Towards Causal Representation Learning

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

***Abstract*—The two fields of machine learning and graphical causality arose and developed separately. However, there is now cross-pollination and increasing interest in both fields to benefit from the advances of the other. In the present paper, 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, 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.**

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, 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 [101, 140], natural language processing [54], and speech recognition [85], 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 comes from. In computer vision [75, 228], changes in the test distribution may, for instance, come from aberrations like camera blur, noise or compression quality [106, 129, 170, 206], or from shifts, rotations, or viewpoints [7, 11, 63, 282]. Motivated by this, new benchmarks were proposed to specifically test generalization of classification and detection methods with respect to simple algorithmically generated interventions like spatial shifts, blur, changes in brightness or contrast [106, 170], time consistency [94, 227], control over background and rotation [11], as well as images collected in multiple environments [19]. 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, pre-training, self-supervision, and architectures with suitable inductive biases w.r.t. a perturbation of interest [233, 59, 63, 137, 170, 206]. 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 [183, 237, 218, 34, 188, 181].

b) Issue 2 – Learning Reusable Mechanisms: Infants' understanding of physics relies upon objects that can be tracked over time and behave consistently [52, 236]. Such a representation allows children to quickly learn new tasks as their knowledge and intuitive understanding of physics can be re-used [15, 52, 144, 250]. Similarly, intelligent agents that robustly solve real-world tasks need to re-use and re-purpose 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 [14, 10, 16, 84, 197, 212, 8, 274, 26, 76, 83, 141, 157, 177, 211, 245, 258, 272, 57, 182]. 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 [220, 84]. When learning a causal model, one should thus require fewer examples to adapt as most knowledge, i.e.,

†: equal contribution.
⋆: equal contribution.
B. Schölkopf is at the Max-Planck Institute for Intelligent Systems, Max-Planck-Ring 4, 72076 Tübingen, Germany, bs@tuebingen.mpg.de.
F. Locatello is at ETH Zurich, Computer Science Department and the Max Planck Institute for Intelligent Systems. Work partially done while interning at Google Research Amsterdam. francesco.locatello@gmail.com
S. Bauer is at the Max-Planck Institute for Intelligent Systems, stefan.bauer@tuebingen.mpg.de.
N. R. Ke is at Mila and the University of Montreal, *rosemary.nan.ke@gmail.com*.
N. Kalchbrenner is at Google Research Amsterdam, nalk@google.com.
A. Goyal is at Mila and the University of Montreal, anirudhgoyal9119@gmail.com
Y. Bengio is at Mila, the University of Montreal, CIFAR Senior Fellow *yoshua.bengio@mila.quebec*

modules, can be re-used without further training.

c) A 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* [237, 183]. The manipulative definition of causation [237, 183, 118] focuses on the fact that conditional probabilities ("seeing people with open umbrellas suggests that it is raining") cannot reliably predict the outcome of an 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 an observed data distribution and a set of training tasks, and it extends to situations involving forms of reasoning.

Our Contributions: In the present paper, we argue that causality, with its focus on representing structural knowledge about the data generating process that allows interventions and changes, can contribute towards 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, i.e., 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. The present work reviews and synthesizes key contributions that have been made to this end.

- 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.
- We expand on the Independent Causal Mechanisms (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 hypothesis as a consequence of the ICM principle and discuss its implications for learning causal models.
- We review existing approaches to learn causal relations from appropriate descriptors (or features) in Section V. We cover both classical approaches and modern re-interpretations based on deep neural networks, with a focus on the underlying principles that enable causal discovery.
- 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.
- 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, self-supervised learning, data augmentation, and pre-training. 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.

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

II. Levels of Causal Modeling

The gold standard for modeling natural phenomena is a set of coupled differential equations modeling physical mechanisms responsible for the 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

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

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 $\frac{d\mathbf{x}}{dt}$ (i.e., the derivative only appears on the left-hand side), then its causal structure can be directly read off.[1]

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 prediction of others as long as experimental conditions do not change. E.g., 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 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

[1] 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 [207]), then (2) may not be causally meaningful [221, 190].

TABLE I
A SIMPLE TAXONOMY OF MODELS. THE MOST DETAILED MODEL (TOP) IS A MECHANISTIC OR PHYSICAL ONE, USUALLY IN TERMS OF DIFFERENTIAL EQUATIONS. AT THE OTHER END OF THE SPECTRUM (BOTTOM), WE HAVE A PURELY STATISTICAL MODEL; THIS CAN BE LEARNED FROM DATA, 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

situation is summarized in Table I, adapted from [188]. Below, we address some of the tasks listed in Table I 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 like: "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 under-explored 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 those 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 conditions. Performing an intervention changes the data distribution, which may lead to (arbitrarily) inaccurate predictions [183, 237, 218, 188].

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 to train 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 a 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 as well as understanding in retrospect what led to certain outcomes, at least to some degree of approximation.[2] We have above mentioned the example of statistical predictions of heart failure. 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 below, counterfactuals, or approximations thereof, are especially critical in reinforcement learning. 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, (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.

[2]Note that the two types of questions occupy a continuum: to this end, consider a probability which is both conditional and interventional $P(A|B, do(C))$. If B is the 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 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).

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.

Hand Engineered Data vs. Raw Data: especially in classical AI, data is 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, is unstructured and does 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, speech etc. On the other hand, although methods for learning causal structure from observations exist [237, 188, 229, 113, 174, 187, 139, 17, 246, 277, 175, 123, 186, 176, 36, 82, 161], 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).[3] 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 meta-learning 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 variables depends on the tasks of interest and on the type of data 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 [148, 171, 223, 231, 53]. 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 finally (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 datasets). Alternatively, problems can be made approximately i.i.d., e.g.. by carefully collecting the right training set for a given application problem, or by methods such as "experience replay" [171] where a reinforcement learning agent stores observations in order to later permute them for the purpose of re-training.

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 [257, 217, 239, 66]. 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 [9, 106, 129, 170, 206]. An example of this is the impact of socio-economic factors in clinics in Thailand on the accuracy of a detection system for Diabetic Retinopathy [18]. 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 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 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.[4] 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 dis-

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

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

tribution where certain variables or mechanisms have changed.

B. The Reichenbach Principle: From Statistics to Causality

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

> *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 $X \to Y$. If babies attract storks, it is $X \leftarrow Y$. If there is some other variable that causes both (such as economic development), we have $X \leftarrow Z \to Y$.

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 [187], the case of more observables is surprisingly easier, the reason being that in that case, there are nontrivial conditional independence properties [238, 51, 74] implied by causal structure. These generalize the Reichenbach Principle and can be described by using the language of causal graphs or structural causal models, merging probabilistic graphical models and the notion of interventions [237, 183]. 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 (SCMs)

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

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

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 $U_1, \ldots, U_n$ are 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 non-descendants.

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.

a) 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, e.g.,

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

the factorization (4) yields practical computational advantages during inference, which is in general hard, even when it comes to non-trivial approximations [210]. But more interestingly, it is the only one that decomposes the joint distribution into conditionals corresponding to the structural assignments (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.

b) 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.

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

Several types of interventions may be possible [62] which can be categorized as: *No intervention:* only observational

data is obtained from the causal model. *Hard/perfect:* the function in the structural assignment (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. *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). *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 structural causal model. 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 [207, 267] 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 Figure 1. A statistical model may be defined for instance through a graphical model, i.e., 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 [97]. 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, i.e., if a distribution is Markovian w.r.t. one of the graphs, then it also is w.r.t. 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, cf. below and [237, 183, 188].

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"), cf. (3). This allows to compute interventional distributions as depicted in Figure 1. When a variable is intervened upon, we disconnect it from its parents, fix its value, and perform ancestral sampling on its children.

A structural causal model is composed of (i) a set of causal variables and (ii) 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 to compute interventional distributions, only the SCMs allow to compute 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).

a) 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 (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 to draw conclusions on the effect of interventions, counterfactuals and potential outcomes. In contrast, statistical models only allow to reason about the outcome of i.i.d. experiments.

IV. Independent Causal Mechanisms

We now return to the disentangled factorization (4) of the joint distribution $P(X_1, \ldots, X_n)$. This factorization according to the causal graph is always possible when the U_i are independent, but we will now consider an additional notion of independence relating the factors in (4) 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 3D structures lining up in 2D, 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 3D information even without stereo vision ("structure from motion").

For another example, consider a dataset 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 the altitude has a causal effect on temperature. Suppose we had two such datasets, 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 $A \to T$. 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.

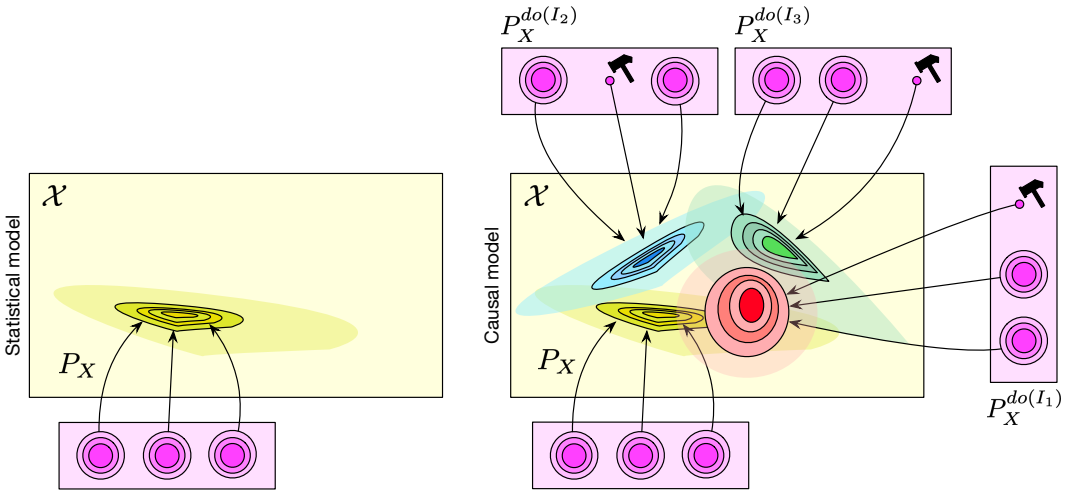

Fig. 1. Difference between statistical (left) and causal models (right) on a given set of three variables. While a statistical model specifies a single probability distribution, a causal model represents a set of distributions, one for each possible intervention (indicated with a 🔨 in the figure).

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 [218, 188]:

> **Independent Causal Mechanisms (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 an independence between the cause distribution and the mechanism producing the effect distribution.

Applied to the causal factorization (4), the principle tells us that the factors should be independent in the sense that

(a) changing (or performing an intervention upon) one mechanism $P(X_i|\mathbf{PA}_i)$ does not change any of the other mechanisms $P(X_j|\mathbf{PA}_j)$ $(i \neq j)$ [218], and
(b) knowing some other mechanisms $P(X_i|\mathbf{PA}_i)$ $(i \neq j)$ does not give us information about a mechanism $P(X_j|\mathbf{PA}_j)$ [120].

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 [99, 71, 111, 183, 120, 240, 188]. Early work on this was done by Haavelmo [99], stating the assumption that changing one of the structural assignments leaves the other ones invariant. Hoover [111] attributes to Herb Simon the *invariance criterion*: the true causal order is the one that is invariant under the right sort of intervention. Aldrich [4] discusses the historical development of these ideas in economics. He argues that the "most basic question one can ask about a relation should be: How autonomous is it?" [71, preface]. Pearl [183] discusses autonomy in detail, arguing that a causal mechanism remains invariant when other mechanisms are subjected to external influences. He points 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, Sec. 2.2]. These seemingly different notions can be unified [120, 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 (a) of the ICM Principle, we state the following hypothesis:

> **Sparse Mechanism Shift (SMS).** *Small distribution changes tend to manifest themselves in a sparse or local way in the causal/disentangled factorization (4), i.e., they should usually not affect all factors simultaneously.*

In contrast, if we consider a non-causal factorization, e.g., (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 [23, 109, 158, 247].

The SMS hypothesis was stated in [181, 24, 221, 115], and in earlier form in [218, 279, 220]. An intellectual ancestor is Simon's invariance criterion, i.e., 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 [69, 270]. It has recently been used for

learning causal models [131], modular architectures [84, 28] 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 (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 non-deterministic,[5] 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 (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 Kolmogorov complexity [120]. 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; i.e., two conditionals are considered independent if knowing (the shortest compression of) one does not help us achieve a shorter compression of the other.

Algorithmic information theory provides a natural framework for non-statistical graphical models [120, 126]. Just like the latter are obtained from structural causal models 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 structural causal model [120]. 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.

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

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[6] from observational data by performing conditional independence tests. However, there are several problems with this approach. One is that our datasets are always finite in practice, and conditional independence testing is a notoriously difficult problem, especially if conditioning sets are continuous and multi-dimensional. So 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 datasets, 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 [90, 91, 73, 278, 60, 191, 42]. 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, i.e., approaching minimal expected error in the infinite sample limit, there are always cases where this convergence is arbitrarily slow. So for a given sample size, it will depend on the problem being learned whether we achieve low expected error, and statistical learning theory provides probabilistic guarantees in terms of measures of complexity of function classes [55, 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 $X \to Y$. 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 non-smooth way, it should be hard to glean information about the SCM from a finite dataset, given that V is not observed and its value randomly selects among arbitrarily different f_v.

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

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) [113, 174, 187, 139, 17], cf. also [246]. This is subject to certain genericity assumptions, and notable exceptions include the case where U, V are Gaussian and f is linear. It generalizes results of Shimizu et al. [229] for linear functions, and it can be generalized to include non-linear rescalings [277], loops [175], confounders [123], and multi-variable settings [186]. Empirically, there is a number of methods that can detect causal direction better than chance [176], some of them building on the above Kolmogorov complexity model [36], some on generative models [82], and some directly learning to classify bivariate distributions into causal vs. anticausal [161].

While restrictions of function classes are one possibility to allow to identify 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 [251, 188]. These contexts can come from interventions [218, 189, 192], non-stationary time series [117, 100, 193] or multiple views [89, 115]. The contexts can likewise be interpreted as different tasks, which provide a connection to meta-learning [22, 67, 213].

The work of Bengio et al. [24] ties the generalization in meta-learning 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 [24, 131] focus on learning a causal model using neural networks with an unsupervised loss, the work of Dasgupta et al. [50] explores learning a causal model using a reinforcement learning 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, e.g., 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 (cf. Bonet and Geffner [33]). 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 non-linear function. An example can be seen in Figure 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 [41, 207]. 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 [207], ordinary differential equations [173, 208], and temporally aggregated time series [78]. The task of identifying suitable units that admit causal models is challenging for both human and machine intelligence. Still, it aligns with the general goal of modern machine learning to learn meaningful representations of data, where meaningful can include *robust*, *explainable*, or *fair* [142, 133, 276, 130, 260].

To combine structural causal modeling (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 re-purposed for new tasks [181, 84] and the SMS hypothesis can be used to enforce the appropriate structure. We visualize an example in Figure 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.

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

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

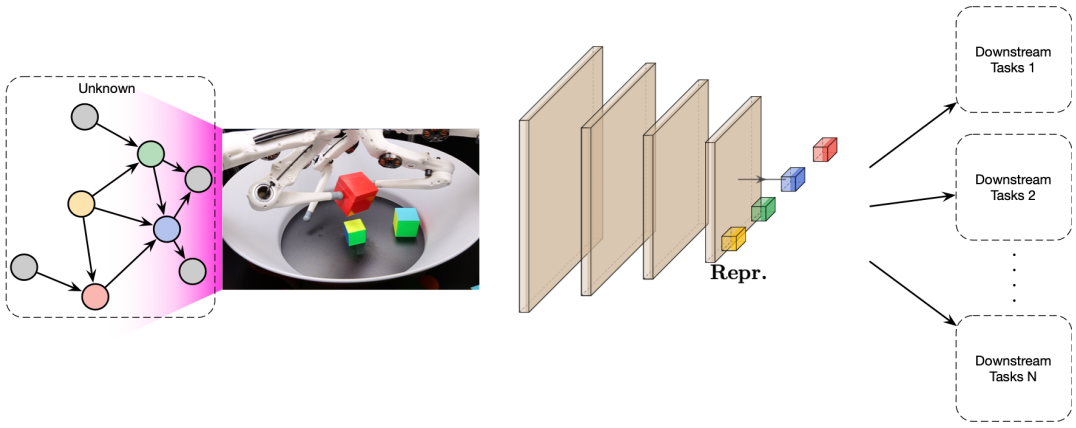

Fig. 2. Illustration of the causal representation learning problem setting. Perceptual data, such as images or other high-dimensional sensor measurements, can be thought of as entangled views of the state of an unknown causal system as described in (10). With the exception of possible task labels, none 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).

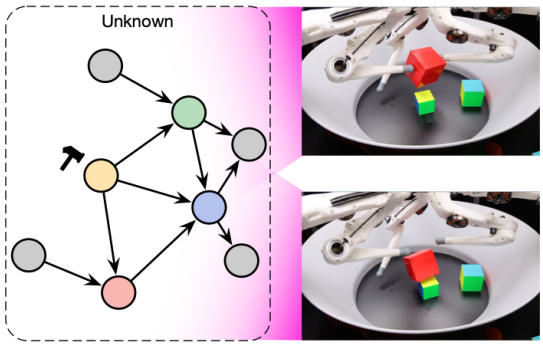

Fig. 3. Example of the SMS hypothesis where an intervention (which may or may not be intentional/observed) changes the position of one finger (↖), and as a consequence, the object falls. The change in pixel space is entangled (or distributed), in contrast to the change in the causal model.

as well as the property that the conditionals $P(S_i \mid \mathbf{PA}_i)$ be independently manipulable and largely invariant across related problems. Suppose we seek to reconstruct such a *disentangled representation using independent mechanisms* (11) from data, but the causal variables S_i are not provided to us a priori. Rather, we are given (possibly high-dimensional) $X = (X_1, \ldots, X_d)$ (below, 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, cf. (3),

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

modeling the causal relationships among the 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 [109, 158, 159, 256, 157, 135, 202, 61] focuses on independent factors of variation. This can be viewed as the special case where the causal graph is trivial, i.e., $\forall i : \mathbf{PA}_i = \emptyset$ in (12). In this case, the factors are functions of the independent exogenous noise variables, and thus independent themselves.[7] However, the ICM Principle is more general and contains statistical independence as a special case.

Note that the problem of *object-centric* representation learning [10, 39, 83, 86, 87, 138, 155, 160, 262, 255] 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[8], 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 distribu-

[7]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.

[8]Objects can be represented at different levels of granularity [207], i.e. as a single entity or as a composition of other causal variables encoding parts, properties, and other factors of variation.

tion [158, 116, 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, i.e., we should make the 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 [221, 21, 29]. Locatello et al. [159] showed that the sparse mechanism shift hypothesis stated above is theoretically sufficient when given suitable training data. Further, 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 [230, 159]. As discussed by Schölkopf et al. [220], 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, i.e., we only have limited data for each task/domain, and thus need to find ways of pooling/re-using 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 re-purposed. 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 re-use components, which requires them to be robust across tasks and environments [220]. An elegant way to do this is to employ a modular structure that mirrors a 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 [84]. For instance, if variations of natural lighting (the position of the sun, clouds, etc.) imply that the visual environment can appear in brightness conditions spanning several orders of magnitude, then visual processing algorithms in our nervous system should employ methods that can factor out these variations, rather than building separate sets of 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 towards 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 signals. A sensible inductive bias to learn such models is to look for independent causal mechanisms [180] and competitive training can play a role in this. For pattern recognition tasks, [181, 84] suggest 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 [23, 148]. However, it does so without taking into account the causal properties of the variables, i.e., it does not care about the interventional properties of the variables it analyzes or reconstructs. Causal representation learning should move beyond the representation of statistical dependence structures towards 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 which has not yet entered the stage of machine learning research although it is at the core of human intelligence [107].

VII. Implications for Machine Learning

All of this discussion calls 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: (a) in many cases, we need to infer abstract causal variables from the available low-level input features; (b) there is no consensus on which aspects of the data reveal causal relations; (c) 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 environment information and interventions; (d) even in the limited cases we understand, we often lack scalable and numerically sound algorithms. Despite these challenges, we argue this endeavor has concrete implications for machine learning and may shed light on desiderata and current practices alike.

A. Semi-Supervised Learning (SSL)

Suppose our underlying causal graph is $X \to Y$, and at the same time we are trying to learn a mapping $X \to Y$. 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. [49] who define a measure of dependence between input $P(X)$ and conditional $P(Y|X)$.[9] 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 of output (cause) given input, i.e., the quantity that machine learning is usually concerned with.

The study [218] 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 [279, 266, 280, 77, 114, 281, 32, 96, 263, 243, 195, 152, 156, 153, 167, 204, 115], complementing the studies of Bareinboim and Pearl [12, 185], and it inspired a thread of work in the statistics community exploiting invariance for causal discovery and other tasks [189, 192, 105, 104, 115].

On the SSL side, subsequent developments include further theoretical analyses [121, 188, Section 5.1.2] and a form of conditional SSL [259]. The view of SSL as exploiting dependencies between a marginal $P(X)$ and a non-causal conditional $P(Y|X)$ is consistent with the common assumptions employed to justify SSL [44]. 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 $P(X)$; and the *semi-supervised smoothness assumption*, applicable also to continuous outputs, states that if two points in a high-density region are close, then 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 *co-training theorem* [31] makes a statement about learnability from unlabelled data, and 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, i.e., 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 which 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 a 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. Clearly, 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 which we might call a confounded one). In a 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 [216, 134].[10] 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 which 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 [95].

C. Robustness and Strong Generalization

We can speculate that structures composed of autonomous modules, such as given by a causal factorization (4), should be relatively robust to swapping out or modifying individual components. Robustness should also play a role when studying *strategic behavior*, i.e., 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

[9] Other dependence measures have been proposed for high-dimensional linear settings and time series [124, 226, 27, 122, 119, 125].

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

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, 204, 169, 189, 218]. 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 small expected risk for a different distribution $P^\dagger(X,Y)$:

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

Clearly, 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, experimental conditions, etc. Environments can be modeled in a causal factorization (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 (for example 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 [20]) to learn statistical models that are stable across training environments and generalize to novel testing environments [189, 204, 167, 5, 2] 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, i.e., $P^\dagger(X,Y) \in \mathbb{P}_\mathcal{G}$ 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, 20].

D. Pre-training, 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 dataset [244, 53], for example, through pre-training on a huge and diverse corpus [196, 54, 112, 137, 59, 35, 45, 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 [9, 234, 140]. 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, etc. 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 $\mathcal{E}$ the model should be robust to (e.g., random crops/interpolations/translation/rotations, etc). 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 [38], using a search algorithm like reinforcement learning [48] or an algorithm based on density matching [154].

The third approach is to rely on self-supervision to learn about $P(X)$. Certain pre-training methods [196, 54, 112, 35, 45, 253] have shown that it is possible to achieve good results using only very few class labels by first pre-training on a large unlabeled dataset and then fine-tuning on few labeled examples. Similarly, pre-training on large unlabeled image datasets can improve performance by learning representations that can efficiently transfer to a downstream task, as demonstrated by [179, 110, 102, 46, 92]. 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 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 [79]. In that case,

$\mathbb{P}_{\mathcal{G}}$ would contain a set of attacks 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

Reinforcement Learning (RL) is closer to causality research than the machine learning mainstream in that it sometimes effectively directly estimates do-probabilities. E.g., 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 [164, 81]. An emerging line of work devoted to the intersection of RL and causality includes [13, 21, 164, 37, 50, 275, 1]. Causal learning applied to reinforcement learning 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, 26, 258].

a) World Models: Model-based RL [248, 67] 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 [127, 248, 98, 47, 271, 178, 232, 214, 268]. 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 [58, 265, 43, 264, 14, 136]. However, many of the current approaches (regardless of structure), only build partial models of the environment [88]. 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].

b) 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. Further, 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 towards 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 [80]. This can be modelled as an RL environment, where the agent can discover causal factors through interventions and observing their effects. Further, causal models may allow to model the environment as a set of underlying independent causal mechanisms such that, if there is a change in distribution, not all the mechanisms need to be re-learned. 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.

c) Counterfactuals: Counterfactual reasoning has been found to improve the data efficiency of RL algorithms [37, 165], improve performance [50], and it has been applied to communicate about past experiences in the multi-agent setting [68, 241]. These findings are consistent with work in cognitive psychology [64], arguing that counterfactuals allow to reason about the usefulness of past actions and transfer these insights to corresponding behavioral intentions in future scenarios [203, 199, 145].

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.

d) Offline RL: The success of deep learning methods in the case of supervised learning can be largely attributed to the availability of large datasets and methods that can scale to large amounts of data. In the case of reinforcement learning, collecting large amounts of high-fidelity diverse data from scratch can be expensive and hence becomes a bottleneck. Offline RL [72, 150] tries to address this concern by learning a policy from a *fixed* dataset 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 offline data makes offline 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 independent causal mechanisms can help make progress towards the offline 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 [93], which can substantially increase the efficiency of hand-engineered simulators [103, 143, 269, 211, 264]. 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 re-trained if the conditions change.

On the other hand, the lack of systematic experimental conditions may become problematic in other applications such as healthcare. One example is personalized medicine, where we may wish to build a model of a patient health state through a multitude of data sources, like electronic health records and genetic information [65, 108]. 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 [18]. 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 [201, 242, 224, 273, 6, 3, 30, 165].

Causality also has significant potential in helping understand medical phenomena, e.g., in the current Covid-19 pandemic, where causal mediation analysis helps disentangle different effects contributing towards case fatality rates when a textbook example of Simpson's paradox was observed [261].

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 [219], which is critical especially in the context of partial technical failures as experienced in the Kepler exoplanet search mission. The application of [219] lead to the discovery of 36 planet candidates [70], 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, i.e., allowing for liquid water [25, 254]. This planet turned out to be one that had first been detected in [70, exoplanet candidate EPIC 201912552].

G. Multi-Task Learning and Continual Learning

State-of-the-art AI is relatively *narrow*, i.e., 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.

Multi-task learning refers to building a system that can solve multiple tasks across different environments [40, 209]. These tasks usually share some common traits. By learning similarities across tasks, a system could utilize 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 [220]. 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 multi-task 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 [35] to beating human world champions in games such as chess, poker and Go [223], improving medical diagnoses [166], and generating music [56]. A critical question thus arises: *"Why can't we just train a huge model that learns environments' dynamics (e.g. in a 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 was 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 datasets, 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 Independent Causal Mechanisms principle. When these assumptions are valid, a learner that does not use them should fare worse than one that does. Further, 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 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. Until now, machine learning has neglected a full integration of causality, and this paper argues that it would indeed benefit from integrating causal concepts. We argue that combining the strengths of both fields, i.e., current deep learning methods as well as tools and ideas from causality, may be a necessary step on the path towards 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 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 semi-supervised learning, domain generalization, and adversarial robustness.

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

a) Learning Non-Linear Causal Relations at Scale: Not all real-world data is 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 non-linear 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 meta- and multi-task learning are close to the assumptions and desiderata of causal modeling, and future work should consider (1) understanding under which conditions non-linear 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 (non-causal) statistical representations in terms of generalization, re-purposing, 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. Further, structured and modular representation should also arise when a network is trained for (sets of) specific tasks, not only auteoncoding. 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 we are robust to in predictive tasks.

c) Understanding the Biases of Existing Deep Learning Approaches: Scaling to massive data sets, 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 datasets in order to build a taxonomy of inductive biases. In particular, it is desirable to understand how design choices in pre-training (e.g., which datasets/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 reinforcement learning (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 reinforcement learning 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 reinforcement learning.

IX. Acknowledgments

Many thanks to 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 [221]. The text has also benefitted from discussions with Elias Bareinboim, Christoph Bohle, Leon Bottou, Isabelle Guyon, Judea Pearl, and Vladimir Vapnik. Thanks to Wouter van Amsterdam for pointing out typos in the first version. We also thank Thomas Kipf, Klaus Greff, and Alexander d'Amour for the useful discussions. Finally, we thank the thorough anonymous reviewers for highly valuable feedback and suggestions.

References

[1] Ossama Ahmed, Frederik Träuble, Anirudh Goyal, Alexander Neitz, Manuel Wuthrich, Yoshua Bengio, Bernhard Schölkopf, and Stefan Bauer. Causalworld: A robotic manipulation benchmark for causal structure and transfer learning. In *International Conference on Learning Representations*, 2021.

[2] Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving rubik's cube with a robot hand. *arXiv preprint 1910.07113*, 2019.

[3] Ahmed Alaa and Mihaela Schaar. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *International Conference on Machine Learning*, pages 129–138, 2018.

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

[5] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint 1907.02893*, 2019.

[6] Onur Atan, James Jordon, and Mihaela van der Schaar. Deep-treat: Learning optimal personalized treatments from observational data using neural networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[7] Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, 20(184):1–25, 2019.

[8] Dzmitry Bahdanau, Shikhar Murty, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries, and Aaron Courville. Systematic generalization: what is required and can it be learned? *arXiv preprint 1811.12889*, 2018.

[9] H. Baird. Document image defect models. In *Proc., IAPR Workshop on Syntactic and Structural Pattern Recognition*, pages 38–46, Murray Hill, NJ, 1990.

[10] Victor Bapst, Alvaro Sanchez-Gonzalez, Carl Doersch, Kimberly Stachenfeld, Pushmeet Kohli, Peter Battaglia, and Jessica Hamrick. Structured agents for physical construction. In *International Conference on Machine Learning*, pages 464–474, 2019.

[11] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, pages 9448–9458, 2019.

[12] E. Bareinboim and J. Pearl. Transportability from multiple environments with limited experiments: Completeness results. In *Advances in Neural Information Processing Systems 27*, pages 280–288, 2014.

[13] E. Bareinboim, A. Forney, and J. Pearl. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems 28*, pages 1342–1350, 2015.

[14] Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. In *Advances in neural information processing systems*, pages 4502–4510, 2016.

[15] Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45): 18327–18332, 2013.

[16] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint 1806.01261*, 2018.

[17] S. Bauer, B. Schölkopf, and J. Peters. The arrow of time in multivariate time series. In *Proceedings of the 33nd International Conference on Machine Learning*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 2043–2051, 2016.

[18] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[19] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473, 2018.

[20] S. Ben-David, T. Lu, T. Luu, and D. Pál. Impossibility theorems for domain adaptation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics 13 (AISTATS)*, pages 129–136, 2010.

[21] Emmanuel Bengio, Valentin Thomas, Joelle Pineau, Doina Precup, and Yoshua Bengio. Independently controllable features. *arXiv preprint 1703.07718*, 2017.

[22] Yoshua Bengio, Samy Bengio, and Jocelyn Cloutier. *Learning a synaptic learning rule*. IJCNN-91-Seattle International Joint Conference on Neural Networks (Vol. 2, pp. 969-vol). IEEE., 1990.

[23] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *arXiv preprint 1206.5538*, 2012.

[24] Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint 1901.10912*, 2019.

[25] Björn Benneke, Ian Wong, Caroline Piaulet, Heather A. Knutson, Ian J. M. Crossfield, Joshua Lothringer, Caroline V. Morley, Peter Gao, Thomas P. Greene, Courtney Dressing, Diana Dragomir, Andrew W. Howard, Peter R. McCullough, Eliza M. R. Kempton Jonathan J. Fortney, and Jonathan Fraine. Water vapor on the habitable-zone exoplanet K2-18b. *arXiv preprint 1909.04642*, 2019.

[26] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dkebiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint 1912.06680*, 2019.

[27] M. Besserve, N. Shajarisales, B. Schölkopf, and D. Janzing. Group invariance principles for causal generative models. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 557–565, 2018.

[28] M. Besserve, R. Sun, D. Janzing, and B. Schölkopf. A theory of independent mechanisms for extrapolation in generative models. In *35th AAAI Conference on Artificial Intelligence: A Virtual Conference*, February 2021.

[29] Michel Besserve, Rémy Sun, and Bernhard Schölkopf. Counterfactuals uncover the modular structure of deep generative models. *arXiv preprint 1812.03253, published at ICLR 2020*, 2018.

[30] Ioana Bica, Ahmed M Alaa, and Mihaela van der Schaar. Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders. *arXiv preprint 1902.00450*, 2019.

[31] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 92–100, New York, NY, USA, 1998. ACM.

[32] Patrick Blöbaum, Takashi Washio, and Shohei Shimizu. Error asymmetry in causal and anticausal regression. *arXiv preprint 1610.03263*, 2016.

[33] Blai Bonet and Hector Geffner. Learning first-order symbolic representations for planning from the structure of the state space. *arXiv preprint 1909.05546*, 2019.

[34] L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugualy, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14:3207–3260, 2013.

[35] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint 2005.14165*, 2020.

[36] Kailash Budhathoki and Jilles Vreeken. Causal inference by compression. In *IEEE 16th International Conference on Data Mining*, 2016.

[37] Lars Buesing, Theophane Weber, Yori Zwols, Sebastien Racaniere, Arthur Guez, Jean-Baptiste Lespiau, and Nicolas Heess. Woulda, coulda, shoulda: Counterfactually-guided policy search. *arXiv preprint 1811.06272*, 2018.

[38] C. J. C. Burges and B. Schölkopf. Improving the accuracy and speed of support vector learning machines. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, pages 375–381, Cambridge, MA, USA, 1997. MIT Press.

[39] Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. *arXiv preprint 1901.11390*, 2019.

[40] Rich Caruana. Multitask learning. *Machine learning*, 28(1): 41–75, 1997.

[41] Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Multi-level cause-effect systems. *arXiv preprint 1512.07942*, 2015.

[42] Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Fast conditional independence test for vector variables with large sample sizes. *arXiv preprint 1804.02747*, 2018.

[43] Michael B Chang, Tomer Ullman, Antonio Torralba, and Joshua B Tenenbaum. A compositional object-based approach to learning physical dynamics. In *5th International Conference on Learning Representations (ICLR)*, 2017.

[44] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, USA, 2006. URL http://www.kyb.tuebingen.mpg.de/ssl-book/.

[45] Mark Chen, Alec Radford, Rewon Child, Jeff Wu, Heewoo Jun, Prafulla Dhariwal, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[46] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint 2002.05709*, 2020.

[47] Silvia Chiappa, Sébastien Racaniere, Daan Wierstra, and Shakir Mohamed. Recurrent environment simulators. In *5th International Conference on Learning Representations (ICLR)*, 2017.

[48] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 113–123, 2019.

[49] P. Daniušis, D. Janzing, J. M. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf. Inferring deterministic causal relations. In *Proceedings of the 26th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 143–150, 2010.

[50] Ishita Dasgupta, Jane Wang, Silvia Chiappa, Jovana Mitrovic, Pedro Ortega, David Raposo, Edward Hughes, Peter Battaglia, Matthew Botvinick, and Zeb Kurth-Nelson. Causal reasoning from meta-reinforcement learning. *arXiv preprint 1901.08162*, 2019.

[51] A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society B*, 41(1):1–31, 1979.

[52] Stanislas Dehaene. *How We Learn: Why Brains Learn Better Than Any Machine... for Now*. Penguin, 2020.

[53] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[54] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint 1810.04805*, 2018.

[55] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Applications of Mathematics*. Springer, New York, NY, 1996.

[56] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint 2005.00341*, 2020.

[57] Andrea Dittadi, Frederik Träuble, Francesco Locatello, Manuel Wüthrich, Vaibhav Agrawal, Ole Winther, Stefan Bauer, and Bernhard Schölkopf. On the transfer of disentangled representations in realistic settings. In *International Conference on Learning Representations*, 2021.

[58] Carlos Diuk, Andre Cohen, and Michael L Littman. An object-oriented representation for efficient reinforcement learning. In *Proceedings of the 25th international conference on Machine learning*, pages 240–247, 2008.

[59] Josip Djolonga, Jessica Yung, Michael Tschannen, Rob Romijnders, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Matthias Minderer, Alexander D'Amour, Dan Moldovan, et al. On robustness and transferability of convolutional neural networks. *arXiv preprint 2007.08558*, 2020.

[60] G. Doran, K. Muandet, K. Zhang, and B. Schölkopf. A permutation-based kernel conditional independence test. In N. L. Zhang and J. Tian, editors, *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, pages 132–141, Corvallis, OR, 2014. AUAI Press. URL http://auai.org/uai2014/proceedings/individuals/194.pdf.

[61] Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

[62] Daniel Eaton and Kevin Murphy. Exact Bayesian structure learning from uncertain interventions. In *Artificial Intelligence and Statistics*, pages 107–114, 2007.

[63] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. *arXiv preprint 1712.02779*, 2017.

[64] Kai Epstude and Neal J Roese. The functional theory of counterfactual thinking. *Personality and social psychology review*, 12(2):168–192, 2008.

[65] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24–29, 2019.

[66] András Faragó and Gábor Lugosi. Strong universal consistency of neural network classifiers. *IEEE Transactions on Information Theory*, 39(4):1146–1151, 2006.

[67] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint 1703.03400*, 2017.

[68] Jakob N Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[69] Peter Földiák. Learning invariance from transformation sequences. *Neural Computation*, 3(2):194–200, 1991.

[70] 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. *The Astrophysical Journal*, 806(2), 2015. URL http://stacks.iop.org/0004-637X/806/i=2/a=215.

[71] R. Frisch, T. Haavelmo, T.C. Koopmans, and J. Tinbergen. *Autonomy of economic relations*. Universitets Socialøkonomiske Institutt, Oslo, Norway, 1948.

[72] Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pages 2052–2062, 2019.

[73] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems 20*, pages 489–496, 2008.

[74] D. Geiger and J. Pearl. Logical and algorithmic properties of independence and their application to Bayesian networks. *Annals of Mathematics and Artificial Intelligence*, 2:165–178, 1990.

[75] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint 1811.12231*, 2018.

[76] Muhammad Waleed Gondal, Manuel Wüthrich, Djordje Miladinović, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. In *Advances in Neural Information Processing Systems*, pages 15740–15751, 2019.

[77] M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour, and B. Schölkopf. Domain adaptation with conditional transferable components. In *Proceedings of the 33nd International Conference on Machine Learning*, pages 2839–2848, 2016.

[78] M. Gong, K. Zhang, B. Schölkopf, C. Glymour, and D. Tao. Causal discovery from temporally aggregated time series. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence (UAI)*, page ID 269, 2017.

[79] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint 1412.6572*, 2014.

[80] Alison Gopnik, Clark Glymour, David M Sobel, Laura E Schulz, Tamar Kushnir, and David Danks. A theory of causal learning in children: causal maps and Bayes nets. *Psychological review*, 111(1):3, 2004.

[81] Omer Gottesman, Fredrik Johansson, Joshua Meier, Jack Dent, Donghun Lee, Srivatsan Srinivasan, Linying Zhang, Yi Ding, David Wihl, Xuefeng Peng, Jiayu Yao, Isaac Lage, Christopher Mosch, Li wei H. Lehman, Matthieu Komorowski, Matthieu Komorowski, Aldo Faisal, Leo Anthony Celi, David Sontag, and Finale Doshi-Velez. Evaluating reinforcement learning algorithms in observational health settings. *arXiv preprint 1805.12298*, 2018.

[82] Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michèle Sebag. Causal generative neural networks. *arXiv preprint 1711.08936*, 2017.

[83] Anirudh Goyal, Alex Lamb, Phanideep Gampa, Philippe Beaudoin, Sergey Levine, Charles Blundell, Yoshua Bengio, and Michael Mozer. Object files and schemata: Factorizing declarative and procedural knowledge in dynamical systems. *arXiv preprint 2006.16225*, 2020.

[84] Anirudh Goyal, Alex Lamb, Jordan Hoffmann, Shagun Sodhani, Sergey Levine, Yoshua Bengio, and Bernhard Schölkopf. Recurrent independent mechanisms. In *International Conference on Learning Representations*, 2021.

[85] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE, 2013.

[86] Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *International Conference on Machine Learning*, pages 2424–2433, 2019.

[87] Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv preprint 2012.05208*, 2020.

[88] Karol Gregor, Danilo Jimenez Rezende, Frederic Besse, Yan Wu, Hamza Merzic, and Aaron van den Oord. Shaping belief states with generative environment models for rl. In *Advances in Neural Information Processing Systems*, pages 13475–13487, 2019.

[89] Luigi Gresele, Paul K Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ica. *arXiv preprint 1905.06642*, 2019.

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

[91] A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 2005.

[92] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint 2006.07733*, 2020.

[93] Radek Grzeszczuk, Demetri Terzopoulos, and Geoffrey Hinton. Neuroanimator: Fast neural network emulation and control of physics-based models. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 9–20, 1998.

[94] Keren Gu, Brandon Yang, Jiquan Ngiam, Quoc Le, and Jonathan Shlens. Using videos to evaluate image model robustness. *arXiv preprint 1904.10076*, 2019.

[95] Shixiang Gu and Luca Rigazio. Towards deep neural network architectures robust to adversarial examples, 2014. arXiv:1412.5068.

[96] Ruocheng Guo, Lu Cheng, Jundong Li, P. Richard Hahn, and Huan Liu. A survey of learning causality with data: Problems and methods. *arXiv preprint 1809.09337*, 2018.

[97] I. Guyon, D. Janzing, and B. Schölkopf. Causality: Objectives and assessment. In I. Guyon, D. Janzing, and B. Schölkopf, editors, *JMLR Workshop and Conference Proceedings: Volume 6*, pages 1–42, Cambridge, MA, USA, 2010. MIT Press.

[98] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint 1803.10122*, 2018.

[99] T. Haavelmo. The probability approach in econometrics. *Econometrica*, 12:S1–S115 (supplement), 1944.

[100] Hermanni Hälvä and Aapo Hyvärinen. Hidden markov nonlinear ica: Unsupervised learning from nonstationary time series. *arXiv preprint 2006.12107*, 2020.

[101] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[102] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[103] Siyu He, Yin Li, Yu Feng, Shirley Ho, Siamak Ravanbakhsh, Wei Chen, and Barnabás Póczos. Learning to predict the cosmological structure formation. *Proceedings of the National Academy of Sciences*, 116(28):13825–13832, 2019.

[104] Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *arXiv preprint 1710.11469*, 2017.

[105] Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *arXiv preprint 1706.08576*, 2017.

[106] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint 1903.12261*, 2019.

[107] Joseph Henrich. *The Secret of our Success*. Princeton University Press, 2016.

[108] Katharine E Henry, David N Hager, Peter J Pronovost, and Suchi Saria. A targeted real-time early warning score (trewscore) for septic shock. *Science translational medicine*, 7(299): 299ra122–299ra122, 2015.

[109] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2016.

[110] R Devon Hjelm and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pages 15535–15545, 2019.

[111] K. D. Hoover. Causality in economics and econometrics. In S. N. Durlauf and L. E. Blume, editors, *The New Palgrave Dictionary of Economics*. Palgrave Macmillan, Basingstoke, UK, 2nd edition, 2008.

[112] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint 1801.06146*, 2018.

[113] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21 (NIPS)*, pages 689–696, 2009.

[114] B. Huang, K. Zhang, J. Zhang, R. Sanchez-Romero, C. Glymour, and B. Schölkopf. Behind distribution shift: Mining driving forces of changes and causal arrows. In *IEEE 17th International Conference on Data Mining (ICDM 2017)*, pages 913–918, 2017.

[115] Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53, 2020. URL http://jmlr.org/papers/v21/19-232.html.

[116] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

[117] AJ Hyvarinen and Hiroshi Morioka. Nonlinear ica of temporally dependent stationary sources. In *Proceedings of Machine Learning Research*, 2017.

[118] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[119] D. Janzing. Causal regularization. In *Advances in Neural Information Processing Systems 33*, 2019.

[120] D. Janzing and B. Schölkopf. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

[121] D. Janzing and B. Schölkopf. Semi-supervised interpolation in an anticausal learning scenario. *Journal of Machine Learning Research*, 16:1923–1948, 2015.

[122] D. Janzing and B. Schölkopf. Detecting non-causal artifacts in multivariate linear regression models. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 2250–2258, 2018.

[123] D. Janzing, J. Peters, J. M. Mooij, and B. Schölkopf. Identifying confounders using additive noise models. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 249–257, 2009.

[124] D. Janzing, P. Hoyer, and B. Schölkopf. Telling cause from effect based on high-dimensional observations. In J. Fürnkranz and T. Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning*, pages 479–486, 2010.

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

[126] D. Janzing, R. Chaves, and B. Schölkopf. Algorithmic independence of initial condition and dynamical law in thermodynamics and causal inference. *New Journal of Physics*, 18(9), 2016. URL http://stacks.iop.org/1367-2630/18/i=9/a=093052.

[127] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

[128] Daniel Kahneman. *Thinking, fast and slow*. Farrar, Straus and Giroux, New York, 2011. ISBN 9780374275631 0374275637. URL https://www.amazon.de/Thinking-Fast-Slow-Daniel-Kahneman/dp/0374275637/ref=wl_it_dp_o_pdT1_nS_nC?ie=UTF8&colid=151193SNGKJT9&coliid=I3OCESLZCVDFL7.

[129] Samil Karahan, Merve Kilinc Yildirum, Kadir Kirtac, Ferhat Sukru Rende, Gultekin Butun, and Hazim Kemal Ekenel. How image degradations affect deep cnn-based face recognition? In *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5. IEEE, 2016.

[130] Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *arXiv 2006.06831*, 2020. Published at NeurIPS.

[131] Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Bernhard Schölkopf, Michael Mozer, Chris Pal, and Yoshua Bengio. Learning neural causal models from unknown interventions. *arXiv preprint 1910.01075v2*, 2020.

[132] Moein Khajehnejad, Behzad Tabibian, Bernhard Schölkopf, Adish Singla, and Manuel Gomez-Rodriguez. Optimal decision making under strategic behavior. *arXiv preprint 1905.09239*, 2019.

[133] N. Kilbertus, M. Rojas Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems 30*, pages 656–666, 2017.

[134] Niki Kilbertus, Giambattista Parascandolo, and Bernhard Schölkopf. Generalization in anti-causal learning. *arXiv preprint 1812.00524*, 2018.

[135] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, 2018.

[136] Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *International Conference on Machine Learning*, pages 2688–2697, 2018.

[137] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. *arXiv preprint 1912.11370*, 2019.

[138] Adam Kosiorek, Hyunjik Kim, Yee Whye Teh, and Ingmar Posner. Sequential attend, infer, repeat: Generative modelling of moving objects. *Advances in Neural Information Processing Systems*, 31:8606–8616, 2018.

[139] S. Kpotufe, E. Sgouritsa, D. Janzing, and B. Schölkopf. Consistency of causal inference under the additive noise model. In *Proceedings of the 31th International Conference on Machine Learning*, pages 478–486, 2014.

[140] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[141] Tejas D Kulkarni, Ankush Gupta, Catalin Ionescu, Sebastian Borgeaud, Malcolm Reynolds, Andrew Zisserman, and Volodymyr Mnih. Unsupervised learning of object keypoints for perception and control. In *Advances in Neural Information Processing Systems*, pages 10723–10733, 2019.

[142] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems 30*, pages 4066–4076. Curran Associates, Inc., 2017.

[143] L'ubor Ladickỳ, SoHyeon Jeong, Barbara Solenthaler, Marc Pollefeys, and Markus Gross. Data-driven fluid simulations using regression forests. *ACM Transactions on Graphics (TOG)*, 34(6):1–9, 2015.

[144] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

[145] Janet Landman, Elizabeth A Vandewater, Abigail J Stewart, and Janet E Malley. Missed opportunities: Psychological ramifications of counterfactual thought in midlife women. *Journal of Adult Development*, 2(2):87–97, 1995.

[146] Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In Marco Wiering and Martijn van Otterlo, editors, *Reinforcement Learning: State-of-the-Art*, pages 45–73. Springer, Berlin, Heidelberg, 2012.

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

[148] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[149] Felix Leeb, Yashas Annadani, Stefan Bauer, and Bernhard Schölkopf. Structural autoencoders improve representations for generation and transfer. *arXiv preprint 2006.07796*, 2020.

[150] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint 2005.01643*, 2020.

[151] David Lewis. Causation. *The journal of philosophy*, 70(17): 556–567, 1974.

[152] Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representation. *arXiv preprint 1807.08479*, 2018.

[153] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *The European Conference on Computer Vision (ECCV)*, 2018.

[154] Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim. Fast autoaugment. In *Advances in Neural Information Processing Systems*, pages 6665–6675, 2019.

[155] Zhixuan Lin, Yi-Fu Wu, Skand Vishwanath Peri, Weihao Sun, Gautam Singh, Fei Deng, Jindong Jiang, and Sungjin Ahn. Space: Unsupervised object-oriented scene representation via spatial attention and decomposition. In *International Conference on Learning Representations*, 2019.

[156] Zachary C. Lipton, Yu-Xiang Wang, and Alex Smola. Detecting and correcting for label shift with black box predictors. *arXiv*

preprint 1802.03916, 2018.

[157] Francesco Locatello, Gabriele Abbati, Tom Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. In *Advances in Neural Information Processing Systems*, pages 14544–14557, 2019.

[158] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[159] Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[160] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, 2020.

[161] D. Lopez-Paz, K. Muandet, B. Schölkopf, and I. Tolstikhin. Towards a learning theory of cause-effect inference. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1452–1461, 2015.

[162] D. Lopez-Paz, R. Nishihara, S. Chintala, B. Schölkopf, and L. Bottou. Discovering causal signals in images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 58–66, 2017.

[163] K. Lorenz. *Die Rückseite des Spiegels*. R. Piper & Co. Verlag, 1973.

[164] Chaochao Lu, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Deconfounding reinforcement learning in observational settings. *arXiv preprint 1812.10576*, 2018.

[165] Chaochao Lu, Biwei Huang, Ke Wang, José Miguel Hernández-Lobato, Kun Zhang, and Bernhard Schölkopf. Sample-efficient reinforcement learning via counterfactual-based data augmentation. *arXiv preprint 2012.09092*, 2020.

[166] Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019.

[167] Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M. Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Proc. NeurIPS*, 2018.

[168] Robert Matthews. Storks deliver babies (p= 0.008). *Teaching Statistics*, 22(2):36–38, 2000.

[169] Nicolai Meinshausen. Causality from a distributional robustness point of view. In *2018 IEEE Data Science Workshop (DSW)*, pages 6–10. IEEE, 2018.

[170] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint 1907.07484*, 2019.

[171] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[172] B. T. Montet, T. D. Morton, D. Foreman-Mackey, J. A. Johnson, D. W. Hogg, B. P. Bowler, D. W. Latham, A. Bieryla, and A. W. Mann. Stellar and planetary properties of K2 campaign 1 candidates and validation of 17 planets, including a planet receiving earth-like insolation. *The Astrophysical Journal*, 809 (1):25, 2015.

[173] J. Mooij, D. Janzing, and B. Schölkopf. From ordinary differential equations to structural causal models: the deterministic case. In A. Nicholson and P. Smyth, editors, *Proceedings of the Twenty-Ninth Conference Annual Conference on Uncertainty in Artificial Intelligence*, pages 440–448, Corvallis, OR, 2013. AUAI Press. URL http://www.is.tuebingen.mpg.de/fileadmin/user_upload/files/publications/2013/MooijJS2013-uai.pdf.

[174] J. M. Mooij, D. Janzing, J. Peters, and B. Schölkopf. Regression by dependence minimization and its application to causal inference. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 745–752, 2009.

[175] J. M. Mooij, D. Janzing, T. Heskes, and B. Schölkopf. On causal discovery with cyclic additive noise models. In *Advances in Neural Information Processing Systems 24 (NIPS)*, 2011.

[176] J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016.

[177] Damian Mrowca, Chengxu Zhuang, Elias Wang, Nick Haber, Li Fei-Fei, Josh Tenenbaum, and Daniel L K Yamins. Flexible neural representation for physics prediction. In *Advances in Neural Information Processing Systems*, pages 8799–8810, 2018.

[178] Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard L Lewis, and Satinder Singh. Action-conditional video prediction using deep networks in atari games. In *Advances in neural information processing systems*, pages 2863–2871, 2015.

[179] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint 1807.03748*, 2018.

[180] G. Parascandolo, M. Rojas-Carulla, N. Kilbertus, and B. Schölkopf. Learning independent causal mechanisms. In *Workshop: Learning Disentangled Representations: from Perception to Control at the 31st Conference on Neural Information Processing Systems (NIPS)*, 2017.

[181] G. Parascandolo, N. Kilbertus, M. Rojas-Carulla, and B. Schölkopf. Learning independent causal mechanisms. In *Proceedings of the 35th International Conference on Machine Learning, PMLR 80:4036-4044*, 2018.

[182] Giambattista Parascandolo, Alexander Neitz, ANTONIO ORVIETO, Luigi Gresele, and Bernhard Schölkopf. Learning explanations that are hard to vary. In *International Conference on Learning Representations*, 2021.

[183] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, 2nd edition, 2009.

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

[185] Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. *arXiv preprint 1503.01603*, 2015.

[186] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Identifiability of causal graphs using functional models. In *Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 589–598, 2011.

[187] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014. URL http://jmlr.org/papers/v15/peters14a.html.

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

[189] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

[190] Jonas Peters, Stefan Bauer, and Niklas Pfister. Causal models for dynamical systems. *arXiv preprint 2001.06208*, 2020.

[191] N. Pfister, P. Bühlmann, B. Schölkopf, and J. Peters. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1): 5–31, 2018.

[192] Niklas Pfister, Stefan Bauer, and Jonas Peters. Learning stable

and predictive structures in kinetic systems. *Proceedings of the National Academy of Sciences*, 116(51):25405–25411, 2019.

[193] Niklas Pfister, Peter Bühlmann, and Jonas Peters. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527):1264–1276, 2019.

[194] Rémi Le Priol, Reza Babanezhad Harikandeh, Yoshua Bengio, and Simon Lacoste-Julien. An analysis of the adaptation speed of causal models. *arXiv preprint 2005.09136*, 2020.

[195] Stephan Rabanser, Stephan Günnemann, and Zachary C. Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. *arXiv preprint 1810.11953*, 2018.

[196] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018.

[197] Nasim Rahaman, Anirudh Goyal, Muhammad Waleed Gondal, Manuel Wuthrich, Stefan Bauer, Yash Sharma, Yoshua Bengio, and Bernhard Schölkopf. Spatially structured recurrent modules. In *International Conference on Learning Representations*, 2021.

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

[199] Laine K Reichert and John R Slate. Reflective learning: The use of "if only..." statements to improve performance. *Social Psychology of Education*, 3(4):261–275, 1999.

[200] Danilo J Rezende, Ivo Danihelka, George Papamakarios, Nan Rosemary Ke, Ray Jiang, Theophane Weber, Karol Gregor, Hamza Merzic, Fabio Viola, Jane Wang, et al. Causally correct partial models for reinforcement learning. *arXiv preprint 2002.02836*, 2020.

[201] Jonathan G Richens, Ciarán M Lee, and Saurabh Johri. Improving the accuracy of medical diagnosis with causal machine learning. *Nature Communications*, 11(1):3923, 2020.

[202] Karl Ridgeway and Michael C Mozer. Learning deep disentangled embeddings with the f-statistic loss. In *Advances in Neural Information Processing Systems*, pages 185–194, 2018.

[203] Neal J Roese. The functional basis of counterfactual thinking. *Journal of personality and Social Psychology*, 66(5):805, 1994.

[204] M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

[205] Michal Rolinek, Dominik Zietlow, and Georg Martius. Variational autoencoders pursue PCA directions (by accident). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12406–12415, 2019.

[206] Prasun Roy, Subhankar Ghosh, Saumik Bhattacharya, and Umapada Pal. Effects of degradations on deep neural network architectures. *arXiv preprint 1807.10108*, 2018.

[207] P. K. Rubenstein, S. Weichwald, S. Bongers, J. M. Mooij, D. Janzing, M. Grosse-Wentrup, and B. Schölkopf. Causal consistency of structural equation models. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*, pages 808–817, 2017.

[208] P. K. Rubenstein, S. Bongers, B. Schölkopf, and J. M. Mooij. From deterministic ODEs to dynamic structural causal models. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.

[209] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint 1706.05098*, 2017.

[210] Stuart Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Prentice Hall, 2002.

[211] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter W Battaglia. Learning to simulate complex physics with graph networks. *arXiv preprint 2002.09405*, 2020.

[212] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, pages 4967–4976, 2017.

[213] Jürgen Schmidhuber. *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*. PhD thesis, Technische Universität München, 1987.

[214] Jürgen Schmidhuber. Curious model-building control systems. In *Proc. international joint conference on neural networks*, pages 1458–1463, 1991.

[215] B. Schölkopf. Artificial intelligence: Learning to see and act. *Nature*, 518(7540):486–487, 2015.

[216] B. Schölkopf. Causal learning, 2017. Invited Talk, 34th International Conference on Machine Learning (ICML), https://vimeo.com/238274659.

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

[218] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. M. Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1255–1262, 2012.

[219] B. Schölkopf, D. Hogg, D. Wang, D. Foreman-Mackey, D. Janzing, C.-J. Simon-Gabriel, and J. Peters. Modeling confounding by half-sibling regression. *Proceedings of the National Academy of Science (PNAS)*, 113(27):7391–7398, 2016.

[220] B. Schölkopf, D. Janzing, and D. Lopez-Paz. Causal and statistical learning. In *Oberwolfach Reports*, volume 13(3), pages 1896–1899, 2016. doi: 10.14760/OWR-2016-33. URL https://publications.mfo.de/handle/mfo/3537.

[221] Bernhard Schölkopf. Causality for machine learning. *arXiv preprint 1911.10500*, 2019.

[222] Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. Towards the first adversarially robust neural network model on MNIST. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1EHOsC9tX.

[223] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *arXiv preprint 1911.08265*, 2019.

[224] Peter Schulam and Suchi Saria. Reliable decision support using counterfactual models. In *Advances in Neural Information Processing Systems*, pages 1697–1708, 2017.

[225] Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *arXiv preprint 1804.07203*, 2018.

[226] N. Shajarisales, D. Janzing, B. Schölkopf, and M. Besserve. Telling cause from effect in deterministic linear dynamical systems. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 285–294, 2015.

[227] Vaishaal Shankar, Achal Dave, Rebecca Roelofs, Deva Ramanan, Benjamin Recht, and Ludwig Schmidt. Do image classifiers generalize across time? *arXiv preprint 1906.02168*, 2019.

[228] Rakshith Shetty, Bernt Schiele, and Mario Fritz. Not using the car to see the sidewalk–quantifying and controlling the effects of context in classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8218–8226, 2019.

[229] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. J. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

[230] Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. *arXiv preprint 1910.09772*, 2019.

[231] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[232] David Silver, Hado Hasselt, Matteo Hessel, Tom Schaul, Arthur Guez, Tim Harley, Gabriel Dulac-Arnold, David Reichert, Neil Rabinowitz, Andre Barreto, et al. The predictron: End-to-end

learning and planning. In *International Conference on Machine Learning*, pages 3191–3199. PMLR, 2017.

[233] Patrice Simard, Bernard Victorri, Yann LeCun, and John Denker. Tangent prop - a formalism for specifying selected invariances in an adaptive network. In J. Moody, S. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4, pages 895–903. Morgan-Kaufmann, 1992. URL https://proceedings.neurips.cc/paper/1991/file/65658fde58ab3c2b6e5132a39fae7cb9-Paper.pdf.

[234] Patrice Y Simard, David Steinkraus, John C Platt, et al. Best practices for convolutional neural networks applied to visual document analysis. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)*, volume 3, 2003.

[235] H. A. Simon. Causal ordering and identifiability. In W. C. Hood and T. C. Koopmans, editors, *Studies in Econometric Methods*, pages 49–74. John Wiley & Sons, New York, NY, 1953. Cowles Commission for Research in Economics, Monograph No. 14.

[236] Elizabeth S Spelke. Principles of object perception. *Cognitive science*, 14(1):29–56, 1990.

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

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

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

[240] B. Steudel, D. Janzing, and B. Schölkopf. Causal Markov condition for submodular information measures. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, pages 464–476, 2010.

[241] Jianyu Su, Stephen Adams, and Peter A Beling. Counterfactual multi-agent reinforcement learning with graph convolution communication. *arXiv preprint 2004.00470*, 2020.

[242] Adarsh Subbaswamy and Suchi Saria. Counterfactual normalization: Proactively addressing dataset shift and improving reliability using causal mechanisms. *arXiv preprint 1808.03253*, 2018.

[243] Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. *arXiv preprint 1812.04597*, 2018.

[244] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.

[245] Chen Sun, Per Karlsson, Jiajun Wu, Joshua B Tenenbaum, and Kevin Murphy. Stochastic prediction of multi-agent interactions from partial observations. *arXiv preprint 1902.09641*, 2019.

[246] X. Sun, D. Janzing, and B. Schölkopf. Causal inference by choosing graphs with most plausible Markov kernels. In *Proceedings of the 9th International Symposium on Artificial Intelligence and Mathematics*, 2006.

[247] Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *International Conference on Machine Learning*, pages 6056–6065. PMLR, 2019.

[248] Richard S Sutton, Andrew G Barto, et al. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.

[249] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint 1312.6199*, 2013.

[250] Ernő Téglás, Edward Vul, Vittorio Girotto, Michel Gonzalez, Joshua B Tenenbaum, and Luca L Bonatti. Pure reasoning in 12-month-old infants as probabilistic inference. *Science*, 332 (6033):1054–1059, 2011.

[251] J. Tian and J. Pearl. Causal discovery from changes. In *Proceedings of the 17th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 512–522, 2001.

[252] Frederik Träuble, Elliot Creager, Niki Kilbertus, Anirudh Goyal, Francesco Locatello, Bernhard Schölkopf, and Stefan Bauer. Is independence all you need? on the generalization of representations learned from correlated data. *arXiv preprint 2006.07886*, 2020.

[253] Michael Tschannen, Josip Djolonga, Marvin Ritter, Aravindh Mahendran, Neil Houlsby, Sylvain Gelly, and Mario Lucic. Self-supervised learning of video-induced visual invariances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13806–13815, 2020.

[254] Angelos Tsiaras, Ingo Waldmann, G. Tinetti, Jonathan Tennyson, and Sergei Yurchenko. Water vapour in the atmosphere of the habitable-zone eight-earth-mass planet K2-18b. *Nature Astronomy*, 2019. doi: 10.1038/s41550-019-0878-9.

[255] Sjoerd Van Steenkiste, Michael Chang, Klaus Greff, and Jürgen Schmidhuber. Relational neural expectation maximization: Unsupervised discovery of objects and their interactions. In *6th International Conference on Learning Representations (ICLR)*, 2018.

[256] Sjoerd van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are disentangled representations helpful for abstract visual reasoning? In *Advances in Neural Information Processing Systems*, pages 14178–14191, 2019.

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

[258] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

[259] J. von Kügelgen, A. Mey, M. Loog, and B. Schölkopf. Semi-supervised learning, causality and the conditional cluster assumption. *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2020.

[260] Julius von Kügelgen, Umang Bhatt, Amir-Hossein Karimi, Isabel Valera, Adrian Weller, and Bernhard Schölkopf. On the fairness of causal algorithmic recourse. *arXiv 2010.06529*, 2020.

[261] Julius von Kügelgen, Luigi Gresele, and Bernhard Schölkopf. Simpson's paradox in Covid-19 case fatality rates: a mediation analysis of age-related causal effects. *arXiv 2005.07180*, 2020.

[262] Julius von Kügelgen, Ivan Ustyuzhaninov, Peter Gehler, Matthias Bethge, and Bernhard Schölkopf. Towards causal generative scene models via competition of experts. *arXiv 2004.12906*, 2020.

[263] Haohan Wang, Zexue He, Zachary C. Lipton, and Eric P. Xing. Learning robust representations by projecting superficial statistics out. *arXiv preprint 1903.06256*, 2019.

[264] Nicholas Watters, Daniel Zoran, Theophane Weber, Peter Battaglia, Razvan Pascanu, and Andrea Tacchetti. Visual interaction networks: Learning a physics simulator from video. In *Advances in neural information processing systems*, pages 4539–4547, 2017.

[265] Nicholas Watters, Loic Matthey, Matko Bosnjak, Christopher P Burgess, and Alexander Lerchner. Cobra: Data-efficient model-based rl through unsupervised object discovery and curiosity-driven exploration. *arXiv preprint 1905.09275*, 2019.

[266] S. Weichwald, B. Schölkopf, T. Ball, and M. Grosse-Wentrup. Causal and anti-causal learning in pattern recognition for neuroimaging. In *4th International Workshop on Pattern Recognition in Neuroimaging (PRNI)*. IEEE, 2014.

[267] Sebastian Weichwald. *Pragmatism and Variable Transformations in Causal Modelling*. PhD thesis, ETH Zurich, 2019.

[268] Marco Wiering and Martijn Van Otterlo. *Reinforcement learning*, volume 12. Springer, 2012.

[269] Steffen Wiewel, Moritz Becher, and Nils Thuerey. Latent space physics: Towards learning the temporal evolution of fluid flow. In *Computer Graphics Forum*, volume 38, pages 71–82. Wiley

Online Library, 2019.

[270] Laurenz Wiskott and Terrence J Sejnowski. Slow feature analysis: unsupervised learning of invariances. *Neural computation*, 14(4):715–70, April 2002. ISSN 0899-7667.

[271] Chris Xie, Sachin Patil, Teodor Moldovan, Sergey Levine, and Pieter Abbeel. Model-based reinforcement learning with parametrized physical models and optimism-driven exploration. In *2016 IEEE international conference on robotics and automation (ICRA)*, pages 504–511. IEEE, 2016.

[272] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. CLEVRER: Collision events for video representation and reasoning. *arXiv preprint 1910.01442*, 2019.

[273] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.

[274] Vinicius Zambaldi, David Raposo, Adam Santoro, Victor Bapst, Yujia Li, Igor Babuschkin, Karl Tuyls, David Reichert, Timothy Lillicrap, Edward Lockhart, et al. Deep reinforcement learning with relational inductive biases. In *International Conference on Learning Representations*, 2018.

[275] J. Zhang and E. Bareinboim. Near-optimal reinforcement learning in dynamic treatment regimes. In *Advances in Neural Information Processing Systems 33*, pages 13401–13411, 2019.

[276] Junzhe Zhang and Elias Bareinboim. Fairness in decision-making - the causal explanation formula. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA*, pages 2037–2045, 2018.

[277] K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 647–655, 2009.

[278] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 804–813, 2011.

[279] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 819–827, 2013.

[280] K. Zhang, M. Gong, and B. Schölkopf. Multi-source domain adaptation: A causal view. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 3150–3157, 2015.

[281] K. Zhang, B. Huang, J. Zhang, C. Glymour, and B. Schölkopf. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017)*, pages 1347–1353, 2017.

[282] Richard Zhang. Making convolutional networks shift-invariant again. *arXiv preprint 1904.11486*, 2019.