite different, "out-of-distribution" settings.

Multitask learning refers to building a system that can solve multiple tasks across different environments [41], [209]. These tasks usually share some common traits. By learning similarities across tasks, a system could utilize the knowledge acquired from previous tasks more efficiently when encountering a new task. One possibility of learning such similarities across tasks is to learn a shared underlying data-generating process as a causal generative model whose components satisfy the SMS hypothesis [219]. In certain cases, causal models adapt faster to sparse interventions in distribution [131], [194].

At the same time, we have clearly come a long way already without explicitly treating the multitask problem as a causal one. Fuelled by abundant data and compute, AI has made remarkable advances in a wide range of applications, from image processing and natural language processing [36] to beating human world champions in games, such as chess, poker, and Go [223], improving medical diagnoses [166], and generating music [57]. A critical question thus arises: *why cannot we just train a huge model that learns environments' dynamics (e.g., in an RL setting) including all possible interventions? After all, distributed representations can generalize to unseen examples, and if we train over a large number of interventions, we may expect that a big neural network will generalize across them.* To address this, we make several points. To begin with, if data were not sufficiently diverse (which is an untestable assumption *a priori*), the worst case error to unseen shifts may still be arbitrarily high (see Section VII-C). While, in the short term, we can often beat "out-of-distribution" benchmarks by training bigger models on bigger data sets, causality offers an important complement. The generalization capabilities of a model are tied to its assumptions (e.g., how the model is structured and how it was trained). The causal approach makes these assumptions more explicit and aligned with our understanding of physics and human cognition, for instance, by relying on the ICM principle. When these assumptions are valid, a learner that does not use them should fare worse than one that does. Furthermore, if we had a model that was successful in all interventions over a certain environment, we may want to use it in different environments that share similar albeit not necessarily identical dynamics. The causal approach and, in particular, the ICM principle, point to the need to decompose knowledge about the world into independent and recomposable pieces (recomposable depending on the interventions or changes in the environment), which suggests more work on modular ML architectures and other ways to enforce the ICM principle in future ML approaches.

At its core, i.i.d. pattern recognition is but a mathematical abstraction, and causality may be essential to most forms of animate learning. Up until now, machine learning has neglected a full integration of causality, and this article argues that it would indeed benefit from integrating causal concepts. We argue that combining the strengths of both fields, that is, current deep learning methods and tools and ideas from causality, may be a necessary step on the path toward versatile AI systems.

VIII. CONCLUSION

In this work, we discussed different levels of models, including causal and statistical ones. We argued that this spectrum builds upon a range of assumptions, both in terms of modeling and data collection. In an effort to bring together causality and machine learning research programs, we first presented a discussion on the fundamentals of causal inference. Second, we discussed how the independent mechanism assumptions and related notions, such as invariance, offer a powerful bias for causal learning. Third, we discussed how causal relations might be learned from observational and interventional data when causal variables are observed. Fourth, we discussed the open problem of causal representation learning, including its relation to the recent interest in the concept of disentangled representations in deep learning. Finally, we discussed how some open research questions in the machine learning community may be better understood and tackled within the causal framework, including SSL, domain generalization, and adversarial robustness.

Based on this discussion, we list some critical areas for future research.

A. Learning Nonlinear Causal Relations at Scale

Not all real-world data are unstructured and the effect of interventions can often be observed, for example, by stratifying the data collection across multiple environments. The approximation abilities of modern machine learning methods may prove useful to model nonlinear causal relations among large numbers of variables. For practical applications, classical tools are not only limited in the linearity assumptions often made, but also in their scalability. The paradigms of metalearning and multitask learning are close to the assumptions and desiderata of causal modeling, and future work should consider: 1) understanding under which conditions nonlinear causal relations can be learned; 2) which training frameworks allow to best exploit the scalability of machine learning approaches; and 3) providing compelling evidence on the advantages over (noncausal) statistical representations in terms of generalization, repurposing, and transfer of causal modules on real-world tasks.

B. Learning Causal Variables

"Disentangled" representations learned by state-of-the-art neural network methods are still distributed in the sense that they are represented in a vector format with an arbitrary ordering in the dimensions. This fixed-format implies that the representation size cannot be dynamically changed; for example, we cannot change the number of objects in a scene. Furthermore, structured and modular representations should also arise when a network is trained for (sets of) specific tasks, not only autoencoding. Different high-level variables may be extracted depending on the task and affordances at hand. Understanding under which conditions causal variables can be recovered could provide insights into which interventions are robust to predictive tasks.

C. Understanding the Biases of Existing Deep Learning Approaches

Scaling to massive data sets and relying on data augmentation and self-supervision have all been successfully explored to improve the robustness of the predictions of deep learning models. It is nontrivial to disentangle the benefits of the individual components, and it is often unclear which "trick" should be used when dealing with a new task, even if we have an intuition about useful invariances. The notion of strong generalization over a specific set of interventions may be used to probe existing methods, training schemes, and data sets in order to build a taxonomy of inductive biases. In particular, it is desirable to understand how design choices in pretraining (e.g., which data sets/tasks) positively impact both transfer and robustness downstream in a causal sense.

D. Learning Causally Correct Models of the World and the Agent

In many real-world RL settings, abstract state representations are not available. Hence, the ability to derive abstract causal variables from high-dimensional, low-level pixel representations and then recover causal graphs is important for causal induction in real-world RL settings. Moreover, building a causal description for both a model of the agent and the environment (world models) should be essential for robust and versatile model-based RL. ■

Acknowledgment

The authors thank the past and present members of the Tübingen Causality Team, without whose work and insights, this article would not exist, in particular, to Dominik Janzing, Chaochao Lu, and Julius von Kügelgen who gave helpful comments on [217]. The text has also benefitted from discussions with Elias Bareinboim, Christoph Bohle, Leon Bottou, Isabelle Guyon, Judea Pearl, and Vladimir Vapnik. The authors would like to thank Wouter van Amsterdam for pointing out typos in the first version. They also thank Thomas Kipf, Klaus Greff, and Alexander d'Amour for the useful discussions. Finally, they thank the thorough anonymous reviewers for highly valuable feedback and suggestions.

REFERENCES

[1] O. Ahmed *et al.*, "Causalworld: A robotic manipulation benchmark for causal structure and transfer learning," in *Proc. Int. Conf. Learn. Represent.*, 2021.

[2] OpenAI *et al.*, "Solving Rubik's cube with a robot hand," 2019, *arXiv:1910.07113*. [Online]. Available: http://arxiv.org/abs/1910.07113

[3] A. Alaa and M. Schaar, "Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 129–138.

[4] J. Aldrich, "Autonomy," *Oxford Econ. Papers*, vol. 41, no. 1, pp. 15–34, 1989.

[5] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," 2019, *arXiv:1907.02893*. [Online]. Available: http://arxiv.org/abs/1907.02893

[6] O. Atan, J. Jordon, and M. van der Schaar, "Deep-treat: Learning optimal personalized treatments from observational data using neural networks," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018.

[7] A. Azulay and Y. Weiss, "Why do deep convolutional networks generalize so poorly to small image transformations?" *J. Mach. Learn. Res.*, vol. 20, no. 184, pp. 1–25, 2019.

[8] A. L. Rezaabad and S. Vishwanath, "Learning representations by maximizing mutual information in variational autoencoders," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 15535–15545.

[9] D. Bahdanau, S. Murty, M. Noukhovitch, T. H. Nguyen, H. de Vries, and A. Courville, "Systematic generalization: What is required and can it be learned," 2018, *arXiv:1811.12889*. [Online]. Available: https://arxiv.org/abs/1811.12889

[10] H. Baird, "Document image defect models," in *Proc. IAPR Workshop Syntactic Structural Pattern Recognit.*, Murray Hill, NJ, USA, 1990, pp. 38–46.

[11] V. Bapst *et al.*, "Structured agents for physical construction," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 464–474.

[12] A. Barbu *et al.*, "Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 9448–9458.

[13] E. Bareinboim, A. Forney, and J. Pearl, "Bandits with unobserved confounders: A causal approach," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1342–1350.

[14] E. Bareinboim and J. Pearl, "Transportability from multiple environments with limited experiments: Completeness results," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 280–288.

[15] P. Battaglia *et al.*, "Interaction networks for learning about objects, relations and physics," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4502–4510.

[16] P. W. Battaglia *et al.*, "Relational inductive biases, deep learning, and graph networks," 2018, *arXiv:1806.01261*. [Online]. Available: http://arxiv.org/abs/1806.01261

[17] P. W. Battaglia, J. B. Hamrick, and J. B. Tenenbaum, "Simulation as an engine of physical scene understanding," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 45, pp. 18327–18332, Nov. 2013.

[18] S. Bauer, B. Schölkopf, and J. Peters, "The arrow of time in multivariate time series," in *Proc. 33rd Int. Conf. Mach. Learn.*, vol. 48, 2016, pp. 2043–2051.

[19] E. Beede *et al.*, "A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2020, pp. 1–12.

[20] S. Beery, G. Van Horn, and P. Perona, "Recognition in terra incognita," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 456–473.

[21] S. Ben-David, T. Lu, T. Luu, and D. Pál, "Impossibility theorems for domain adaptation," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, 2010, pp. 129–136.

[22] E. Bengio, V. Thomas, J. Pineau, D. Precup, and Y. Bengio, "Independently controllable features," 2017, *arXiv:1703.07718*. [Online]. Available: http://arxiv.org/abs/1703.07718

[23] Y. Bengio, S. Bengio, and J. Cloutier, "Learning a synaptic learning rule," in *Proc. Seattle Int. Joint Conf. Neural Netw. (IJCNN)*, vol. 2. IEEE, Jul. 1991, p. 969.

[24] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," 2012, *arXiv:1206.5538*. [Online]. Available: http://arxiv.org/abs/1206.5538

[25] Y. Bengio *et al.*, "A meta-transfer objective for learning to disentangle causal mechanisms," 2019, *arXiv:1901.10912*. [Online]. Available: http://arxiv.org/abs/1901.10912

[26] B. Benneke *et al.*, "Water vapor on the habitable-zone exoplanet K2-18b," 2019, *arXiv:1909.04642*. [Online]. Available: https://arxiv.org/abs/1909.04642

[27] OpenAI *et al.*, "Dota 2 with large scale deep reinforcement learning," 2019, *arXiv:1912.06680*. [Online]. Available: http://arxiv.org/abs/1912.06680

[28] M. Besserve, N. Shajarisales, B. Schölkopf, and D. Janzing, "Group invariance principles for causal generative models," in *Proc. 21st Int. Conf. Artif. Intell. Statist. (AISTATS)*, 2018, pp. 557–565.

[29] M. Besserve, R. Sun, D. Janzing, and B. Schölkopf, "A theory of independent mechanisms for extrapolation in generative models," in *Proc. 35th AAAI Conf. Artif. Intell. Virtual Conf.*, Feb. 2021.

[30] M. Besserve, A. Mehrjou, R. Sun, and B. Schölkopf, "Counterfactuals uncover the modular structure of deep generative models," 2018, *arXiv:1812.03253*. [Online]. Available: http://arxiv.org/abs/1812.03253

[31] I. Bica, A. M. Alaa, and M. van der Schaar, "Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders," 2019, *arXiv:1902.00450*. [Online]. Available: http://arxiv.org/abs/1902.00450

[32] P. Blöbaum, T. Washio, and S. Shimizu, "Error asymmetry in causal and anticausal regression," 2016, *arXiv:1610.03263*. [Online]. Available: http://arxiv.org/abs/1610.03263

[33] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. 11th Annu. Conf. Comput. Learn. Theory*, New York, NY, USA, 1998, pp. 92–100.

[34] B. Bonet and H. Geffner, "Learning first-order symbolic representations for planning from the structure of the state space," 2019, *arXiv:1909.05546*. [Online]. Available: http://arxiv.org/abs/1909.05546

[35] L. Bottou *et al.*, "Counterfactual reasoning and learning systems: The example of computational advertising," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 3207–3260, Jan. 2013.

[36] T. B. Brown *et al.*, "Language models are few-shot learners," 2020, *arXiv:2005.14165*. [Online]. Available: https://arxiv.org/abs/2005.14165

[37] K. Budhathoki and J. Vreeken, "Causal inference by compression," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Dec. 2016, pp. 41–50.

[38] L. Buesing *et al.*, "Woulda, coulda, shoulda: Counterfactually-guided policy search," 2018, *arXiv:1811.06272*. [Online]. Available: http://arxiv.org/abs/1811.06272

[39] C. J. C. Burges and B. Schölkopf, "Improving the accuracy and speed of support vector learning machines," in *Advances in Neural Information Processing Systems*, vol. 9, M. Mozer, M. Jordan, and T. Petsche, Eds. Cambridge, MA, USA: MIT Press, 1997, pp. 375–381.

[40] C. P. Burgess *et al.*, "MONet: Unsupervised scene decomposition and representation," 2019, *arXiv:1901.11390*. [Online]. Available: http://arxiv.org/abs/1901.11390

[41] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[42] K. Chalupka, P. Perona, and F. Eberhardt, "Multi-level cause-effect systems," 2015, *arXiv:1512.07942*. [Online]. Available: http://arxiv.org/abs/1512.07942

[43] K. Chalupka, P. Perona, and F. Eberhardt, "Fast conditional independence test for vector variables with large sample sizes," 2018, *arXiv:1804.02747*. [Online]. Available: http://arxiv.org/abs/1804.02747

[44] M. B. Chang, T. Ullman, A. Torralba, and J. B. Tenenbaum, "A compositional object-based approach to learning physical dynamics," in *Proc. 5th Int. Conf. Learn. Represent. (ICLR)*, 2017.

[45] O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-Supervised Learning*. Cambridge, MA, USA: MIT Press, 2006.

[46] M. Chen *et al.*, "Generative pretraining from pixels," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 1691–1703.

[47] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020, *arXiv:2002.05709*. [Online]. Available: http://arxiv.org/abs/2002.05709

[48] S. Chiappa, S. Racaniere, D. Wierstra, and S. Mohamed, "Recurrent environment simulators," in *Proc. 5th Int. Conf. Learn. Represent. (ICLR)*, 2017.

[49] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning augmentation strategies from data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 113–123.

[50] P. Daniušis *et al.*, "Inferring deterministic causal relations," in *Proc. 26th Annu. Conf. Uncertainty Artif. Intell. (UAI)*, 2010, pp. 143–150.

[51] I. Dasgupta *et al.*, "Causal reasoning from meta-reinforcement learning," 2019, *arXiv:1901.08162*. [Online]. Available: http://arxiv.org/abs/1901.08162

[52] A. P. Dawid, "Conditional independence in statistical theory," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 41, no. 1, pp. 1–31, 1979.

[53] S. Dehaene, *How We Learn: Why Brains Learn Better Than Any Machine … for Now*. Baltimore, MD, USA: Penguin, 2020.

[54] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[55] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805

[56] L. Devroye, L. Györfi, and G. Lugosi, "A probabilistic theory of pattern recognition," in *Applications of Mathematics*, vol. 31. New York, NY, USA: Springer, 1996.

[57] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," 2020, *arXiv:2005.00341*. [Online]. Available: http://arxiv.org/abs/2005.00341

[58] A. Dittadi *et al.*, "On the transfer of disentangled representations in realistic settings," in *Proc. Int. Conf. Learn. Represent.*, 2021.

[59] C. Diuk, A. Cohen, and M. L. Littman, "An object-oriented representation for efficient reinforcement learning," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, 2008, pp. 240–247.

[60] J. Djolonga *et al.*, "On robustness and transferability of convolutional neural networks," 2020, *arXiv:2007.08558*. [Online]. Available: https://arxiv.org/abs/2007.08558

[61] G. Doran, K. Muandet, K. Zhang, and B. Schölkopf, "A permutation-based kernel conditional independence test," in *Proc. 30th Conf. Uncertainty Artif. Intell.*, N. L. Zhang and J. Tian, Eds. Corvallis, OR, USA: AUAI Press, 2014, pp. 132–141.

[62] C. Eastwood and C. K. Williams, "A framework for the quantitative evaluation of disentangled

representations," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[63] D. Eaton and K. Murphy, "Exact Bayesian structure learning from uncertain interventions," in *Proc. Artif. Intell. Statist.*, 2007, pp. 107–114.

[64] L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry, "Exploring the landscape of spatial robustness," 2017, *arXiv:1712.02779*. [Online]. Available: http://arxiv.org/abs/1712.02779

[65] K. Epstude and N. J. Roese, "The functional theory of counterfactual thinking," *Personality Social Psychol. Rev.*, vol. 12, no. 2, pp. 168–192, May 2008.

[66] A. Esteva *et al.*, "A guide to deep learning in healthcare," *Nature Med.*, vol. 25, no. 1, pp. 24–29, 2019.

[67] A. Farago and G. Lugosi, "Strong universal consistency of neural network classifiers," *IEEE Trans. Inf. Theory*, vol. 39, no. 4, pp. 1146–1151, Jul. 1993.

[68] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," 2017, *arXiv:1703.03400*. [Online]. Available: http://arxiv.org/abs/1703.03400

[69] J. N. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018.

[70] P. Földiák, "Learning invariance from transformation sequences," *Neural Comput.*, vol. 3, no. 2, pp. 194–200, Jun. 1991.

[71] 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.*, vol. 806, no. 2, p. 215, 2015.

[72] R. Frisch, T. Haavelmo, T. Koopmans, and J. Tinbergen, "Autonomy of economic relations," Universitets Socialøkonomiske Institutt, Oslo, Norway, Tech. Rep., 1948.

[73] S. Fujimoto, D. Meger, and D. Precup, "Off-policy deep reinforcement learning without exploration," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2052–2062.

[74] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf, "Kernel measures of conditional dependence," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 489–496.

[75] D. Geiger and J. Pearl, "Logical and algorithmic properties of independence and their application to Bayesian networks," *Ann. Math. Artif. Intell.*, vol. 2, nos. 1–4, pp. 165–178, Mar. 1990.

[76] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," 2018, *arXiv:1811.12231*. [Online]. Available: http://arxiv.org/abs/1811.12231

[77] M. W. Gondal *et al.*, "On the transfer of inductive bias from simulation to the real world: A new disentanglement dataset," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 15740–15751.

[78] M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour, and B. Schölkopf, "Domain adaptation with conditional transferable components," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, pp. 2839–2848.

[79] M. Gong, K. Zhang, B. Schölkopf, C. Glymour, and D. Tao, "Causal discovery from temporally aggregated time series," in *Proc. 33rd Conf. Uncertainty Artif. Intell. (UAI)*, 2017, p. 269.

[80] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*. [Online]. Available: http://arxiv.org/abs/1412.6572

[81] A. Gopnik, C. Glymour, D. M. Sobel, L. E. Schulz, T. Kushnir, and D. Danks, "A theory of causal learning in children: Causal maps and Bayes nets," *Psychol. Rev.*, vol. 111, no. 1, pp. 3–32, 2004.

[82] O. Gottesman *et al.*, "Evaluating reinforcement learning algorithms in observational health settings," 2018, *arXiv:1805.12298*. [Online]. Available: http://arxiv.org/abs/1805.12298

[83] O. Goudet, D. Kalainathan, P. Caillou, I. Guyon, D. Lopez-Paz, and M. Sebag, "Causal generative neural networks," 2017, *arXiv:1711.08936*. [Online]. Available: http://arxiv.org/abs/1711.08936

[84] A. Goyal *et al.*, "Object files and schemata: Factorizing declarative and procedural knowledge in dynamical systems," 2020, *arXiv:2006.16225*. [Online]. Available: http://arxiv.org/abs/2006.16225

[85] A. Goyal, A. Lamb, J. Hoffmann, S. Sodhani, S. Levine, Y. Bengio, and B. Schölkopf, "Recurrent independent mechanisms," in *Proc. Int. Conf. Learn. Represent.*, 2021.

[86] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6645–6649.

[87] K. Greff *et al.*, "Multi-object representation learning with iterative variational inference," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2424–2433.

[88] K. Greff, S. van Steenkiste, and J. Schmidhuber, "On the binding problem in artificial neural networks," 2020, *arXiv:2012.05208*. [Online]. Available: http://arxiv.org/abs/2012.05208

[89] K. Gregor, D. J. Rezende, F. Besse, Y. Wu, H. Merzic, and A. van den Oord, "Shaping belief states with generative environment models for RL," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 13475–13487.

[90] L. Gresele, P. K. Rubenstein, A. Mehrjou, F. Locatello, and B. Schölkopf, "The incomplete Rosetta stone problem: Identifiability results for multi-view nonlinear ICA," 2019, *arXiv:1905.06642*. [Online]. Available: https://arxiv.org/abs/1905.06642

[91] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, "Measuring statistical dependence with Hilbert–Schmidt norms," in *Algorithmic Learning Theory*. Springer-Verlag, 2005, pp. 63–78.

[92] A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf, "Kernel methods for measuring independence," *J. Mach. Learn. Res.*, vol. 6, pp. 2075–2129, Dec. 2005.

[93] J.-B. Grill *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," 2020, *arXiv:2006.07733*. [Online]. Available: http://arxiv.org/abs/2006.07733

[94] R. Grzeszczuk, D. Terzopoulos, and G. Hinton, "NeuroAnimator: Fast neural network emulation and control of physics-based models," in *Proc. 25th Annu. Conf. Comput. Graph. Interact. Techn.*, 1998, pp. 9–20.

[95] K. Gu, B. Yang, J. Ngiam, Q. Le, and J. Shlens, "Using videos to evaluate image model robustness," 2019, *arXiv:1904.10076*. [Online]. Available: http://arxiv.org/abs/1904.10076

[96] S. Gu and L. Rigazio, "Towards deep neural network architectures robust to adversarial examples," 2014, *arXiv:1412.5068*. [Online]. Available: http://arxiv.org/abs/1412.5068

[97] R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu, "A survey of learning causality with data: Problems and methods," 2018, *arXiv:1809.09337*. [Online]. Available: http://arxiv.org/abs/1809.09337

[98] I. Guyon, D. Janzing, and B. Schölkopf, "Causality: Objectives and assessment," in *Proc. JMLR Workshop Conf.*, vol. 6, I. Guyon, D. Janzing, and B. Schölkopf, Eds. Cambridge, MA, USA: MIT Press, 2010, pp. 1–42.

[99] D. Ha and J. Schmidhuber, "World models," 2018, *arXiv:1803.10122*. [Online]. Available: http://arxiv.org/abs/1803.10122

[100] T. Haavelmo, "The probability approach in econometrics," *Econometrica*, vol. 12, pp. S1–S115, Jul. 1944.

[101] H. Hälvä and A. Hyvärinen, "Hidden Markov nonlinear ICA: Unsupervised learning from nonstationary time series," 2020, *arXiv:2006.12107*. [Online]. Available: http://arxiv.org/abs/2006.12107

[102] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.

[103] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[104] S. He *et al.*, "Learning to predict the cosmological structure formation," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 28, pp. 13825–13832, Jul. 2019.

[105] C. Heinze-Deml and N. Meinshausen, "Conditional variance penalties and domain shift robustness," 2017, *arXiv:1710.11469*. [Online]. Available: http://arxiv.org/abs/1710.11469

[106] C. Heinze-Deml, J. Peters, and N. Meinshausen, "Invariant causal prediction for nonlinear models," 2017, *arXiv:1706.08576*. [Online]. Available: http://arxiv.org/abs/1706.08576

[107] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," 2019, *arXiv:1903.12261*. [Online]. Available: http://arxiv.org/abs/1903.12261

[108] J. Henrich, *The Secret our Success*. Princeton, NJ, USA: Princeton Univ. Press, 2016.

[109] K. E. Henry, D. N. Hager, P. J. Pronovost, and S. Saria, "A targeted real-time early warning score (TREWScore) for septic shock," *Sci. Transl. Med.*, vol. 7, no. 299, 2015, Art. no. 299ra122.

[110] I. Higgins *et al.*, "Beta-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. Int. Conf. Learn. Represent.*, 2016.

[111] K. D. Hoover, "Causality in economics and econometrics," in *The New Palgrave Dictionary of Economics*, S. N. Durlauf and L. E. Blume, Eds., 2nd ed. Basingstoke, U.K.: Palgrave Macmillan, 2008.

[112] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," 2018, *arXiv:1801.06146*. [Online]. Available: http://arxiv.org/abs/1801.06146

[113] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, "Nonlinear causal discovery with additive noise models," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2009, pp. 689–696.

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

[115] 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 *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2017, pp. 913–918.

[116] A. Hyvarinen and H. Morioka, "Nonlinear ICA of temporally dependent stationary sources," in *Proc. Mach. Learn. Res.*, 2017, pp. 460–469.

[117] A. Hyvärinen and P. Pajunen, "Nonlinear independent component analysis: Existence and uniqueness results," *Neural Netw.*, vol. 12, no. 3, pp. 429–439, Apr. 1999.

[118] G. W. Imbens and D. B. Rubin, *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge, U.K.: Cambridge Univ. Press, 2015.

[119] D. Janzing, "Causal regularization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 12704–12714.

[120] D. Janzing, R. Chaves, and B. Schölkopf, "Algorithmic independence of initial condition and dynamical law in thermodynamics and causal inference," *New J. Phys.*, vol. 18, no. 9, Sep. 2016, Art. no. 093052.

[121] D. Janzing, P. Hoyer, and B. Schölkopf, "Telling cause from effect based on high-dimensional observations," in *Proc. 27th Int. Conf. Mach. Learn.*, J. Fürnkranz and T. Joachims, Eds., 2010, pp. 479–486.

[122] D. Janzing *et al.*, "Information-geometric approach to inferring causal directions," *Artif. Intell.*, vols. 182–183, pp. 1–31, May 2012.

[123] D. Janzing, J. Peters, J. M. Mooij, and B. Schölkopf, "Identifying confounders using additive noise models," in *Proc. 25th Annu. Conf.*

Uncertainty Artif. Intell. (UAI), 2009, pp. 249–257.

[124] D. Janzing and B. Schölkopf, "Causal inference using the algorithmic Markov condition," *IEEE Trans. Inf. Theory*, vol. 56, no. 10, pp. 5168–5194, Oct. 2010.

[125] D. Janzing and B. Schölkopf, "Semi-supervised interpolation in an anticausal learning scenario," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 1923–1948, 2015.

[126] D. Janzing and B. Schölkopf, "Detecting non-causal artifacts in multivariate linear regression models," in *Proc. 35th Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 2250–2258.

[127] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *J. Artif. Intell. Res.*, vol. 4, pp. 237–285, Jan. 1996.

[128] D. Kahneman, *Thinking, Fast Slow*. New York, NY, USA: Farrar, Straus and Giroux, 2011.

[129] S. Karahan, M. K. Yildirum, K. Kirtac, F. S. Rende, G. Butun, and H. K. Ekenel, "How image degradations affect deep CNN-based face recognition," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, 2016, pp. 1–5.

[130] A.-H. Karimi, J. von Kügelgen, B. Schölkopf, and I. Valera, "Algorithmic recourse under imperfect causal knowledge: A probabilistic approach," 2020, *arXiv:2006.06831*. [Online]. Available: http://arxiv.org/abs/2006.06831

[131] N. R. Ke *et al.*, "Learning neural causal models from unknown interventions," 2019, *arXiv:1910.01075*. [Online]. Available: http://arxiv.org/abs/1910.01075

[132] S. Tsirtsis, B. Tabibian, M. Khajehnejad, A. Singla, B. Schölkopf, and M. Gomez-Rodriguez, "Optimal decision making under strategic behavior," 2019, *arXiv:1905.09239*. [Online]. Available: http://arxiv.org/abs/1905.09239

[133] N. Kilbertus, G. Parascandolo, and B. Schölkopf, "Generalization in anti-causal learning," 2018, *arXiv:1812.00524*. [Online]. Available: http://arxiv.org/abs/1812.00524

[134] N. Kilbertus, M. R. Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, "Avoiding discrimination through causal reasoning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 656–666.

[135] H. Kim and A. Mnih, "Disentangling by factorising," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2649–2658.

[136] T. Kipf, E. Fetaya, K.-C. Wang, M. Welling, and R. Zemel, "Neural relational inference for interacting systems," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2688–2697.

[137] A. Kolesnikov *et al.*, "Big transfer (BiT): General visual representation learning," 2019, *arXiv:1912.11370*. [Online]. Available: http://arxiv.org/abs/1912.11370

[138] A. Kosiorek, H. Kim, Y. W. Teh, and I. Posner, "Sequential attend, infer, repeat: Generative modelling of moving objects," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 8606–8616.

[139] S. Kpotufe, E. Sgouritsa, D. Janzing, and B. Schölkopf, "Consistency of causal inference under the additive noise model," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, pp. 478–486.

[140] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[141] T. D. Kulkarni *et al.*, "Unsupervised learning of object keypoints for perception and control," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 10723–10733.

[142] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates, 2017, pp. 4066–4076.

[143] L. Ladický, S. Jeong, B. Solenthaler, M. Pollefeys, and M. Gross, "Data-driven fluid simulations using regression forests," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 1–9, Nov. 2015.

[144] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *Behav. Brain Sci.*, vol. 40, p. e253, Jan. 2017.

[145] J. Landman, E. A. Vandewater, A. J. Stewart, and J. E. Malley, "Missed opportunities: Psychological ramifications of counterfactual thought in midlife women," *J. Adult Develop.*, vol. 2, no. 2, pp. 87–97, Apr. 1995.

[146] S. Lange, T. Gabel, and M. Riedmiller, "Batch reinforcement learning," in *Reinforcement Learning: State-of-the-Art*, M. Wiering and M. van Otterlo, Eds. Berlin, Germany: Springer, 2012, pp. 45–73.

[147] S. L. Lauritzen, *Graphical Models*. New York, NY, USA: Oxford Univ. Press, 1996.

[148] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[149] F. Leeb, Y. Annadani, S. Bauer, and B. Schölkopf, "Structural autoencoders improve representations for generation and transfer," 2020, *arXiv:2006.07796*. [Online]. Available: http://arxiv.org/abs/2006.07796

[150] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," 2020, *arXiv:2005.01643*. [Online]. Available: http://arxiv.org/abs/2005.01643

[151] D. Lewis, "Causation," *J. Philosophy*, vol. 70, no. 17, pp. 556–567, 1974.

[152] Y. Li, M. Gong, X. Tian, T. Liu, and D. Tao, "Domain generalization via conditional invariant representation," 2018, *arXiv:1807.08479*. [Online]. Available: http://arxiv.org/abs/1807.08479

[153] Y. Li *et al.*, "Deep domain generalization via conditional invariant adversarial networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 624–639.

[154] S. Lim, I. Kim, T. Kim, C. Kim, and S. Kim, "Fast autoaugment," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 6665–6675.

[155] Z. Lin *et al.*, "Space: Unsupervised object-oriented scene representation via spatial attention and decomposition," in *Proc. Int. Conf. Learn. Represent.*, 2019.

[156] Z. C. Lipton, Y.-X. Wang, and A. Smola, "Detecting and correcting for label shift with black box predictors," 2018, *arXiv:1802.03916*. [Online]. Available: http://arxiv.org/abs/1802.03916

[157] F. Locatello, G. Abbati, T. Rainforth, S. Bauer, B. Schölkopf, and O. Bachem, "On the fairness of disentangled representations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 14544–14557.

[158] F. Locatello *et al.*, "Challenging common assumptions in the unsupervised learning of disentangled representations," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 4114–4124.

[159] F. Locatello, B. Poole, G. Rätsch, B. Schölkopf, O. Bachem, and M. Tschannen, "Weakly-supervised disentanglement without compromises," in *Proc. 37th Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 6348–6359.

[160] F. Locatello *et al.*, "Object-centric learning with slot attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020.

[161] D. Lopez-Paz, K. Muandet, B. Schölkopf, and I. Tolstikhin, "Towards a learning theory of cause-effect inference," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 1452–1461.

[162] D. Lopez-Paz, R. Nishihara, S. Chintala, B. Scholkopf, and L. Bottou, "Discovering causal signals in images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 58–66.

[163] K. Lorenz, *Die Rückseite des Spiegels*. Munich, Germany: Piper Verlag, 1973.

[164] C. Lu, B. Huang, K. Wang, J. Miguel Hernández-Lobato, K. Zhang, and B. Schölkopf, "Sample-efficient reinforcement learning via counterfactual-based data augmentation," 2020, *arXiv:2012.09092*. [Online]. Available: http://arxiv.org/abs/2012.09092

[165] C. Lu, B. Schölkopf, and J. M. Hernández-Lobato, "Deconfounding reinforcement learning in observational settings," 2018, *arXiv:1812.10576*. [Online]. Available: http://arxiv.org/abs/1812.10576

[166] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, May 2019.

[167] S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij, "Domain adaptation by using causal inference to predict invariant conditional distributions," in *Proc. NeurIPS*, 2018, pp. 10869–10879.

[168] R. Matthews, "Storks deliver babies (p= 0.008)," *Teaching Statist.*, vol. 22, no. 2, pp. 36–38, 2000.

[169] N. Meinshausen, "Causality from a distributional robustness point of view," in *Proc. IEEE Data Sci. Workshop (DSW)*, Jun. 2018, pp. 6–10.

[170] C. Michaelis *et al.*, "Benchmarking robustness in object detection: Autonomous driving when winter is coming," 2019, *arXiv:1907.07484*. [Online]. Available: http://arxiv.org/abs/1907.07484

[171] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[172] B. T. Montet *et al.*, "Stellar and planetary properties of K2 campaign 1 candidates and validation of 17 planets, including a planet receiving earth-like insolation," *Astrophys. J.*, vol. 809, no. 1, p. 25, 2015.

[173] J. Mooij, D. Janzing, and B. Schölkopf, "From ordinary differential equations to structural causal models: The deterministic case," in *Proc. 29th Conf. Annu. Conf. Uncertainty Artif. Intell.*, A. Nicholson and P. Smyth, Eds. Corvallis, OR, USA: AUAI Press, 2013, pp. 440–448.

[174] J. M. Mooij, D. Janzing, T. Heskes, and B. Schölkopf, "On causal discovery with cyclic additive noise models," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2011, pp. 639–647.

[175] J. M. Mooij, D. Janzing, J. Peters, and B. Schölkopf, "Regression by dependence minimization and its application to causal inference," in *Proc. 26th Int. Conf. Mach. Learn. (ICML)*, 2009, pp. 745–752.

[176] 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.*, vol. 17, no. 32, pp. 1–102, 2016.

[177] D. Mrowca *et al.*, "Flexible neural representation for physics prediction," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 8799–8810.

[178] J. Oh, X. Guo, H. Lee, R. L. Lewis, and S. Singh, "Action-conditional video prediction using deep networks in Atari games," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2863–2871.

[179] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*. [Online]. Available: http://arxiv.org/abs/1807.03748

[180] G. Parascandolo, N. Kilbertus, M. Rojas-Carulla, and B. Schölkopf, "Learning independent causal mechanisms," in *Proc. 35th Int. Conf. Mach. Learn. (PMLR)*, vol. 80, 2018, pp. 4036–4044.

[181] G. Parascandolo, A. Neitz, A. ORVIETO, L. Gresele, and B. Schölkopf, "Learning explanations that are hard to vary," in *Proc. Int. Conf. Learn. Represent.*, 2021.

[182] G. Parascandolo, M. Rojas-Carulla, N. Kilbertus, and B. Schölkopf, "Learning independent causal mechanisms," in *Proc. Workshop Learn. Disentangled Represent. From Perception Control 31st Conf. Neural Inf. Process. Syst. (NIPS)*, 2017.

[183] J. Pearl, *Causality: Models, Reasoning, Inference*, 2nd ed. New York, NY, USA: Cambridge Univ. Press, 2009.

[184] J. Pearl, "Giving computers free will," Forbes, 2009.

[185] J. Pearl and E. Bareinboim, "External validity: From do-calculus to transportability across populations," 2015, *arXiv:1503.01603*. [Online]. Available: http://arxiv.org/abs/1503.01603

[186] J. Peters, S. Bauer, and N. Pfister, "Causal models for dynamical systems," 2020, *arXiv:2001.06208*. [Online]. Available: http://arxiv.org/abs/2001.06208

[187] J. Peters, P. Bühlmann, and N. Meinshausen, "Causal inference by using invariant prediction:

Identification and confidence intervals," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 78, no. 5, pp. 947–1012, Nov. 2016.

[188] J. Peters, D. Janzing, and B. Schölkopf," *Elements of Causal Inference—Foundations and Learning Algorithms*. Cambridge, MA, USA: MIT Press, 2017.

[189] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf, "Identifiability of causal graphs using functional models," in *Proc. 27th Annu. Conf. Uncertainty Artif. Intell. (UAI)*, 2011, pp. 589–598.

[190] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf, "Causal discovery with continuous additive noise models," *J. Mach. Learn. Res.*, vol. 15, pp. 2009–2053, Jan. 2014.

[191] N. Pfister, S. Bauer, and J. Peters, "Learning stable and predictive structures in kinetic systems," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 51, pp. 25405–25411, Dec. 2019.

[192] N. Pfister, P. Bühlmann, and J. Peters, "Invariant causal prediction for sequential data," *J. Amer. Stat. Assoc.*, vol. 114, no. 527, pp. 1264–1276, Jul. 2019.

[193] N. Pfister, P. Bühlmann, B. Schölkopf, and J. Peters, "Kernel-based tests for joint independence," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 80, no. 1, pp. 5–31, Jan. 2018.

[194] R. L. Priol, R. B. Harikandeh, Y. Bengio, and S. Lacoste-Julien, "An analysis of the adaptation speed of causal models," 2020, *arXiv:2005.09136*. [Online]. Available: http://arxiv.org/abs/2005.09136

[195] S. Rabanser, S. Günnemann, and Z. C. Lipton, "Failing loudly: An empirical study of methods for detecting dataset shift," 2018, *arXiv:1810.11953*. [Online]. Available: http://arxiv.org/abs/1810.11953

[196] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," Tech. Rep., 2018.

[197] N. Rahaman *et al.*, "Spatially structured recurrent modules," in *Proc. Int. Conf. Learn. Represent.*, 2021.

[198] H. Reichenbach, *The Direction Time*. Berkeley, CA, USA: Univ. of California Press, 1956.

[199] L. K. Reichert and J. R. Slate, "Reflective learning: The use of 'if only ...' statements to improve performance," *Social Psychol. Educ.*, vol. 3, no. 4, pp. 261–275, 1999.

[200] D. J. Rezende *et al.*, "Causally correct partial models for reinforcement learning," 2020, *arXiv:2002.02836*. [Online]. Available: http://arxiv.org/abs/2002.02836

[201] J. G. Richens, C. M. Lee, and S. Johri, "Improving the accuracy of medical diagnosis with causal machine learning," *Nature Commun.*, vol. 11, no. 1, p. 3923, Dec. 2020.

[202] K. Ridgeway and M. C. Mozer, "Learning deep disentangled embeddings with the f-statistic loss," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 185–194.

[203] N. J. Roese, "The functional basis of counterfactual thinking," *J. Personality Social Psychol.*, vol. 66, no. 5, p. 805, 1994.

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

[205] M. Rolinek, D. Zietlow, and G. Martius, "Variational autoencoders pursue PCA directions (by Accident)," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12406–12415.

[206] P. Roy, S. Ghosh, S. Bhattacharya, and U. Pal, "Effects of degradations on deep neural network architectures," 2018, *arXiv:1807.10108*. [Online]. Available: http://arxiv.org/abs/1807.10108

[207] P. K. Rubenstein, S. Bongers, B. Schölkopf, and J. M. Mooij, "From deterministic ODEs to dynamic structural causal models," in *Proc. 34th Conf. Uncertainty Artif. Intell. (UAI)*, 2018, pp. 114–123.

[208] P. K. Rubenstein *et al.*, "Causal consistency of structural equation models," in *Proc. Thirty-Third Conf. Uncertainty Artif. Intell.*, 2017, pp. 808–817.

[209] S. Ruder, "An overview of multi-task learning in deep neural networks," 2017, *arXiv:1706.05098*. [Online]. Available: http://arxiv.org/abs/1706.05098

[210] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ, USA: Prentice-Hall, 2002.

[211] A. Sanchez-Gonzalez, J. Godwin, T. Pfaff, R. Ying, J. Leskovec, and P. W. Battaglia, "Learning to simulate complex physics with graph networks," 2020, *arXiv:2002.09405*. [Online]. Available: http://arxiv.org/abs/2002.09405

[212] A. Santoro *et al.*, "A simple neural network module for relational reasoning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4967–4976.

[213] J. Schmidhuber, "Evolutionary principles in self-referential learning, or on learning how to learn: The meta-meta-... hook," Ph.D. dissertation, Technische Universität München, München, Germany, 1987.

[214] J. Schmidhuber, "Curious model-building control systems," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Nov. 1991, pp. 1458–1463.

[215] B. Schölkopf, "Artificial intelligence: Learning to see and act," *Nature*, vol. 518, no. 7540, pp. 486–487, 2015.

[216] B. Schölkopf, "Causal learning, 2017," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, 2017. [Online]. Available: https://vimeo.com/238274659

[217] B. Schölkopf, "Causality for machine learning," 2019, *arXiv:1911.10500*. [Online]. Available: https://arxiv.org/abs/1911.10500

[218] B. Schölkopf *et al.*, "Modeling confounding by half-sibling regression," *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 27, pp. 7391–7398, Jul. 2016.

[219] B. Schölkopf, D. Janzing, and D. Lopez-Paz, "Causal and statistical learning," *Oberwolfach Rep.*, vol. 13, no. 3, pp. 1896–1899, 2016.

[220] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. M. Mooij, "On causal and anticausal learning," in *Proc. 29th Int. Conf. Mach. Learn. (ICML)*, 2012, pp. 1255–1262.

[221] B. Schölkopf and A. J. Smola, *Learning With Kernels*. Cambridge, MA, USA: MIT Press, 2002.

[222] L. Schott, J. Rauber, M. Bethge, and W. Brendel, "Towards the first adversarially robust neural network model on MNIST," in *Proc. Int. Conf. Learn. Represent.*, 2019.

[223] J. Schrittwieser *et al.*, "Mastering Atari, go, chess and Shogi by planning with a learned model," 2019, *arXiv:1911.08265*. [Online]. Available: http://arxiv.org/abs/1911.08265

[224] P. Schulam and S. Saria, "Reliable decision support using counterfactual models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1697–1708.

[225] R. D. Shah and J. Peters, "The hardness of conditional independence testing and the generalised covariance measure," 2018, *arXiv:1804.07203*. [Online]. Available: http://arxiv.org/abs/1804.07203

[226] N. Shajarisales, D. Janzing, B. Schölkopf, and M. Besserve, "Telling cause from effect in deterministic linear dynamical systems," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 285–294.

[227] V. Shankar, A. Dave, R. Roelofs, D. Ramanan, B. Recht, and L. Schmidt, "Do image classifiers generalize across time," 2019, *arXiv:1906.02168*. [Online]. Available: https://arxiv.org/abs/1906.02168

[228] R. Shetty, B. Schiele, and M. Fritz, "Not using the car to see the sidewalk—Quantifying and controlling the effects of context in classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8218–8226.

[229] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. J. Kerminen, "A linear non-Gaussian acyclic model for causal discovery," *J. Mach. Learn. Res.*, vol. 7, no. 10, pp. 2003–2030, 2006.

[230] R. Shu, Y. Chen, A. Kumar, S. Ermon, and B. Poole, "Weakly supervised disentanglement with guarantees," 2019, *arXiv:1910.09772*. [Online]. Available: http://arxiv.org/abs/1910.09772

[231] D. Silver *et al.*, "The predictron: End-to-end learning and planning," in *Proc. Int. Conf. Mach. Learn. (PMLR)*, 2017, pp. 3191–3199.

[232] D. Silver *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.

[233] P. Simard, B. Victorri, Y. LeCun, and J. Denker, "Tangent prop—A formalism for specifying selected invariances in an adaptive network," in *Advances in Neural Information Processing Systems*, vol. 4, J. Moody, S. Hanson, R. P. Lippmann, Eds. San Mateo, CA, USA: Morgan Kaufmann, 1992, pp. 895–903.

[234] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *Proc. 7th Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 3, 2003.

[235] H. A. Simon, "Causal ordering and identifiability," in *Studies in Econometric Methods*, W. C. Hood and T. C. Koopmans, Eds. New York, NY, USA: Wiley, 1953, pp. 49–74.

[236] E. S. Spelke, "Principles of object perception," *Cognit. Sci.*, vol. 14, no. 1, pp. 29–56, Jan. 1990.

[237] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*, 2nd ed. Cambridge, MA, USA: MIT Press, 2000.

[238] W. Spohn, *Grundlagen der Entscheidungstheorie*. Berlin, Germany: Scriptor-Verlag, 1978.

[239] I. Steinwart and A. Christmann, *Support Vector Machines*. New York, NY, USA: Springer, 2008.

[240] B. Steudel, D. Janzing, and B. Schölkopf, "Causal Markov condition for submodular information measures," in *Proc. 23rd Annu. Conf. Learn. Theory (COLT)*, 2010, pp. 464–476.

[241] J. Su, S. Adams, and P. A. Beling, "Counterfactual multi-agent reinforcement learning with graph convolution communication," 2020, *arXiv:2004.00470*. [Online]. Available: http://arxiv.org/abs/2004.00470

[242] A. Subbaswamy and S. Saria, "Counterfactual normalization: Proactively addressing dataset shift and improving reliability using causal mechanisms," 2018, *arXiv:1808.03253*. [Online]. Available: http://arxiv.org/abs/1808.03253

[243] A. Subbaswamy, P. Schulam, and S. Saria, "Preventing failures due to dataset shift: Learning predictive models that transport," 2018, *arXiv:1812.04597*. [Online]. Available: http://arxiv.org/abs/1812.04597

[244] C. Sun, P. Karlsson, J. Wu, J. B Tenenbaum, and K. Murphy, "Stochastic prediction of multi-agent interactions from partial observations," 2019, *arXiv:1902.09641*. [Online]. Available: http://arxiv.org/abs/1902.09641

[245] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 843–852.

[246] X. Sun, D. Janzing, and B. Schölkopf, "Causal inference by choosing graphs with most plausible Markov kernels," in *Proc. 9th Int. Symp. Artif. Intell. Math.*, 2006, pp. 1–11.

[247] R. Suter, D. Miladinovic, B. Schölkopf, and S. Bauer, "Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness," in *Proc. Int. Conf. Mach. Learn. (PMLR)*, 2019, pp. 6056–6065.

[248] R. S. Sutton *et al.*, *Introduction to Reinforcement Learning*, vol. 135. Cambridge, MA, USA: MIT Press, 1998.

[249] C. Szegedy *et al.*, "Intriguing properties of neural networks," 2013, *arXiv:1312.6199*. [Online]. Available: http://arxiv.org/abs/1312.6199

[250] E. Téglás, E. Vul, V. Girotto, M. Gonzalez, J. B. Tenenbaum, and L. L. Bonatti, "Pure reasoning in 12-month-old infants as probabilistic inference," *Science*, vol. 332, no. 6033, pp. 1054–1059, May 2011.

[251] J. Tian and J. Pearl, "Causal discovery from changes," in *Proc. 17th Annual Conf. Uncertainty Artif. Intell. (UAI)*, 2001, pp. 512–522.

[252] F. Träuble *et al.*, "Is independence all you need? On the generalization of representations learned from correlated data," 2020, *arXiv:2006.07886*.

[Online]. Available: http://arxiv.org/abs/2006.07886

[253] M. Tschannen *et al.*, "Self-supervised learning of video-induced visual invariances," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13806–13815.

[254] 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," *Nature Astron.*, vol. 3, pp. 1086–1091, Sep. 2019.

[255] S. Van Steenkiste, M. Chang, K. Greff, and J. Schmidhuber, "Relational neural expectation maximization: Unsupervised discovery of objects and their interactions," in *Proc. 6th Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.

[256] S. van Steenkiste, F. Locatello, J. Schmidhuber, and O. Bachem, "Are disentangled representations helpful for abstract visual reasoning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 14178–14191.

[257] V. N. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, 1998.

[258] O. Vinyals *et al.*, "Grandmaster level in StarCraft II using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, Nov. 2019.

[259] J. von Kügelgen, U. Bhatt, A.-H. Karimi, I. Valera, A. Weller, and B. Schölkopf, "On the fairness of causal algorithmic recourse," 2020, *arXiv:2010.06529*. [Online]. Available: http://arxiv.org/abs/2010.06529

[260] 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," 2020, *arXiv:2005.07180*. [Online]. Available: http://arxiv.org/abs/2005.07180

[261] J. von Kügelgen, A. Mey, M. Loog, and B. Schölkopf, "Semi-supervised learning, causality and the conditional cluster assumption," in *Proc. Conf. Uncertainty Artif. Intell. (UAI)*, 2020, pp. 1–10.

[262] J. von Kügelgen, I. Ustyuzhaninov, P. Gehler, M. Bethge, and B. Schölkopf, "Towards causal generative scene models via competition of experts," 2020, *arXiv:2004.12906*. [Online]. Available: http://arxiv.org/abs/2004.12906

[263] H. Wang, Z. He, Z. C. Lipton, and E. P. Xing, "Learning robust representations by projecting superficial statistics out," 2019, *arXiv:1903.06256*. [Online]. Available: http://arxiv.org/abs/1903.06256

[264] N. Watters, L. Matthey, M. Bosnjak, C. P. Burgess, and A. Lerchner, "COBRA: Data-efficient model-based RL through unsupervised object discovery and curiosity-driven exploration," 2019, *arXiv:1905.09275*. [Online]. Available: http://arxiv.org/abs/1905.09275

[265] N. Watters, D. Zoran, T. Weber, P. Battaglia, R. Pascanu, and A. Tacchetti, "Visual interaction networks: Learning a physics simulator from video," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4539–4547.

[266] S. Weichwald, "Pragmatism and variable transformations in causal modelling," Ph.D. dissertation, ETH Zurich, Zürich, Switzerland, 2019.

[267] S. Weichwald, B. Schölkopf, T. Ball, and M. Grosse-Wentrup, "Causal and anti-causal learning in pattern recognition for neuroimaging," in *Proc. 4th Int. Workshop Pattern Recognit. Neuroimag. (PRNI)*, 2014, pp. 1–4.

[268] M. Wiering and M. Van Otterlo, *Reinforcement Learning*, vol. 12. Springer, 2012.

[269] S. Wiewel, M. Becher, and N. Thuerey, "Latent space physics: Towards learning the temporal evolution of fluid flow," in *Computer Graphics Forum*, vol. 38. Hoboken, NJ, USA: Wiley, 2019, pp. 71–82.

[270] L. Wiskott and T. J. Sejnowski, "Slow feature analysis: Unsupervised learning of invariances," *Neural Comput.*, vol. 14, no. 4, pp. 715–770, Apr. 2002.

[271] C. Xie, S. Patil, T. Moldovan, S. Levine, and P. Abbeel, "Model-based reinforcement learning with parametrized physical models and optimism-driven exploration," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2016, pp. 504–511.

[272] K. Yi *et al.*, "CLEVRER: CoLlision events for video REpresentation and reasoning," 2019, *arXiv:1910.01442*. [Online]. Available: http://arxiv.org/abs/1910.01442

[273] J. Yoon, J. Jordon, and M. van der Schaar, "GANITE: Estimation of individualized treatment effects using generative adversarial nets," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[274] V. Zambaldi *et al.*, "Deep reinforcement learning with relational inductive biases," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[275] J. Zhang and E. Bareinboim, "Fairness in decision-making—The causal explanation formula," in *Proc. 32nd AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, 2018, pp. 2037–2045.

[276] J. Zhang and E. Bareinboim, "Near-optimal reinforcement learning in dynamic treatment regimes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 13401–13411.

[277] K. Zhang, M. Gong, and B. Schölkopf, "Multi-source domain adaptation: A causal view," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 3150–3157.

[278] K. Zhang, B. Huang, J. Zhang, C. Glymour, and B. Schölkopf, "Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 1347–1353.

[279] K. Zhang and A. Hyvärinen, "On the identifiability of the post-nonlinear causal model," in *Proc. 25th Annu. Conf. Uncertainty Artif. Intell. (UAI)*, 2009, pp. 647–655.

[280] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf, "Kernel-based conditional independence test and application in causal discovery," in *Proc. 27th Annu. Conf. Uncertainty Artif. Intell. (UAI)*, 2011, pp. 804–813.

[281] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang, "Domain adaptation under target and conditional shift," in *Proc. 30th Int. Conf. Mach. Learn. (ICML)*, 2013, pp. 819–827.

[282] R. Zhang, "Making convolutional networks shift-invariant again," 2019, *arXiv:1904.11486*. [Online]. Available: http://arxiv.org/abs/1904.11486