An Adversarial Perspective on Accuracy, Robustness, Fairness, and Privacy: Multilateral-Tradeoffs in Trustworthy ML

ALEX GITTENS\textsuperscript{1}, (Member, IEEE), BÜLENT YENER\textsuperscript{1}, (Fellow, IEEE), AND MOTI YUNG\textsuperscript{2}, (Fellow, IEEE)
\textsuperscript{1}Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180, USA
\textsuperscript{2}Google LLC, New York, NY 10011, USA

ABSTRACT Model accuracy is the traditional metric employed in machine learning (ML) applications. However, privacy, fairness, and robustness guarantees are crucial as ML algorithms increasingly pervade our lives and play central roles in socially important systems. These four desiderata constitute the pillars of Trustworthy ML (TML) and may mutually inhibit or reinforce each other. It is necessary to understand and clearly delineate the trade-offs among these desiderata in the presence of adversarial attacks. However, threat models for the desiderata are different and the defenses introduced for each leads to further trade-offs in a multilateral adversarial setting (i.e., a setting attacking several pillars simultaneously). The first half of the paper reviews the state of the art in TML research, articulates known multilateral trade-offs, and identifies open problems and challenges in the presence of an adversary that may take advantage of such multilateral trade-offs. The fundamental shortcomings of statistical association-based TML are discussed, to motivate the use of causal methods to achieve TML. The second half of the paper, in turn, advocates the use of causal modeling in TML. Evidence is collected from across the literature that causal ML is well-suited to provide a unified approach to TML. Causal discovery and causal representation learning are introduced as essential stages of causal modeling, and a new threat model for causal ML is introduced to quantify the vulnerabilities introduced through the use of causal methods. The paper concludes with pointers to possible next steps in the development of a causal TML pipeline.

INDEX TERMS Security, adversarial robustness, privacy, fairness, machine learning, causal models, causal representation, trustworthy machine learning.

I. PILLARS OF TRUSTWORTHY MACHINE LEARNING

Traditional applications of machine learning (ML) have included tasks with low social impact, such as email spam filtering, fraud detection, product recommendation, and image recognition. In these applications, accuracy, viz. fidelity to the training data and low error on similar test data, has been the primary performance metric that ML systems were designed to optimize. However, as the availability of massive data sets and powerful models has increased, ML is increasingly deployed in socially sensitive domains including health-care, law enforcement, and credit lending.

As ML becomes ubiquitous and plays a more influential role in the safety, well-being, and health of our society, it has become evident that other metrics are equally as relevant in assessing the performance of ML systems. Specifically, in order for an ML system to be trustworthy, in addition to accuracy, it must exhibit robustness, privacy, and fairness. It is evident that these goals may be mutually supportive, or mutually detrimental. This paper surveys the known trade-offs between these four desiderata, discusses the vulnerabilities exposed by optimizing for multiple desiderata, and highlights what trade-offs remain unknown.

The associate editor coordinating the review of this manuscript and approving it for publication was Dost Muhammad Khan.
or underexplored. Further, we note that unlike accuracy, the goals of privacy, fairness, and robustness cannot be achieved by increasing the expressive power of models or the amount of training data; instead, they require appropriate forms of regularization. Different regularization techniques are generally used to achieve these three separate desiderata. The latter portion of the paper argues that causal learning provides a form of regularization suited to each of these desiderata.

Another important desiderata of trustworthy ML is interpretability: to what extent can the predictions of the model be explained to human stakeholders? Similarly, a recent work [1] links the quality of explanations to fairness. In particular they show that fidelity of explainability methods may vary in performance between subgroups, indicating fairness problems. They also argue that methods for improving robustness also improve explainability and fairness.

Explainability and interpretability are not considered in this paper, as the quantification of interpretability is application-specific, and requirements vary from stakeholder to stakeholder even in the same arena. Although interpretability can to some extent be addressed using causal modeling, the existence of a unified and practically useful theoretical framework for interpretability seems highly improbable.

In this introduction we briefly describe the basic desiderata of trustworthy ML (TML) other than accuracy, and how they are achieved.

Notation: Throughout, bold uppercase roman letters denote random variables while lowercase roman letters denote their values, e.g. \( X = x \). We consider supervised machine learning problems with input, sensitive, and target variables \((X, S, Y)\) distributed according to the population (joint distribution) \( \mathbb{P} \). Expectations over the population are denoted by \( \mathbb{E} \); e.g., the performance of a hypothesis function \( f \) is typically measured using the expected loss \( \mathbb{E}(f(X), Y) \), where \( f \) is a loss function.

Organization of This Paper: This paper is organized in five sections. Section I introduces the four pillars of trustworthy machine learning: robustness, fairness, privacy, and accuracy. It shows that directly jointly optimizing over all these desiderata is not a feasible approach. Section II focuses on multiway trade-offs among these desiderata. While some pairwise trade-offs are discussed in the literature, three or more way trade-offs remain under-investigated. Section III defines the adversarial models relevant to these desiderata and trade-offs. It discusses the basic attack and potential defense mechanisms for each one of the desiderata. We conclude that there are fundamental vulnerabilities inherent to association-based learning, which does not capture the cause-effect relationships between the predictor and predicted variables. Thus, Section IV advocates a causality-based paradigm of trustworthy machine learning. In this section, the basic background for causal learning is given and each of the desiderata is discussed from a causal perspective. This section concludes with the introduction of an adversarial model for attacks on causal learning. Finally, in Section V, we discuss open problems and future research directions in employing causality to achieve trustworthy machine learning.

A. ROBUSTNESS

Robustness of statistical learning procedures has always been of practical interest. In classical robust statistics, it is motivated by (i) uncertainty around the quality of the data, so we desire model-fitting procedures that are robust to errors in the data; and (ii) the fact that model misspecification errors that arise from fitting simple models to complex phenomenon are exacerbated by non-robust model-fitting procedures.

In modern ML, robust learning has taken on an additional implication beyond classical robust statistics [2], adversarial robustness, to reflect the vulnerability of machine learning models to imperceptible perturbations of their input [3]. Given a data point \((x, y)\), an adversarial example \(x'\) is a data point that is similar to \(x\), but causes the model to predict a label other than \(y\); adversarial attacks are either targeted, when the goal is to obtain a specific label, or untargeted, when the goal is to obtain any false label. An adversarially robust model is invulnerable to such adversarial examples. Adversarial examples can be generated under different attack models, including rotations, translations, and deformations, but the canonical \(\varepsilon\)-bounded attack model is that \(x'\) is \(\varepsilon\)-close to \(x\) in an \(\ell_p\) norm, for \(p \in \{1, 2, \infty\}\).

Adversarial robustness in the \(\varepsilon\)-bounded attack model is quantified by the adversarial accuracy

\[
\mathbb{E} \max_{\|\delta\|_p \leq \varepsilon} \ell(f(X + \delta), Y).
\] (1)

Section III discusses attack and defense techniques aimed at adversarial robustness in more detail.

B. FAIRNESS

As learned models are increasingly deployed in socially sensitive contexts such as hiring, loan decisions, and healthcare, fairness has emerged as a central concern in machine learning. Fairness requires that the model’s output must not discriminate upon the basis of some protected attribute. This intuition is quantified using different mathematical formulations, as appropriate to the specific application.

We distinguish between individual and group fairness. Individual fairness [4], [5] ensures that similar \(x\) are assigned similar outcomes, and is strongly related to robustness as discussed later in section II-D. However individual fairness is more computationally expensive to ensure, and requires an application-specific metric. In contrast, group fairness quantifies and explains fairness in a manner that lends itself to multiple domains of application. Thus, we mainly consider the widely used class of group fairness metrics, which impose different variants of independence between the model’s output and the sensitive variable [6].

The following definitions were introduced for classification problems, so we take \(f\) to be binary-valued, so that e.g. \(\mathbb{E}[f(X)] = \mathbb{P}[f(X) = 1]\), but several group fairness definitions have also been generalized to the setting of regression
problems. The literature has converged on several widely applicable formulations of group fairness; of these, the most often seen are [6]:

(i) **Demographic Parity** ensures that the outcome of the model is independent of the sensitive attributes: \( f(X) \perp S \). Intuitively, it aims at similar outcomes for different groups. Demographic parity is achieved once the independence of \( f(X) \) and \( S \) “holds in expectation”, i.e. if \( \mathbb{E}[f(X) | S = a] = \mathbb{E}[f(X) | S = b] \) for all potential values \( a \) and \( b \) of \( S \) [1]. This is a very strong notion of fairness, as it says that the sensitive attribute must not influence the model’s predictions, even if \( X \) and \( S \) are dependent. More refined notions of group fairness allow \( S \) to affect \( f(X) \) in controlled manners.

(ii) **Equal Opportunity** ensures that the model prediction is independent of the sensitive variable when the true label is positive: \( f(X) \perp S | Y = 1 \). This type of fairness is appropriate when the model \( f \) is used to predict who to hire; if we know the person is capable of doing the job \( (Y = 1) \), then the decision to hire should be independent of gender and race \((S)\).

(iii) **Equalized Odds** ensures that both the probability of true and false positives are independent of the sensitive variable: \( f(X) \perp S | Y \). This type of fairness is appropriate when \( f \) is used to classify a person’s credit-worthiness; applicants with good credit \( (Y = 1) \) or applicants with bad credit \( (Y = 0) \) should be classified similarly, regardless of their gender or race \((S)\).

Multiple metrics have been proposed to quantify each form of group fairness; it is conventional for researchers to introduce novel metrics to capture the notion of group fairness of interest, then develop algorithms targeted at that specific metric [6]. For instance, we say that a model has demographic parity gap \( \phi \) with respect to \( S \) if \( \max_{a,b} | \mathbb{E}[f(X) | S = a] - \mathbb{E}[f(X) | S = b] | \leq \phi \). Demographically fair models can then be learned by minimizing the demographic parity gap.

Algorithms achieving group fairness can be categorized as either preprocessing, in-processing, or post-processing methods [7]. Preprocessing methods attempt data repair, modifying the training data so that model learning is independent of the sensitive variables. In-processing methods, by far the most common approach to achieving group fairness, add constraints to the model or modify the training objective to ensure fairness; these approaches can be model-specific. Post-processing methods can also be model-specific, and alter already fitted models in order to ensure fairness.

**C. PRIVACY**

Machine learning fundamentally depends on understanding the distribution of the relevant population: recommendation systems, for instance, aggregate ratings across users to learn to predict user preferences. However, because the training data is finite, fitting ML models potentially discloses sensitive information on individual users, even after seemingly clever anonymization techniques are used on the data set.

Differential privacy (DP) provides strong privacy guarantees for each individual’s data and has become the standard framework for privacy-preserving ML [8]. A randomized algorithm \( A \) is said to be \((\varepsilon, \delta)\)-DP if it guarantees that for any two training data sets \( D \) and \( D' \) that differ by inclusion or exclusion of a single example, and any set \( S \) in the output space,

\[
\mathbb{P} \left[ A(D) \in S \right] \leq \exp(\varepsilon) \mathbb{P} \left[ A(D') \in S \right] + \delta.
\]

This guarantee means that even if all the other examples used in fitting a model are known, one cannot infer the presence or absence of any particular example in the training set \( D \) from the learned model \( A(D) \), because it is very likely to be similar to a model \( A(D') \) learned using a training set that does not contain that example.

DP machine learning is achieved in practice using noise insertion via the Laplacian, Gaussian, or exponential mechanisms, or by randomly perturbing the model output or the learning objective [9]. By far the most popular and ubiquitous approach to privacy-preserving model learning is differentially private SGD (DP-SGD), which modifies the standard SGD algorithm by clipping the model gradient and adding noise to the update steps to lower disclosure risk [10]. DP-SGD and its variants are locally differentially private, because the differential privacy is ensured by adding noise locally: in the case of DP-SGD, the data owners add noise locally to the gradients on each of their training points, thereby limiting the risk of disclosure when they return their gradients to be used in updating the model. By way of contrast, methods that perturb the learning objective are not locally DP because the data owners completely expose their data during the training process. Local DP is more stringent than general DP, and should be employed when the owners of the data do not trust the consumers of the data; thus local differential privacy is particularly desirable in the settings of federated, distributed, and edge learning, where the individual owners of data locally modify their data to limit disclosure risk before passing it along potentially compromised communication channels [11].

The differential privacy of iterative learning algorithms (e.g. DP-SGD) is guaranteed by the application of DP composition theorems which track the loss of privacy per iteration [8], [10]. The theoretical bounds obtained in this fashion for non-trivial optimization algorithms such as DP-SGD applied to deep learning tend to be loose, so the ML community has developed numerical estimation procedures [12] and heuristic proxies. More recently, vulnerability to membership inference attacks – determining whether a given data record was used in the training of a model — has become a popular and easily estimable proxy metric for disclosure risk [13].

**D. ACHIEVING MULTIPLE DESIDERATA JOINTLY IS HARD**

Fairness, privacy, and robustness must be achieved using regularization, as these properties are independent of the training data and thus must be explicitly and externally imposed upon
II. MULTILATERAL TRADE-OFFS BETWEEN THE DESIDERATA OF TML

Understanding the interactions between the desiderata is of fundamental importance to ensuring ML that is trustworthy along multiple axes. This gives rise to two important questions: (i) What are the fundamental, algorithm-independent, trade-offs between robustness, privacy, fairness, and accuracy? (ii) How do the optimization algorithms and regularizations used in practice to guarantee these desiderata positively or negatively interact with each other? Figure 1 summarizes some of the trade-offs discussed in more detail in this section.

A. THREE OR MORE WAY TRADE-OFFS

The explicit study of three- or more way trade-offs is rare, due to the difficulty of characterizing and optimizing for the four desiderata. Because theoretical quantification of these multi-way trade-offs has largely proven elusive, most results in this space are empirical characterizations of the trade-offs incurred by using specific algorithms to achieve the desiderata.

Jagielski et al. [14] modify algorithms that explore the trade-off between fairness and accuracy by inserting noise to make them DP, and show corresponding three-way trade-offs between these three desiderata. The recent work [15] provides a heuristic that balances between accuracy, privacy, and fairness by using DP-SGD and employing an early stopping criteria which monitors the change in fairness and accuracy on a validation set.

In comparison to the case of three- and more way trade-offs, there is a large body of accumulated knowledge on pairwise trade-offs.

B. ACCURACY VS FAIRNESS, PRIVACY, ROBUSTNESS

1) PRIVACY

Differential privacy requires that each data point has little impact on the final learned model; this is achieved through the insertion of well-calibrated noise to mask the influence of influential data points. The insertion of noise decreases the accuracy of the model; optimal trade-offs between differential privacy and accuracy are known for many machine learning problems [16], [17].

2) FAIRNESS

The trade-off between fairness and accuracy is typically quantified using the trade-off between fairness and accuracy on held-out test data, as there may not even be a fair generative process relating the target to the predictor variables. By presumption the test data comes from observations that implicitly convey some unfair bias, so attempts to increase the fairness of the learned model conflict with the desire to accurately fit the test data. These trade-offs have been quantified for a variety of models for differing forms of group fairness. The main theme is that, as the input data is distorted away from its empirical distribution to satisfy...
fairness, the misfit between the predictions and the true targets increases [18], [19].

3) ROBUSTNESS
Early work shows provably that there exist unavoidable trade-offs between adversarial robustness and accuracy: [20], [21] exploited the behavior of high-dimensional functions and functions defined on manifolds, respectively, to argue that adversarial robustness is generically unachievable. But as more structure is introduced to the target function and hypothesis space, we see that robustness and accuracy are more reconcilable; e.g. [22] shows that if finite margins exist between classes, then robust learning is achievable using locally Lipschitz continuous functions. In general, adversarial robustness is achieved through various forms of regularization; with other forms of regularization, adversarial regularization has been shown to help in the case of small amounts of training data, but eventually a cross-over occurs where adversarial regularization hurts accuracy [23].

C. PRIVACY AND FAIRNESS
Randomness is, by definition, essential to DP. This randomness can be inserted into the learning process at the input, in the objective, to the gradients, to the outputs, and to the labels [24]. Due to the non-convexity of the learning process, the standard approach to privacy in training DL models is to add noise to the gradient; specifically, the DP-SGD algorithm clips the gradient before adding noise, and trades off accuracy in the final model for increased differential privacy.

It is known that the decrease is accuracy induced through the use of DP-SGD is unevenly distributed across classes: when DP-SGD is used, the “poor become poorer”, in the sense that those classes for which non-DP models perform worse suffer a greater decrease in accuracy [25]. Thus, the use of DP training tends to explicitly decrease fairness (as measured by accuracy parity). Similarly, interventions that increase fairness also increase disclosure risk on the minority classes [26]. Prior work has considered the interplay of DP and fairness [14], introducing a particular min-max algorithmic approach to obtaining provably private and fair classifiers, where noise is inserted at each step to ensure privacy.

One approach to understanding the trade-off between fairness and privacy from DP-SGD is to note that imposing fairness requirements raises the influence of underrepresented groups on the model being fit. This in turn leads to information leakage in the model about the more influential data points; indeed, the privacy leakage is known to be more heavily incurred on the minority classes [26]. Similarly, adding noise and clipping gradients to increase DP is intended precisely to reduce the influence of any individual data point on the fitted model.

On the theoretical side, there is evidence that fairness and privacy are competing goals. Agarwal’s work provides a strong impossibility theorem for binary classification: it is not possible to have a learning algorithm that both guarantees ($\epsilon$, 0)-differential privacy and fairness and returns a non-constant classifier [27]. This work suggests that if accuracy is required, both ($\epsilon$, 0)-DP and fairness cannot be obtained by a general algorithm.

D. ROBUSTNESS AND FAIRNESS
Individual fairness and robustness are strongly coupled since individual fairness implies algorithmic robustness and one can satisfy individual fairness through Lipschitz regularization [28], [29]. However, the applicability of individual fairness is limited by the need for an application-specific metric to define similar individuals [29]. Group fairness, by way of contrast, does not require application-specific metrics, simply the choice of an appropriate form of group fairness. However, group fairness can significantly adversely impact individual fairness.

Current algorithms used to achieve group fairness are brittle: different test-train splits or preprocessing choices can lead to significant changes in the trade-off between fairness and accuracy [7]; it is reasonable to infer that group fairness algorithms also have a poor trade-off with robustness. We refer to two recent surveys and the references therein for extensive listings of algorithms for fair ML and their performance [30], [31].

On one hand, robustness essentially downweights the importance of fitting any particular subset of the training data set accurately, for an increase in the robustness to adversarial perturbations. Indeed, the concept of robustness bias is introduced in [32], which establishes that it may be easier for adversaries to attack certain classes in adversarially robust models, even when the training data is balanced. [33] and [34] confirmed this finding and introduced algorithms that decrease the robustness bias by introducing class-weighted losses in the popular PGD-AT [35] and TRADES [36] algorithms for adversarial robust machine learning.

On the other hand, fair ML methods increase the importance of minority classes in order to equalize performance across classes. This increase in the dependence of particular subgroups in the training data set conflicts with robustness [37].

E. ROBUSTNESS AND PRIVACY
Several forms of regularization have been used to guarantee robust and private ML models. The theoretical trade-offs between these two desiderata are unknown, but empirical evidence suggests that they are mutually detrimental. [38] considers the effect of using adversarial defense methods during training on the privacy of the learned models, and found that six state-of-the-art defenses increase the vulnerability to membership inference attacks. This work postulates that this occurs because of the well-established fact that overfitting causes privacy loss [39]: adversarial defense methods encourage models to stay constant around the training data, which increases the influence of the training data points on the fitted model. The claim that adversarial training leads to overfitting is supported by [40], which found that, unlike with standard training where it is difficult to overfit deep learned

120854
models, there is a point in adversarial training at which the robust loss on the training data continues to decrease while the robust loss on the validation data increases.

The work [41] considers the impact of training private models, using DP-SGD, on the robustness of the fitted models. It uses five measures of model robustness—accuracy under two classes of adversarial attacks, distance to the decision surface, curvature of the decision surface, and susceptibility to common corruptions—and found that all measures of robustness decrease when DP-SGD is used to train networks instead of SGD. Similar empirical results are presented in [42], along with the observation that adversarial attacks transfer better between DP models than between non-DP models.

III. ADVERSARIAL SETTING AND THREAT MODELS

The multilateral trade-offs among the desiderata introduces a new threat model in which the adversary may aim to exploit the trade-offs introduced by specific algorithms and regularization techniques used to ensure the trustworthiness desiderata. For example, it is known that differential privacy makes the “poor become poorer”, in the sense that those classes for which non-DP models perform worse suffer a greater decrease in accuracy when differential privacy is imposed [10]. Thus, the use of DP training tends to explicitly decrease fairness (as measured by accuracy parity).

In the new multilateral adversarial model, the attacker may target such dependencies in the desiderata. It may focus on adversarial input that achieves multiple objectives of degrading the desiderata. This is a particularly important problem for the fairness and privacy desiderata, which are not independent. While for some definitions of fairness and privacy, ensuring one can help the other, in other settings there may be a detrimental trade-off. For example, [4] shows how differential privacy tools can be used to ensure individual fairness. In contrast, [26] argues that ensuring group fairness increases the success of membership inference attacks on unprivileged subgroups: the distribution of data for each subgroup may be different and compensating for this difference to ensure group fairness enables attackers to infer if a particular group member was used in the training.

In this section we discuss threat models for the desiderata as well as potential defense mechanisms. We start by first examining the fault lines in association-based machine learning that enable attacks.

A. APPROXIMATION VS GENERALIZATION

ML algorithms that aim to learn a function \( f \approx f(X) \) using statistical associations (e.g., correlations) between the domain of training set data \( X \in \mathcal{R}^d \) and a target variable \( Y \) rely on two major assumptions. First, the observations \( \{(x_i, y_i)\}_{i=1}^{n} \) used in training are assumed to be i.i.d.; as the training data set size increases, the training data set better approximates the distribution of \( (X, Y) \). Second, it is assumed that no covariate shift occurs between the training and testing distributions. Covariate shift limits the generalization of the function \( f \) and can be mitigated by including some samples in the training set that come from the test data distribution. Transfer learning techniques have demonstrated limited success in generalizing a model to new distributions.

From an adversarial perspective, the approximation–generalization trade-off creates several vulnerabilities. In order to fit the training data well, the model may focus on attributes that may cause privacy and fairness concerns. Such attributes may be individual input features, or combinations of the input features. An adversary that discovers the identities of these attributes and how the model employs them can exploit privacy and fairness vulnerabilities of the model. Similarly, a model that does not generalize well to new distributions is vulnerable to adversarial examples [43].

The literature includes a rich set of work presenting remedies to these two fundamental issues within the context of association-based learning in the presence of a smart adversary [44].

B. CONFLICTING REQUIREMENTS

The applicability of the mathematical formalizations of the desiderata should always be questioned.

Robustness is situational: the standard \( \ell_p \)-norm perturbation model was introduced specifically in the context of computer vision, where it corresponds intuitively to human-imperceptible perturbations. But in other domains, other definitions of perturbations are more appropriate: in malware detection, perturbations may correspond to the insertion of no-op codes. How then do the approaches developed for fitting robust computer vision models translate over to fitting robust models in other domains?

Differential privacy is a very strong property, and may not be required in all applications; for this reason, different attempts have been made to weaken the definition of DP in order to provide intuitively satisfying quantifications of privacy which are more practically achievable.

It is also clear that the various definitions of group fairness all quantify different intuitive notions of fairness; there are multiple such definitions because each individual definition leaves significant gaps. Unfortunately, a series of works have established that, in general, multiple forms of group fairness cannot simultaneously hold if the classifier is nontrivial [45], [46], [47].

C. CURRENT THREAT MODELS

Adversaries are generally classified into one of three threat models, depending on their capabilities:

1) **White box:** In this threat model, the adversary \( A_w \) has access to the entire model, including the model’s architecture, weights, and gradients.

2) **Black box:** In this threat model, the adversary \( A_b \) has access to the model only at the input and output level. Thus \( A_b \) can only query the model. Typically \( A_b \) is allowed to query the model up to a maximum limit to conduct attacks, e.g. [35]. Substitution attacks, where \( A_b \)
queries the model and uses the outputs to learn a model equivalent to the target one, are a popular type of black box attack.

3) **Gray box:** In this threat model, the adversary $A_g$ has more knowledge than a black box adversary, but less than a white box adversary: e.g., $A_g$ does not know the weights of the model, but may know the model architecture and may be aware of the details of the defenses protecting the model.

### D. ATTACKS ON ROBUSTNESS

The goal of attacks on a model’s robustness is to force the model to commit an error—or arbitrary misclassification or targeted misclassification—by crafting adversarial examples that resemble inputs which the model labels accurately. For example, $A_g$ may optimize to find a perturbation that maximizes the cross-entropy loss between the true label and the predicted label, or to optimize more sophisticated objectives such as those popularized by Carlini and Wagner [48].

1) **ADVERSARIAL EXAMPLES AND DEFENSE MECHANISMS**

The fast sign gradient method (FSGM) is a simple algorithm for constructing adversarial examples under the white box threat model, that calculates the gradient of the loss function $\ell(x, y)$, with respect to the input $x$ to determine the perturbation (adding a perturbation on the scale of $\epsilon$ to $x$) to cause a classification error: $x' = x + \epsilon \text{sign}(\nabla_x \ell(x, y))$. The projected gradient descent attack (PGD) is a more sophisticated alternative to FSGM that takes multiple gradient steps to minimize the $\ell_p$ norm, for $p \in 1, 2, \infty$, between $x$ and the adversarial example $x'$.

A plethora of ingenious adversarial defenses have been introduced [49], [50], [51], [52], [53], [54], [55], [56], [57], but the most popular general purpose state-of-the-art adversarial defenses are variants of the projected gradient descent adversarial training (PGD-AT) algorithm [35], and the TRADES algorithm [36].

The PGD-AT algorithm replaces the usual risk minimization objective used in learning with the adversarial risk objective in (1), and uses a min-max optimization formulation to minimize the adversarial risk [35]. The TRADES algorithm retains the usual risk objective, and regularizes by a term called the boundary loss, that encourages robustness [36].

2) **POISONING ATTACKS**

ML models are also susceptible to data poisoning attacks. Poisoning attacks target the training phase of a ML model by inserting adversarial samples to the training data to manipulate the training set distribution $P$ to shift toward a distribution $Q$ which will cause the model to make mistakes by deceiving it into adjusting the decision function. Poisoning can occur under white, black, or gray box attack models.

Robustness to data poisoning attacks is closely related to classical statistically robust model-fitting, and is an important goal in itself; in fact, a recent survey of 28 companies found that out of eleven potential attacks including poisoning, model stealing, model inversion, and membership inference among others, the top concern was model poisoning [58].

Data poisoning attacks are often categorized as either backdoor or triggerless attacks. In the former, the adversary manipulates the training data and the testing data, typically so that the presence of certain features elicits a desired response from the model: e.g., an image classification model is trained to classify any input image as a stop sign, regardless of the actual content, as long as the $10 \times 10$ region of pixels in the upper-left hand corner is uniformly white. In triggerless attacks, the adversary manipulates only the training data.

Triggerless attack methods poison the training data set so that specific images are misclassified. As an example, given an input $x_i$ with true label not equal to $t$ and a target input $x_t$ with true label $t$, the popular feature collision method [59] generates a poisoned training sample as a minimizer of the objective

$$J(x) = \|f(x) - f(x_i)\|^2 + \beta\|x - x_t\|^2,$$

where $f$ is the feature extraction portion of the ML model. Here $\beta$ balances the goals of keeping $x$ similar to $x_i$ and keeping its features similar to those of an example from the target class. Since $x$ is close to $x_i$ but its feature representation is close to that of an example from class $t$, a model trained using $x$ is likely to misclassify $x_t$.

Backdoor attacks surreptitiously prime the model during training to classify an input with an included trigger as coming from a given target class. At test time, regardless of the content of the example, the presence of the trigger causes it to be classified as an instance of the target class. As an example, the hidden trigger backdoor attack of [60] adds triggers to an example using

$$\tilde{x} = x \odot (1 - m) + p \odot m,$$

where $x$ is the original example (image), $p$ is the patch, $m$ is a binary mask, and $\odot$ denotes pointwise multiplication of vectors. Training examples $z$ are created to have similar features as patched images, by minimizing $\|f(z) - f(\tilde{x})\|_2$, subject to the poisoned examples being perceptually close to images from the target class, $\|z - x_i\|_\infty < \epsilon$. In training, the poisoned examples are labeled with the target class. At test time, patched images will tend to have features similar to patched images at training time, and thus be classified as coming from the target class.

### E. ATTACKS ON PRIVACY

DP utilizes randomized noise addition mechanisms with Laplace, Gaussian, or exponential noise, and ensures privacy as explained before. Recent work shows that DP is vulnerable to poisoning attacks [61], [62], [63], [64] particularly when local differential privacy (LDP) is used in decentralized settings. In LDP the server is not trusted thus privacy constraints are enforced locally by the participants. An adversary can exploit LDP noise insertion to create fake examples (or impersonate real examples), and insert adversarial inputs into the computations performed by the server.
The adversary in attacks on privacy may focus on (i) adaptively querying the model to extract sensitive information, or (ii) DP poisoning the model during model fitting with the goal of distorting the model. In the latter case, the goal may be to corrupt the computation of a statistic at the server, such as frequency or moment estimation, or to perturb the minimizer of a sum of convex functions [61].

1) IMPLEMENTATION VULNERABILITIES
We note that DP has implementation flaws that can be exploited by an adversary [65], [66], [67]. The main concern is the implementation of random noise used in DP, for both the Laplacian and Gaussian mechanisms. In particular, Mironov [65] argues that sampling from all possible real numbers requires infinite precision and an unlimited source of entropy, which is not possible with finite precision computers. As a result, the standard floating-point sampling procedures will concentrate on a small subset of outputs while omitting many possible output values, thereby inducing porous noise distributions. For example, if the sensitive data are protected histogram counts then it is possible to reconstruct an entire database using adaptive queries [65]. Another attack uses timing information [66], [67]: because the duration of floating-point operations varies, an adversary can carefully craft adversarial input to leak information through this side-channel.

F. ATTACKS ON FAIRNESS
The main approaches for providing group fairness are preprocessing, in-processing, and post processing, all of which are subject to data poisoning attacks. Recently, generative models (GANs) are proposed to generate sanitized synthetic data sets to ensure fairness [68]. The synthetic data is used for training ML models in lieu of modifying or debiasing the original dataset. Another approach, by [69], uses variants of domain adaptation technique to enable fair learning.

These approaches do not consider the robustness of the synthetic data generation process and may be exploited through insertion of adversarial samples, because the generated data sets may inherit some properties of the original distribution used for building a GAN. This possibility makes it feasible for an adversary to conduct poisoning attacks through the sanitized data set.

IV. CAUSALITY AND TRUSTWORTHINESS
Our vision is that the trustworthiness of machine learning can be addressed in a uniform manner by moving from simply leveraging statistical associations to discovering and imposing causal relationships in the data. A recent body of work supports this vision and shows that explicitly modeling causality in machine learning can improve accuracy and generalization, robustness, fairness, and privacy — all the desiderata of trustworthy ML.

Causal models have the desirable properties of parsimony and modularity, and allow the answering of counterfactual questions [70], [71]. The first two properties lead to increased generalizability, robustness, and privacy. The ability to answer counterfactual questions, such as “would the model have returned the same classification if the value of the sensitive variable were different”, facilitates strong and subtle counterfactual notions of fairness. Additionally, causal models are attractive as a means for achieving trustworthiness because they aim to understand “true” cause-and-effect relationships between the observed data.

Causal structures can be partially or completely inferred from observational data; a large body of results focuses on this problem. More relevant to trustworthy machine learning, causality may be explicitly imposed to ensure parsimonious and modular models that support counterfactual reasoning. This latter approach is called causal representation learning, and is the subject of active research [72], [73], [74], [75].

A. MODELING CAUSALITY
Arbitrary joint distributions of $p$ random variables can be factorized in the form

$$P[X_1, \ldots, X_p] = \prod_{i=2}^{p} P[X_i | X_1, \ldots, X_{i-1}].$$

In these entangled factorizations, $X_i$ could be caused by variables that are before or after it in the specific ordering used. A causal factorization of the same distribution explicitly identifies the causes of each variable $X_i$, in the form

$$P[X_1, \ldots, X_p] = \prod_{i=1}^{p} P[X_i | PA_i],$$

where $PA_i$ denotes the parents of node $i$ in an associated directed acyclic graph (DAG) $\mathcal{G}$. When $P$ has a causal factorization corresponding to $\mathcal{G}$, it is Markovian with respect to $\mathcal{G}$: conditioned on its parents, each $X_i$ is independent of its non-descendants. Conversely, $P$ is said to be faithful with respect to $\mathcal{G}$ if every conditional independence in $P$ is encoded in $\mathcal{G}$.

A structural causal model (SCM) associated with $\mathcal{G}$ expresses the cause-effect relationships using functional relations of the form

$$X_i := f_i(PA_i, N_i),$$

indicating that each variable is determined by the values of its parent variables, and a noise variable $N_i$. Here, the exogenous noise terms $N_1, \ldots, N_p$ are assumed to be jointly independent of each other and the endogenous variables $X_1, \ldots, X_p$. The SCM model gives us an algorithmic description of how to sample from $P$: first sample the exogenous noise, then pass them through the functional relations to determine the corresponding values of $X_1, \ldots, X_p$. We assume that the DAG is causally sufficient, in that all relevant variables are included in the model; there are no confounders, or unaccounted for common causes. Figure 2 depicts a hypothetical causal graph for predicting student debt default for which the corresponding SCM model is

$$X_1 = N_1$$
$$X_2 = N_2$$
Causal machine learning uses the causal factorization of $P$ and involves the three levels of causal reasoning: (i) observational reasoning, corresponding to simple statements about $P$ such as $P(Y = y | X_6 = x)$ (i.e., probability of a particular loan default conditional on a specified amount of debt), (ii) interventional reasoning such as $P(Y = y | \text{do}(X_6 = x))$, corresponding to statements about the joint distribution that results from forcibly modifying one of random variables $X_i$ and allowing its changes to propagate through to alter the distribution of the other variables according to the causal factorization, i.e. $P(X = (x_1, \ldots, x_p) | \text{do}(X_i = x)) = \prod_{j=1, j\neq i}^{p} P(X_j = x_j | PA_j) \delta_{X_i=x}$, and (iii) counterfactual reasoning, which in essence allows you to specify a “universe” in which observations $X_i$ were made, by constraining the values of the $N_j$ to be consistent with those observations, then to answer interventional questions in this universe, for example $P(Y = y | X_2 = z, \text{do}(X_6 = x))$.

Observational, interventional, and counterfactual distributions are generically not equivalent. For example, observing $X_6$ constrains the likely values of $X_5$ because $X_6$ is an effect of $X_5$, which in turn affects the probability that $Y = y$ because it is a child of both $X_5$ and $X_6$; by contrast, intervening to set $X_6 = x$, does not affect the value of $X_5$, so the intervention and observations lead to different probabilities i.e. $P(Y = y | X_6 = x) \neq P(Y = y | \text{do}(X_6 = x))$. Similarly, if we are in a universe where $X_2 = 2$, then the value of $X_5$ is constrained as a child of $X_2$, so observing that $X_2 = 2$ and then intervening on $X_6$ affects the probability that $Y = y$ in a different manner than simply intervening on $X_6$, i.e. $P(Y = y | \text{do}(X_6 = x)) \neq P(Y = y | X_2 = 2, \text{do}(X_6 = x))$.

Observational reasoning requires only knowledge of $P$. Interventional reasoning requires at least a causal factorization of $P$. Counterfactual reasoning requires an SCM, to enable the following steps [70]: abduction, in which a posterior distribution on $N$ is computed given the evidence, action, in which the structural equations are modified according to the given intervention, then prediction in which the distribution of the non-intervened variables are computed using the posterior on $N$ and the modified structural equations.

### B. DISCOVERING CAUSALITY

In general, causal models for a set of random variables $X_1, X_2, \ldots, X_n$ cannot be learned from their joint distribution $P$ alone unless restricting assumptions are made [70], [71]. This implies that observational data alone is not enough to discover the underlying causal graphs. However, one can discover an equivalence set of causal graphs from the observational data, and use experiments to eliminate candidates to narrow down to the true causal graph.

The key observation used in uncovering causal structures is the equivalence of conditional independences in $P$ with the graphical criteria of $d$-separation. Assume that the distribution is Markovian and faithful with respect to the underlying DAG $G$. A foundational result due to Pearl [70] states that, in this case, $X_i \perp \!\!\!\!\perp X_j | S$ if and only if $X_i$ is d-separated from $X_j$ by the random variables in the set $S$. This result implies that a given DAG has statistical consequences that can be tested for in a given $P$.

One approach for identifying the causal structure is based on independence testing methods. Steps of an generic algorithm based in independence testing can be summarized as follow:

1. Gather i.i.d. samples from $P$
2. Use independence oracles to determine the conditional independences between $X_1, X_2, \ldots, X_p$
3. Assuming Markovianity and faithfulness, construct a DAG $G$ on $X_1, \ldots, X_p$, whose edges are consistent with the $d$-separations implied by the results of the independence tests. Some of these edges may be undirected.

Note that the output of this process is, in general, a partially directed acyclic graph (PDAG). This is because multiple Markov equivalent DAGs, differing only in the directionality of some edges, may be consistent with the results of the independence tests. Edges whose directionality cannot be determined from the conditional independence structures in the observed data become undirected edges in the PDAG.

Interventional experiments can be used to determine the directionality of these edges: e.g. if changing the distribution of $X$ causes that of $Y$ to change, then the edge would be directed from $X$ to $Y$. 

---

1. The simplified notation employed here is ambiguous in that it does not distinguish between counterfactual distributions and conditioned counterfactual distributions, e.g. one in which the counterfactual universe $X_5 = z, \text{do}(X_6 = x)$ is then conditioned on the observation $X_6 = x$. Conditioned counterfactual distributions do not appear in this paper.
Another popular approach is to propose causal graph structures and measure how well they fit the data. See [71] for further details on these score-based methods and hybrid methods that combine scoring and independence testing.

It is important to note that the learning of causal structures is potentially vulnerable to data poisoning attacks. We revisit this consideration later.

C. IMPOSING CAUSALITY: CAUSAL REPRESENTATION LEARNING

Causal discovery is closely related to classical variable selection: given a set of input variables and a target variable, what minimal set of inputs can be used to predict the target as well as the full set of input variables? If the causal graph describing these variables is known, then the answer is given by the intersection of the input variables and the Markov blanket of the target, which consists of the target’s parents, children, and spouses. Conversely, one approach to learning causal structures is to find an appropriate ordering of the variables under consideration and sequentially conduct variable selection to learn the parents of the current variable [71].

In many machine learning applications, however, variable selection directly on the input variables is ineffective. In image classification, for example, it is unlikely that the values of a small, fixed set of pixels allows one to determine whether an image contains a cat or not. In these applications, feature-engineering is essential to create a small set of features that are useful for predicting the target; e.g., in image classification, features that are convolutional functions of the input features have been found to be effective. In non-causal ML, these features are constructed by hand or in an automated manner to maximize accuracy.

A body of work has developed in which features are constructed to satisfy causal structures. The main impetus for these works has been the desire for out-of-distribution generalization. Causal machine learning is appropriate for this goal as it describes the behavior of a system of variables using a collection of independent causal mechanisms; often distributional shifts can be modeled as interventions changing the target, which consists of the target’s parents, children, and spouses. Conversely, one approach to learning causal structures is to learn models robust to adversarial perturbations [79].

Intuitively, each mechanism $f_i$ should reflect an cause-effect relationship that is independent of the other cause-effect relationships governing the transformation $X \rightarrow Z$. A decoder $D$ is trained so that $D(E(X)) \approx X$ on the training distribution. The resulting causal features $Z_i$ contain the same information as the raw features $X_k$, but are lower-dimensional and have favorable causal structure. This approach is popular, but has the disadvantages of not connecting the causal features with the target $Y$, not learning explicitly useful causal relations; and in general, such models are not uniquely identifiable from data.

Invariant risk minimization (IRM) is an alternative approach to causal learning which explicitly learns feature representations that encode environmental invariance [77]. In this paradigm, the data is collected from multiple training environments $D_e := \{(X_e', Y_e')\}$ governed by different laws $P_e(X_e', Y_e')$, with the assumption that the mechanism relating $Y^e$ to $X^e$ is invariant across environments. IRM proposes to learn that representation by finding a data representation $Z = f(X)$ that simultaneously minimizes the risks $\mathbb{R}^e(f) = E_{(X,Y) \sim D_e}(g(f(X)), Y)$ in all environments; here the minimization is done over the representation $f$ and the predictor $g$. The original work showed the existence of invariant predictors across all training environments $f$ when both $f$ and $g$ are constrained to be linear and the training environments satisfy some diversity condition.

The IRM paradigm has been critiqued and expanded significantly, with the most powerful results so far due to [78]. This work considers a general causal model in which $E$, $X$, $Y$, and $Z$ are variables in a causal DAG in which the observed random variables $X$ are caused by the latent variables $Z$; each $Z_i$ depends on one or both of $Y$ and $E$; and there are arbitrary relations between $Y$ and $Z$ as long as the DAG remains acyclic. In this setup, and under the assumption that $P[Z \mid Y, E]$ belongs to an exponential family with arbitrary natural parameters $\lambda(Y, E)$ and sufficient statistics $T(Z)$, the authors establish the identifiability of the latent factors $Z$.

D. ROBUSTNESS THROUGH CAUSALITY

One of the most commonly seen arguments for the use of causal machine learning is the fact that the principle of invariant causal mechanisms implies that causal models better generalize to out-of-sample distributions where the causal mechanism is the same even if the distributions of the causes changes. This is in contradiction to associational ML, which is vulnerable to spurious associations which fail to remain invariant when the test samples are not drawn from the same distribution as the training samples.

Specifically, researchers have advocated the deliberate insertion of interventions in causal representation models to learn models robust to adversarial perturbations [79]. By explicitly modeling changes of distribution using interventions, and requiring the mechanism $\hat{Y} = f(X)$ to be invariant under these interventions, one can expect that adversarial attacks that modify the covariate distribution to leverage spurious correlations will be less effective.
E. PRIVACY THROUGH CAUSALITY

Causal learning positively impacts the trade-off between privacy and accuracy [80]. To build an intuition for this, consider the process of privately learning an estimator \( \hat{Y} = f(X) \approx Y \) of \( Y \) by using gradient descent methods to minimize the risk \( \mathbb{E}(Y, \hat{Y}) \). Given a set of training data \( D \), private learning proceeds by using a randomized mechanism to estimate the gradient of the model loss,

\[
\mathcal{M}(D) \sim g(D) + \mathcal{N}(0, S_g^2 \sigma^2 I).
\]

Here \( g \) is the aggregated gradient of the risk on the training data, \( g(D) = \sum_{(x, y) \in D} g(x, y) \), and \( S_g \) is the sensitivity of the gradient on the training data, \( \max_{(x, y) \in D} \|g(x, y)\|_2 \).

By calibrating the noise to the sensitivity of the gradient, the influence of any specific data point on the fitted model is limited. A larger parameter \( \sigma \) increases the privacy guarantees. The DP-SGD algorithm manipulates \( \sigma \) and \( S_g \) (by clipping the gradients), to achieve a desired level of \((\epsilon, \delta)\)-differential privacy. At the same time, the accuracy of the resulting estimator is negatively impacted if the sensitivity is large or \( \sigma \) is large.

The sensitivity of the gradient is connected to the number of variables used in forming the estimate \( \hat{Y} \): *mutatis mutandis*, an estimator that uses more variables has a higher sensitivity than one that uses fewer. Causal ML models can use fewer variables than generic non-causal models— it is clear from the causal DAG that it suffices to use the Markov blanket of \( Y \) in forming the estimate \( \hat{Y}_c = f(X_{PA(Y) \cup CH(Y)}) \), rather than all available variables. Thus the sensitivity of causal estimators is lower; this implies that for a fixed noise level, the privacy guarantees are stronger for causal models. This in turn implies that there is a better trade-off between privacy and accuracy for causal models, once the causality structure is known.

For end-to-end differentially private causal ML, one must ensure that the model does not leak information about whose data was used in training to construct the causal DAG. The DP paradigm can be adapted here during the construction of the DAG. Thus the first step is to find causal structure in DP manner. For example, noise can be added during the conditional independence tests for PC algorithm. Score-based methods by computing scores in a DP-manner (see [81] for recovering two variable ANM models using DP scoring) are applicable here. Other work includes [82] which uses the exponential mechanism to privatize the PC algorithm for finding causal graphs. A followup work uses a two-stage approach in which some edges are coarsely eliminated using a portion of the privacy budget, then the remainder is used to eliminate more uncertain edges [83].

F. FAIRNESS THROUGH CAUSALITY

The commonly used group fairness approaches to defining and measuring fairness in machine learning define fairness in terms of associations between the output of the machine learning model and the sensitive attribute of its inputs; these forms of fairness are thus entirely based on the observational distribution \( \mathbb{P}_{X, Y, S} \). The strengths of this paradigm include that regularization and constrained optimization can be used to control these correlations, and that group fairness can be easily measured. However, group fairness has significant drawbacks. First, when a model is unfair, this paradigm offers no insight as to why, or how to remedy the unfairness. Second, group fairness gives guarantees at the population level, not at the level of individuals: this implies in particular, that models which are explicitly unfair at the level of individuals can be fair at the group level.

These drawbacks are consequences of the fact that group fairness inherently quantify associational properties of the model, rather than causal properties. Causal reasoning, in particular interventional and counterfactual reasoning, offer strong alternative definitions of fairness [84] that aim to ensure fairness in the causal model underlying the predictions made by ML models. In some cases, these notions of fairness subsume various forms of group fairness. Causal notions of fairness address the above drawbacks of group fairness: causal structures deliver insight into what causal paths result in undesired discrimination, offer remedies for observed unfairness, and can supply fairness at the individual level. A primary drawback to causal fairness approaches is that, because they use interventional and counterfactual distributions, they may not be identifiable: observational data alone may not suffice to ensure causal fairness. Recent work has studied when causal fairness is identifiable, and provided bounds on fairness in the case where it is not identifiable.

The concept of *counterfactual fairness* was formalized in [85] as follows: a predictor \( \hat{Y} \) is counterfactually fair if

\[
\mathbb{P}[\hat{Y} = y | X = x, S = s, \text{do}(S = s')] = \mathbb{P}[\hat{Y} = y | X = x, S = s, \text{do}(S = s)]
\]

for all evidence \( X = x \) and \( S = s \), any value \( y \), and any achievable sensitive value \( s' \). That is, a predictor \( \hat{Y} \) is counterfactually fair if the counterfactual effect of changing the sensitive attribute is zero. Intuitively, this means that \( \hat{Y} \) is counterfactually fair if, in all universes consistent with the observations, forcibly intervening to set the sensitive variable to any value does not affect the distribution of the predictor. In particular, this suggests that one way to get a counterfactual estimator of \( Y \) is to use only non-descendants of \( S \) in forming the predictor \( \hat{Y} \), and choose the predictor to minimize some loss between \( \hat{Y} \) and \( Y \). Indeed, [85] advocates that, given a causal model, one take the predictor to be of the form \( \hat{Y} = \mathbb{E}[g_\theta(n, x_{\neq S}) | x, s] \), where \( x_{\neq S} \) comprises the observed non-descendants of \( S \), the exogeneous random variables \( n \) are sampled from the posterior probability given the evidence, \( \mathbb{P}[N = n | X = x, S = s] \), and the function \( g_\theta \) is chosen to minimize the risk in estimating \( Y \). Thus, to apply this formulation of fairness, one must first have an SCM of the data generation process. An advantage of this approach to fairness is that once the SCM is given, one can readily construct a counterfactually fair predictor; another is that this is a notion of individual fairness.
The notion of counterfactual fairness introduced in [85] does not distinguish the paths along which the sensitive variable can affect $Y$. In many applications, it can be the case that $S$ may affect $Y$ along some causal paths in a manner that may be considered fair, while its effect along another path may be considered unfair. For example, consider the influence of gender ($X_1$) on employment ($X_5$) in Figure 2. The direct path $X_1 \rightarrow X_5$ is unfair, in that gender should not directly affect employment opportunities. However, the path $X_1 \rightarrow X_1 \rightarrow X_5$ in which the effect of gender on employment is mediated through education, may be considered fair, because different genders may choose to be educated in different fields, which then affects employment opportunities. The subtleties of which paths are considered fair mediators of causal effects must be determined by domain experts.

Given that a set of causal paths $\Pi$ has been determined to be unfair, the path-specific counterfactual effect (PSE) quantifies how much they mediate the causal effect of one variable upon another. Specifically, the path-specific counterfactual effect on $Y = y$ of changing the value of $S$ from $s$ to $s'$ along the paths on $\Pi$, given observations $X = x$, is defined as

$$P(Y = y | X = x, do(\Pi(S = s')) - P(Y = y | X = x, do(\Pi(S = s)))$$

where $\Pi$ consists of all the causal paths between $S$ and $Y$ that are not in $\Pi$, and $do(\Pi(S = s))$ indicates that the intervention $S = s$ is propagated only along the causal paths in $\Pi$. In this way, the PSE measures how much of the total causal effect of $S$ on $Y$ is mediated by the paths in $\Pi$. [86] proposes to learn a fair estimator $\hat{Y}$ by constructing a distribution $\hat{P}_{X,Y,S}$ that is close in KL-divergence to $P_{X,Y,S}$ and in which the unfair paths $\Pi$ have bounded PSE. In some instances, the PSE can be identified (i.e. determined from the observational distribution alone) [87], but in practice its computation involves intractable multiple integrals over conditional distributions, thus the constrained optimization involved in the approach taken in [86] is challenging for non-linear models. [88] sidesteps the issue of optimization by directly intervening on $S$, setting it to the baseline value $s$ along unfair paths $\Pi$ to estimate $\hat{Y}$ from the original causal model, thereby avoiding the need to compute or optimize the PSE.

G. ADVERSARIAL ATTACKS ON CAUSAL LEARNING

While causality improves the robustness of a machine learning model by focusing on independent mechanisms and distinguishing causally meaningful structures from spurious correlations, it introduces new mechanisms of attack for malicious actors. In particular, data poisoning during the learning of causal structures is a serious concern [89], [90].

1) CAUSAL ADVERSARY

Let $P$ be the joint distribution of the $p$ variables under consideration, and assume that $P$ is Markov and faithful with respect to a DAG $G$. The goal of the causal adversary is to find a distribution $Q$ that is close to $P$ such that $Q$ is Markov and faithful with respect to a DAG $H$ that is not Markov equivalent to $G$. We assume that adversary can sample $P$ thus DAG $G$ is not secret.

The power of the adversary can be measured in multiple ways, e.g. the KL-divergence $D_{KL}(P||Q)$ or total variation distance $||P - Q||_TV$; the measure should capture the intuition that it must be difficult to distinguish samples from the true distribution from samples from the adversarial distribution. We may also consider the performance of adversaries in the statistical setting, where $H$ is obtained from exact knowledge of the conditional independence structures of $Q$, or the finite data setting, where $H$ is obtained by running a causal discovery algorithm on a finite number of samples from $Q$. Especially in the latter case, it is practical to consider that the adversary can only poison a percentage of the data used for causal structure determination, so $Q = (1-\epsilon)P + \epsilon Z$ for some distribution $Z$ and corruption percentage $\epsilon \in (0, \frac{1}{2})$. Also, in the finite data setting, the adversary may exploit knowledge of the specific causal discovery algorithm to carry out more effective attacks.

The adversary’s goal in constructing $H$ may vary:

- **Model accuracy.** The adversary’s goal may be to ensure low accuracy. The fact that $P$ and $Q$ are close may make it difficult to make the estimation of marginals like $P[X_1, X_{ip}]$ too inaccurate, but the adversary may target either the accuracy of interventional distributions, or the accuracy of the causal graph itself.

- **Model robustness.** The adversary’s goal may be to reduce the robustness of a model, e.g. by replacing the true independent causal mechanisms that allow for domain invariance with inappropriate causal mechanisms. One example is causal ordering attacks in which the adversary aims to create an indistinguishability between $X \rightarrow Y$ vs $Y \rightarrow X$, or to alter the true direction of causality. For example, it can be difficult to determine the direction of causality in linear models with Gaussian noise; this could be utilized by an adversary.

- **Model fairness.** The adversary may introduce spurious causal links between sensitive variables and target variables, to ensure that the trade-off between fairness and accuracy is incorrectly judged when using fair learning procedures.

- **Model privacy.** The adversary may target privacy by interventions to discover if a particular individual is in $C$ from data, this is the only nontrivial independence relation, and they are jointly Gaussian with mean zero. It could be the case that $C$ is a common cause of $A$ and $B$ (i.e. $C \rightarrow A$ and $C \rightarrow B$), $B \rightarrow C \rightarrow A$, or $A \rightarrow C \rightarrow B$. Thus one must sample from the interventional distributions to figure out which case is the true DAG. The adversary can return samples from models with the same observational
distribution but for which the SHD is large (e.g. by returning samples from a jointly Gaussian model that reverses $A \rightarrow C \rightarrow B$ to $B \rightarrow C \rightarrow A$ while preserving the mean and covariances), to thwart the learning of the correct causal model.

Empirical approaches for poisoning causal structure learning have been considered in [89] and [90], respectively for the classic PC-structure learning algorithm and the hybrid CB algorithm. The results indicate that model invalidation attacks, in which the goal is to invalidate the correct causal model, can be achieved with many fewer changes than targeted attacks that attempt to manipulate specific edges. However, data poisoning attacks on causal structure learning are severely understudied, as are algorithms to defend against such attacks.

V. DISCUSSION AND CONCLUSION
We conclude with a few observations and recommendations on the design and analysis of trustworthy ML systems, followed by a review of benchmarks available for empirical characterization of trustworthiness along the multiple axes.

1) When designing and deploying TML systems, several axes of trustworthiness and the trade-offs among them should be considered in a vertically integrated data-to-knowledge ecosystem that includes: (i) rigorous quantification of the four desiderata as appropriate to the application, (ii) acquisition and appropriate representation of data, and (iii) testing and validation of the trustworthiness of the system. This design and validation should be revisited over time to maintain trust in the system.

2) Multiple definitions of fairness and privacy exist. It is important to ask for an appropriately strong measure of these aspects of trustworthiness: too weak guarantees allow vulnerability along these axes, while too strong guarantees implicitly limit achievability. The choice of which definitions and metrics to employ should be done in conjunction with a subject matter expert.

3) Associational machine learning is vulnerable to learning unintended biases or exploiting undesirable correlations between sensitive variables and the target variable, which may increase accuracy but hurt privacy and fairness and increase vulnerability to adversarial attacks. These issues can be mitigated by appropriate forms of regularization, but regularizations introduced to mitigate vulnerabilities in one aspect may increase vulnerabilities in other aspects of trustworthiness.

4) Missing or insufficient data from a sub-population produces biased and inconsistent results from ML and feature selection algorithms, compromising the predictive and discriminative value of the selected features. This reduces the accuracy of ML models and creates unfairness in predictive performance across group. Further, data quality and unbiasedness strongly limit the extent to which data-dependent measures of trustworthiness can be employed. Thus it is imperative that effort be extended to collect representative, unbiased training and validation data sets.

5) We advocate the use of causal machine learning modeling as a uniform approach to achieving the desiderata of trustworthy ML: causal models naturally support the identification and quantification of the relevance of the features used in causal learning problems, often provide more interpretable semantics for machine learning models, allow validation of (counterfactual) fairness of a system, tend to encourage learning models with less disclosure risk, and support the introduction of adversarial robustness through the use of interventions.

A. BENCHMARKS FOR EVALUATION
We noted above the need to test and validate TML systems using rigorous quantifications of the four desiderata. As argued throughout this paper, defenses for vulnerabilities along one axis of trustworthiness potentially introduce vulnerabilities along other axes; for this reason, evaluation of TML systems must be carried out simultaneously along all four desiderata to be convincing. This points to the urgent need for tools and platforms supporting the rigorous evaluation of multilateral TML trade-offs in a standardized manner.

Existing packages focus on accuracy-vs-X trade-offs, e.g. Fairlearn is an open source package which can be used to assess the fairness of ML models [91], and the RobustBench [92] library uses the AutoAttack [93] toolkit to evaluate the robustness-accuracy trade-off on more than 120+ model and data set combinations; RobustBench also maintains an up-to-date leaderboard of submitted defense schemes. However, there are currently no projects that readily enable rigorous and easily reproducible empirical examination of the trade-offs between multiple of these desiderata. Such a project would ideally include a diverse set of data sets, models, and tasks in keeping with the broad range of applications across which ML is employed. We briefly survey the availability of data sets that can currently be used to evaluate TML performance.

Several survey papers point to various data sets used for evaluating privacy attacks [94] and fairness attacks [30], [31] on machine learning systems. In [30] the authors not only evaluate various ML algorithms on a rich set of (tabular) data sets, but also build a Bayesian network to discover unfairness in the data itself, in the form of links from a sensitive variable to the target for prediction. Below we note a few common data sets briefly.

Several standard data sets exist for evaluating the fairness-accuracy trade-off, including the Bank Marketing Data set [95], the Adult and German Credit binary classification data sets, the Law School dataset, and the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) dataset [30].

Privacy-accuracy benchmarks are more varied, spanning from recommendation systems data sets such as Purchase [13] to location check-in data sets such as
Location [13], to medical data sets such as Texas [13], to standard image classification data sets including CIFAR 10 and 100.

To evaluate three-way trade-offs, one may use data sets that naturally call for multiple desiderata: for instance, the Texas data set [13] consists of medical discharge and procedure information for tens of thousands of patients, and is associated with the task of predicting the main procedure a patient underwent based on the patient’s information; for obvious reasons, a model deployed for such a task should exhibit a favorable fairness-privacy-accuracy three-way trade-off.

A recent study empirically evaluated the bias and risk trade-offs of a multi-task benchmarking pipeline for ML models [96]. Both data sets used in the study—MIMIC-III, the Medical Information Mart for Intensive Care, a database comprising deanonymized health care records of patients [97]; and STARR, the Stanford medicine Research data Repository [98]—are available upon request, and are natural benchmarks for evaluating fairness-privacy-accuracy trade-offs.

Facial recognition data sets are natural benchmarks for evaluating robustness-fairness-accuracy and robustness-privacy-accuracy trade-offs: one example is the re-annotated IJB-C data set that was previously used to evaluate race, age, and gender biases in facial recognition models in the FairFace Challenge at ECCV 2020 [99], [100].

REFERENCES

[1] A. Balogopalan, H. Zhang, K. Hamidieh, T. Hartvigsen, F. Rudzicz, and M. Ghassemi, “The road to explainability is paved with bias: Measuring the fairness of explanations,” in Proc. ACM Conf. Fairness, Accountability, Transparency, Jun. 2022, pp. 1194–1206.

[2] S. Yeom and M. Fredrikson, “Individual fairness revisited: Transferring techniques from adversarial robustness,” in Proc. 29th Int. Conf. Artif. Intell., Jul. 2020, pp. 437–443.

[3] S. Verma and J. Rubin, “Fairness definitions explained,” in Proc. Int. Workshop Softw. Fairness, May 2018, pp. 1–7.

[4] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, “A comparative study of fairness-enhancing interventions in machine learning,” in Proc. Conf. Fairness, Accountability, Transparency, Jan. 2019, pp. 329–338.

[5] S. Dwork and A. Roth, “The algorithmic foundations of differential privacy,” Found. Trends Theor. Comput. Sci., vol. 9, nos. 3–4, pp. 211–407, 2014.

[6] M. Gong, Y. Xie, K. Pan, K. Feng, and A. K. Qin, “A survey on differentially private machine learning,” IEEE Intell. Comput. Mag., vol. 15, no. 2, pp. 49–64, May 2020.

[7] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, J. Mironov, K. Talwar, and L. Zhang, “Deep learning with differential privacy,” in Proc. 2016 ACM SIGSAC Conf. Comput. Commun. Secur., 2016, pp. 308–318.

[8] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, “Local privacy and statistical minimax rates,” in Proc. IEEE 54th Ann. Symp. Found. Comput. Sci., Oct. 2013, pp. 429–438.

[9] S. Gopi, Y. T. Lee, and L. Wutschitz, “Numerical composition of differential privacy,” in Proc. Adv. Neural Inf. Process. Syst., vol. 34, 2021, pp. 11631–11642.

[10] M. Jafari and M. Karam, J. Mao, A. Oprea, A. Roth, S. S. Malvajerdi, and J. Ullman, “Differentially private fair learning,” in Proc. 36th Int. Conf. Mach. Learn. (ICML), vol. 97, 2019, pp. 3000–3008.

[11] T. Zhang, T. Zhu, K. Gao, W. Zhou, and P. S. Yu, “Balancing learning model privacy, fairness, and accuracy with early stopping criteria,” IEEE Trans. Neural Netw. Learn. Syst., pp. 1–13, 2021.

[12] R. Bassily, A. Smith, and A. Thakurta, “Private empirical risk minimization: Efficient algorithms and tight error bounds,” in Proc. IEEE 55th Annu. Symp. Found. Comput. Sci., Oct. 2014, pp. 464–473.

[13] D. Wang, M. Ye, and J. Xu, “Differentially private empirical risk minimization revisited: Faster and more general,” in Proc. Adv. Neural Inf. Process. Syst., vol. 30, Red Hook, NY, USA: Curran Associates, 2017, pp. 2722–2731.

[14] P. Gordaliza, E. Del Barrio, G. Fabrice, and J.-M. Loubes, “Obtaining fairness using optimal transport theory,” in Proc. 36th Int. Conf. Mach. Learn. (ICML), 2019, pp. 2357–2365.

[15] R. Jiang, A. Pacchiano, T. Stepelton, H. Jiang, and S. Chiappa, “Wasserstein fair classification,” in Proc. 35th Conf. Uncertainty Artif. Intell. (UAI), vol. 115, PMLR, 2019, pp. 862–872. [Online]. Available: http://proceedings.mlr.press/v115/jiang20a.html

[16] A. Shafahi, W. R. Huang, C. Studer, S. Feizi, and T. Goldstein, “Are adversarial examples inevitable?” in Proc. 7th Int. Conf. Learn. Represent. (ICLR), 2019, pp. 1–17.

[17] A. Farzawi, H. Farawi, and O. Farzawi, “Adversarial vulnerability for any classifier,” in Proc. Adv. Neural Inf. Process. Syst., vol. 31, 2018, pp. 1186–1195.

[18] Y.-Y. Yang, C. Rashchitcan, H. Zhang, R. R. Salakhutdinov, and K. Chaudhuri, “A closer look at accuracy vs robustness,” in Proc. Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 8588–8601.

[19] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, “Robustness may be at odds with accuracy,” in Proc. 7th Int. Conf. Learn. Represent. (ICLR), 2019, pp. 1–24.

[20] F. Miresghallah, M. Taram, P. Vepakomma, A. Singh, R. Raskar, and H. Esmaeilzadeh, “Privacy in deep learning: A survey,” 2020, arXiv:2004.12254.

[21] E. Bagdasaryan, “Differential privacy has disparate impact on model accuracy,” in Proc. Adv. Neural Inf. Process. Syst., vol. 32, 2019, pp. 15453–15462.

[22] H. Chang and R. Shokri, “On the privacy risks of algorithmic fairness,” in Proc. IEEE Eur. Symp. Secur. Privacy (EuroS P), Sep. 2021, pp. 292–303.

[23] S. Agarwal, “Trade-offs between fairness, interpretability, and privacy in machine learning,” Ph.D. dissertation, School Comput. Sci., Univ. Waterloo, Waterloo, ON, USA, 2020.

[24] H. Xu and S. Mannor, “Robustness and generalization,” Mach. Learn., vol. 86, no. 3, pp. 391–423, 2012.

[25] M. Cisse and S. Koyejo, “Fairness and representation learning,” in Proc. 33rd Conf. Neural Inf. Process. Syst., 2019, pp. 1–10.

[26] T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsi, “A survey on datasets for fairness-aware machine learning,” Wiley Interdiscip. Rev. Data Mining Knowl. Discovery, vol. 12, no. 3, 2022.

[27] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” ACM Comput. Surv., vol. 54, no. 6, p. 115, 2021.

[28] V. Nunda, S. Dooley, S. Singla, S. Feizi, and J. P. Dickerson, “Fairness through robustness: Investigating robustness disparity in deep learning,” in Proc. ACM Conf. Fairness, Accountability, Transparency, 2021, pp. 466–477.

[29] H. Xu, X. Liu, Y. Li, A. K. Jain, and J. Tang, “To be robust or to be fair: Towards fairness in adversarial training,” in Proc. 38th Int. Conf. Mach. Learn. (ICML), vol. 139, 2021, pp. 11492–11501.

[30] Q. Tian, K. Kuang, K. Jiang, F. Wu, and Y. Wang, “Analysis and applications of class-wise robustness in adversarial training,” in Proc. 27th ACM SIGKDD Conf. Knowl. Discovery Data Mining, Aug. 2021, pp. 1561–1570.

[31] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” in Proc. 6th Int. Conf. Learn. Represent. (ICLR), 2018, pp. 1–28.

[32] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaioui, and M. Jordan, “Theoretically principled trade-off between robustness and accuracy,” in Proc. Int. Conf. Mach. Learn., vol. 97, 2019, pp. 7472–7482.
ALEX GITTENS (Member, IEEE) received the Ph.D. degree in applied mathematics from Caltech, in 2013. He is currently an Assistant Professor with the Department of Computer Science, Rensselaer Polytechnic Institute, Troy, New York. He is a member of SIAM and AMS.

BÜLENT YENER (Fellow, IEEE) received the Ph.D. degree in computer science from Columbia University. He is currently a Professor with the Department of Computer Science and the Founding Director of the Data Science Research Center, Rensselaer Polytechnic Institute, Troy, New York. He is a Senior Member of the ACM.

MOTI YUNG (Fellow, IEEE) received the Ph.D. degree from Columbia University, in 1988. He is currently a Principal Scientist with Google and an Adjunct Senior Research Faculty at the Computer Science Department, Columbia University. He was a recipient of the IEEE Computer Society 2021 Women of the Eniac Computer Pioneer Award, the IEEE Computer Society 2018 W. Wallace McDowell Award, and the ACM’s SIGSAC 2014 Outstanding Innovation Award. He is a fellow of the ACM, The International Association for Cryptologic Research, and the European Association for Theoretical Computer Science.

* * *