Marrying Fairness and Explainability in Supervised Learning

PRZEMYSLAW GRABOWICZ, University of Massachusetts Amherst, USA

NICHOLAS PERELLO, University of Massachusetts Amherst, USA

AARSHEE MISHRA, University of Massachusetts Amherst, USA

Machine learning algorithms that aid human decision-making may inadvertently discriminate against certain protected groups. We formalize direct discrimination as a direct causal effect of the protected attributes on the decisions, while *induced* discrimination as a change in the causal influence of non-protected features associated with the protected attributes. The measurements of marginal direct effect (MDE) and SHapley Additive exPlanations (SHAP) reveal that state-of-the-art fair learning methods can induce discrimination via association or reverse discrimination in synthetic and real-world datasets. To inhibit discrimination in algorithmic systems, we propose to nullify the influence of the protected attribute on the output of the system, while preserving the influence of remaining features. We introduce and study post-processing methods achieving such objectives, finding that they yield relatively high model accuracy, prevent direct discrimination, and diminishes various disparity measures, e.g., demographic disparity.

1 INTRODUCTION

Discrimination consists of treating somebody unfavorably because of their membership to a particular group, characterized by a *protected attribute*, such as race or gender. Freedom from discrimination is outlined as a basic human right by the Universal Declaration of Human Rights. In the legal [1, 2] and social science [3–5] contexts, a key consideration serving as the basis for identifying discrimination is whether there is a disparate treatment or unjustified disparate impact on the members of some protected group. To prevent disparate treatment, the law often forbids the use of certain protected attributes, Z, such as race or gender, in decision-making, e.g., in hiring. Thus, these decisions, Y, shall be based on a set of relevant attributes, X, and should not depend on the protected attribute, Z, i.e., $P(y|x, z) = P(y|x, z')$ for any z, z', ensuring that there is no *disparate treatment.*[1] We refer to this kind of discrimination as *direct discrimination* (or lack of thereof), because of the direct use of the protected attribute Z.

Historically, the prohibition of direct discrimination was sometimes circumvented by the use of variables correlated with the protected attribute as proxies. For instance, some banks systematically denied loans and services, intentionally or unintentionally, to certain racial groups based on the areas they lived in [6, 7], which is known as the phenomenon of "redlining". In order to prevent such *inducement of discrimination*, the legal system of the United States has established that the impact of a decision-making process should be the same across groups differing in protected attributes [4, 5], that is $P(y|z) = P(y|z')$, unless there is a "justified reason" or "business necessity" for this *disparate impact* [1, 2].

---

[1]Throughout the manuscript we use a shorthand notation for probability: $P(y|x, z) \equiv P(Y = y|X = x, Z = z)$, where X, Y, Z are random variables, x, y, z are their instances, and P is a probability distribution or density.

Authors' addresses: Przemyslaw Grabowicz, University of Massachusetts Amherst, Amherst, Massachusetts, USA, grabowicz@cs.umass.edu; Nicholas Perello, University of Massachusetts Amherst, Amherst, Massachusetts, USA, nperello@umass.edu; Aarshee Mishra, University of Massachusetts Amherst, Amherst, Massachusetts, USA, aarsheemishr@umass.edu.

Example. The following example runs through the manuscript. Consider a hypothetical linear model of loan interest rate, Y. Prior works suggest that interest rates differ by race, Z [8, 9]. Some loan-granting clerks may produce non-discriminatory decisions, $y = \beta_0 - x_1$, while other clerks may discriminate directly, $y_{\mathrm{dir}} = \beta_0 - x_1 - z$ (see the graphical model in the top of Figure 1), where β_0 is a fixed base interest rate, x_1 is a relative salary of a loan applicant, x_2 is an encoding of the zip code (positive for wealth neighbourhoods, negative otherwise), while z encodes race and takes some positive (negative) value for White (non-White) applicants. If the protected attribute is not available (e.g., loan applications are submitted online), then a dis-

Fig. 1. An illustration of the graphical models that result from applying different learning methods to the example scenario: standard learning dropping Z (bottom left), the removal of influence of Z while preserving the influence of X (bottom right). The directed edges correspond to causal relations, while the dashed edge to a potentially unknown relationship, e.g., a non-causal association.

criminating clerk may induce discrimination in the interest rate, by using a proxy for race, $y_{\mathrm{ind}} = \beta_0 - x_1 - x_2$, where x_2 is the proxy. This case corresponds to the aforementioned real-world phenomenon of redlining. If we trained a model on the dataset $D = \{(x_1, x_2, z, y_{\mathrm{dir}})\}$ without using the protected attribute, since it is prohibited by law, then we would induce indirect discrimination in the same way as redlining. To see this point, assume for simplicity that all variables have a zero mean and there's no correlation between X_1 and Z and a positive correlation, $r > 0$, between X_2 and Z. If we applied standard supervised learning under the quadratic loss, then we would learn the model $\hat{y}_1 = \beta_0 - x_1 - z$. If we dropped the protected attribute, Z, before regressing Y_{dir} on the attributes X_1 and X_2, then we would learn the model $\hat{y}_2 = \beta_0 - x_1 - rx_2$, that induces via X_2 indirect discrimination growing in proportion to r.

Interdisciplinary challenge. There is a substantial and quickly growing literature on fairness in machine learning. However, its connection to the legal literature is underdeveloped, e.g., legal background is missing in the first textbook on fair machine learning (as of January 2022) [10], and business necessity is often neglected, which may be slowing down the widespread adoption of fair machine learning methods [11]. In supervised learning, potentially any feature that improves model predictiveness on deployment could be claimed to fulfil a business necessity. However, how does one prevent such features from being used for unintentional inducement of discrimination? This is a particularly acute problem for machine learning data-rich systems, since they often can find surprisingly accurate surrogates for protected attributes when a large enough set of legitimate-looking variables is available, resulting in discrimination via association [12]. Causality-based research offers so-called *path-specific counterfactual fairness* that enables designation of fair paths for business necessity [13–15], but these approaches rely on causal assumptions, arbitrary reference interventions, achieve sub-optimal model accuracy, and do not formally prevent induced discrimination via fair paths. Our study brings the concepts inspired by legal systems to supervised learning, which necessitates less assumptions and is used more widely than causal discovery, e.g., we make no assumptions about the relationship between X_2 and Z (dashed line in Figure 1). The big challenge in introducing non-discriminatory supervised learning algorithms in preventing direct discrimination without inducing indirect discrimination while enabling the necessity of businesses to maximizing model accuracy.
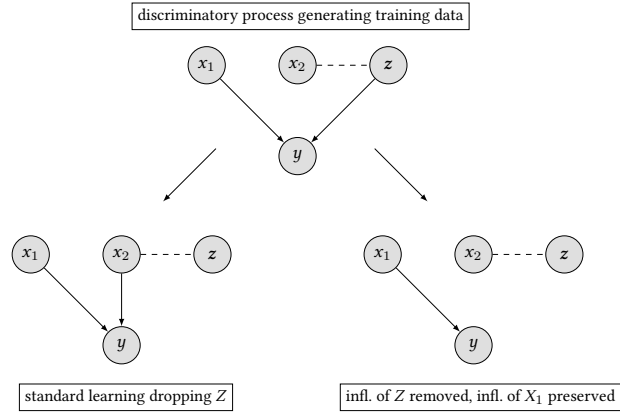
Contributions. To the best of our knowledge, this is the first study that fills the gap between fair supervised learning and legal systems by bridging causal notions of fairness with the literature on explainable artificial intelligence. We propose methods for removing direct discrimination from models that allows a limited use of features that prevents their use as a proxy for the protected attribute (the bottom right part of Figure 1). Specifically, first we define the concepts of direct, indirect, and induced discrimination via the measures of causal influence and tie them legal instruments. While doing so, we point out that induced discrimination can happen both for causal models of real-world decision-making processes and any other models that approximate such processes. Second, we construct loss functions that aim to remove the influence of the protected attribute, Z, while maintaining the influence of the remaining features, X, using a novel measure of marginal direct effect (MDE) and a well-known input influence measure (SHAP). Third, we show that dropping the protected attribute before training in standard supervised learning would result in increased influence of features associated with the protected attribute. Fourth, we introduce marginal interventional mixture models that drop Z while minimizing the inducement of discrimination through X. We show that this method keeps influence of X and Z close to the target values and, in addition, decreases popular disparity measures, while keeping high model accuracy.

2 RELATED WORKS

In machine learning, discrimination is typically defined based on statistical independence [16–27] or causal relations [28–31]. Well-known fairness objectives, such as parity of impact and equalized odds, correspond or are related to the statistical independence between Z and Y [26]. However, legal systems allow for exceptions from this independence through the business necessity clause, which permits usage of an attribute X associated with Z and results in the decisions Y depending on Z through X if it fulfils certain business necessity. Hence, the notions of discrimination based on the statistical independence between Y and Z are misaligned with their legal counterparts [32], which results in shortcomings. For instance, the algorithms that put constraints on the aforementioned disparities in treatment and impact [16–18] could negatively affect females with short hair and/or programming skills, because of those features' (fair or unfair) association with males [33].

A relevant line of research proposes to define direct and indirect discrimination as direct and indirect causal influence of Z on Y, respectively [30, 34]. While this notion of direct discrimination is consistent with the concept of disparate treatment in legal systems, the corresponding indirect discrimination is not consistent with them, since the business necessity clause allows the use of an attribute that depends on the protected feature (causally or otherwise), if the attribute is judged relevant to the decisions made, e.g., as in the seminal court case of Ricci v. DeStefano [35]. *Path-specific* notions of causal fairness address this issue to a limited extent [13–15]. These methods introduce *fair causal paths*, i.e., the paths through which the impact of the protected attribute is permitted, hence enabling business necessity. However, if there is no limit on the influence that can pass through such a path, then the path can be used for discrimination, as in the aforementioned case of *redlining*. This limit is not a focus of prior works [13–15, 28–31], but it is crucial to prevent induced discrimination in machine learning. In addition, for the removal of protected attributes these works rely on causal assumptions and a reference intervention, which is a standard technique in causality literature, but the reference intervention is arbitrary and may decrease model accuracy, as we show in Section 4.3.1. To the best of our knowledge, this work is the first to define and inhibit induced discrimination in supervised learning on the grounds of causality and explainability research.

3 PROBLEM FORMULATION OF FAIR AND EXPLAINABLE LEARNING

Consider decisions Y that are outcomes of a process acting on non-protected variables X and protected variables Z, where $x \in \mathcal{X}$, $z \in \mathcal{Z}$, $y \in \mathcal{Y}$, i.e., the variables can take values from any set, e.g., binary or real. Protected and non-protected features are indexed, e.g., X_i corresponds to the i'th feature (component). The decisions are generated via a function $y = y(x, z, \epsilon)$, where ϵ is an exogenous noise variable. Since the exogenous noise is unpredictable, we focus on the de-noised function $y(x, z) = \mathbb{E}_\epsilon \, y(x, z, \epsilon)$ for notational simplicity. The process generating decisions Y corresponds either to a real-world causal mechanism or its model, while the inducement of indirect discrimination shall be prevented on legal grounds in either case (see Subsection 3.1.2). These decisions can represent any decision-making process, e.g.: i) estimates of recidivism risk for a crime suspect, given some information about their prior offenses x and their race z, or ii) credit score assignments for a customer, given their financial record x and their gender z.

The goal of standard supervised learning is to obtain a function $\hat{y} : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ that minimizes an expected loss, e.g., $\mathbb{E}[\ell(Y, \hat{y}(X, Z))]$, where the expectation is over the set of training samples (x, z, y) and ℓ is a loss function such as quadratic loss, $\ell(y, \hat{y}) = (y - \hat{y})^2$. If the dataset is tainted by discrimination, then a data science practitioner may desire, and, in principle, be obliged by law, to apply an algorithm that does not perpetuate this discrimination. For example, Y could correspond to past hiring decisions, which we now want to automate with model $\hat{Y}$. If historical discrimination in hiring took place, then Y would be tainted, and a suitable fair machine learning algorithm would be needed. In this setting, $\hat{Y}$ can be *altered* w.r.t. the model of the original decisions Y to prevent discrimination. The crucial question is how to drop Z from the model without inducing discrimination, that is, without increasing the impact of attributes X associated with Z in an unjustified and discriminatory way.

We propose that a non-discriminatory model shall remove the influence of the protected features Z on Y, while preserving the influence of the remaining attributes X on Y. This method allows addition of features to the model that increase model predictiveness, while preventing them from being used inadvertently as proxies for the protected features. To preserve influence of non-protected attributes, we define and minimize special loss functions. Such losses can be constructed on the grounds of causal influence (CDE, MDE), or model input influence or feature relevance measures (SHAP). If there are many non-protected attributes, then the influence can be preserved for each of them separately or all of them together; we study both cases.

3.1 Legal notions and responsibility for decision-making models

Before we deep dive into mathematical definitions of respective loss functions, we first define a couple of abstractions of legal instruments by tying them to decision-making models and discuss legal responsibility for a model.

3.1.1 Legality of the influence of protected features and their relationships with other attributes. We define unfair influence and fair relationship between protected attributes Z and decisions Y by tying them to legal instruments, i.e., legal terms of art that formally express a legally enforceable act.

Definition 1. Unfair influence is an influence of protected feature(s) Z on specified type of decisions Y that is judged illegal via some legal instrument.

For instance, the U.S. Civil Rights Act of 1968 (Titles VIII and IX, known as Fair Housing Act) [2] determines that decisions about sale, rental, and financing of housing shall not be influenced by race, creed, and national origin; the U.S. Civil Rights Act of 1964 (Title VII) [1] determines that hiring decisions shall not be influenced by race, color, religion, sex, and national origin.
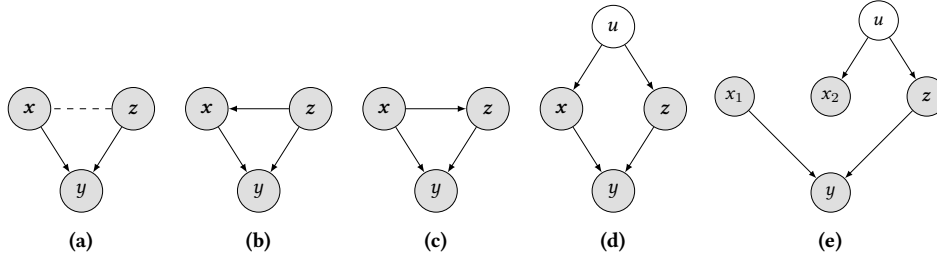
Fig. 2. The considered setting. We make no assumptions about the relations between x and z (marked with a dashed edge), nor their components. Hence, the graph (a) includes all exemplary cases (b-e). The graph (e) depicts the data-generating process from Example and shows no relationship between the components x_1 and x_2 of x. The random variable u is an exogenous noise, i.e., an unmeasured independent random variable.

In the context of making decisions Y using features X, some of the features may be associated with, or affected by, the protected attribute Z. Some of such features may be legally admissible for use in the decision-making if they are not *unfairly influenced*, are relevant to decisions Y, and fulfil a business purpose.

Definition 2. Fair relationship of protected feature(s) Z with non-protected feature(s) X is a relationship in the context of making decisions Y that is judged legal via some legal instrument, e.g., business necessity clause.

For instance, in graduate admissions to University of California Berkeley it was found that females were less often admitted than males [36]. However, females applied to departments with lower admission rates than males and the overall admission process was judged legal. If we represent department choice with X, then we could use this feature in the model of admission decisions Y, despite the fact that X is causally influenced by gender. Prior research shows that features perceived as fair tend to be volitional [37], as in the above example.

From the perspective of supervised learning, the definitions of unfair and fair influence are exclusion and inclusion rules, respectively, determining which features are legally admissible in the model of Y. Legal texts typically clearly define unfair influence, but fair relationships are determined on case-by-case basis. It is a reasonable to assume that the purpose of business is to develop a model that on deployment is the most predictive possible. One could argue that any feature that is predictive of Y and different than Z fulfills business necessity and is fair to use. However, some of such features may be affected by *unfair influence*. In such cases, one can remove Z from the unfairly influenced X and, then, from Y, without inducing discrimination (see Section 4.3).

3.1.2 Legal responsibility for a decision-making model vs. its causal interpretation. To determine responsibility for potentially harmful decisions, legal systems consider the epistemic state of decision-makers [38, 39], e.g., whether an employer knew about discrimination in company's hiring process, and their intentions [4], i.e., the employer may be expected to do their due diligence to identify discrimination and to correct their hiring process given their knowledge. In the context of decision-making models, the epistemic state corresponds to a potentially discriminatory model Y of the respective real-world decision-making, whereas intentions correspond to learning objectives, methods, and feature selection that result in a discriminatory model Y and a desired non-discriminatory $\hat{Y}$. The first step towards developing non-discriminatory models is finding accurate and robust, potentially causal [40–42], models of discriminatory decisions in close collaboration with domain experts. Machine learning models are developed in best faith to maximize accuracy, but often are not causal and not robust to covariate shifts [43, 44], i.e., they constitute an inaccurate epistemic state. Unfortunately, in practice it may be impossible to test causal validity of model $y(x, z)$, because of limited and unobserved

data, privacy concerns, and the infeasibility or prohibitive costs of causal experimentation. In such situations, legal systems may acquit model developers if the intentions and reasoning behind the development process of models of Y and $\hat{Y}$ were legally admissible, despite the incorrect epistemic state. Either way, whether the model at hand does or does not represent causal relations between variables in the real world, the model is causal w.r.t. its own predictions and the parents of these predictions are X and, possibly, Z, as detailed in the causal explainability literature [45]. That model can suffer the effects of training on discriminatory data. In the reminder of this paper, we use Y to refer both to the causal process and its model, since the two are the same in the former "ideal" causal setting, but our reasoning and approach is applicable to the latter "practical" non-causal settings as well, since the induction of indirect discrimination is questionable on legal grounds, i.e., decision-maker's epistemic state may be incorrect, but their intentions shall be good (to identify and prevent discrimination using reasonable methods).

3.2 Problem formulation based on causal influence measures

Formal frameworks for causal models include classic potential outcomes (PO) and structural causal models (SCM) [46]. Other frameworks, such as segregated graphs [47] and extended conditional independence [48] generalize the classic frameworks, e.g., they introduce undirected and bidirectional causal relationships. The methods proposed here rely only on the notion of intervention, which tends to have a consistent meaning across causal frameworks.

The following formulas are for the graphs depicted in Figure 2, where all variables are observed. We assume that there are direct causal links from X and Z to Y. If this assumption does not hold, e.g., because supervised learning is used for nowcasting instead of forecasting, then the following methodology may suffer collider bias (Berkson's paradox) [41, 42]. For instance, if the underlying causal graph is $Y \rightarrow X \leftarrow Z$, then conditioning on X makes Y and Z depend on each other, despite the fact that Z does not causally influence Y, so supervised learning based on samples (x, z, y) would yield a model in which Z unfaithfully influences the model of Y. We make no assumptions about the relations between X and Z and their components (Figure 2a), e.g., these relations may be direct causal links (Figure 2b-2d) or associations (Figure 2e). Finally, it is assumed that there are no unmeasured confounders.

In the notation of SCM and PO, the potential outcome for variable Y after intervention $do(X = x, Z = z)$ is written as $Y_{x,z}$, which is the outcome we would have observed had the variables X and Z been set to the values x and z via an intervention. The causal *controlled direct effect* on Y of changing the value of Z from a reference value z to z' given that X is set to x [46] is

$$\text{CDE}_Y(z', z|x) = \mathbb{E}[Y_{x,z'} - Y_{x,z}]. \tag{1}$$

Next, we define direct, indirect, and induced discrimination by tying the causal concept of controlled direct effect to the notions of *fair influence* and *unfair relationship*, which are abstractions of respective legal concepts.

Definition 3. Direct discrimination is an *unfair influence* of protected attribute(s) Z on the decisions Y and $\exists_{z,z'}\exists_x \text{CDE}_Y(z, z'|x) \neq 0$.

Definition 4. Indirect discrimination is an influence on the decisions Y of feature(s) X whose *relationship* with Z is not *fair* and $\exists_{x,x'}\exists_z \text{CDE}_Y(x, x'|z) \neq 0$.

To remove direct discrimination, one can construct a model $\hat{Y}$ that does not use Z. However, the removal of direct discrimination may induce discrimination via the attributes X_i associated with the protected attributes Z, even if there is no causal link from Z to X_i.

Definition 5. Discrimination induced via X_i is a transformation of the process generating Y not affected by direct and indirect discrimination into a new process $\hat{Y}$ that modifies the influence of certain X_i depending on Z between the processes Y and $\hat{Y}$ in the sense that $\exists_z \exists_{x,x'} \text{CDE}_Y(x, x'|z) \neq \text{CDE}_{\hat{Y}}(x, x'|z)$ given that $P(x|z) \neq P(x)$ or $P(x'|z) \neq P(x')$.

Example. Consider the aforementioned linear models of loan interest rate, $\hat{y}_1$ and $\hat{y}_2$. Note that $CDE_{\hat{y}_1}(x, x'|z) - CDE_{\hat{y}_2}(x, x'|z) = r * (x_2 - x_2')$, since x has two components x_1 and x_2 and the first component is reduced, so the model $\hat{y}_2$, that drops the protected attribute, induces indirect discrimination via X_2, because X_2 serves as a proxy of Z.

Following causal inference literature [46], to measure the extent of induced discrimination, we introduce natural direct and indirect effects using nested counterfactuals, $Y_{X_z, z'}$, denoting the outcome that would have been observed if Z were set to z' and X were set to the value it would have taken if Z were set to z. *Natural direct effect* of changing the value of Z from a reference value z to z' is

$$\text{NDE}_Y(z', z) = \mathbb{E}[Y_{X_z, z'} - Y_{X_z, z}]. \tag{2}$$

However, the measure NDE faces some challenges: to see this, consider the graphs in Figure 2. For the graph in Figure 2b the *causal mediation formula* [46] yields

$$\text{NDE}_Y(z', z) = \mathbb{E}[Y_{X_z, z'} - Y_{X_z, z}] = \underset{X \sim P(X|z)}{\mathbb{E}}[Y|X, z'] - \underset{X \sim P(X|z)}{\mathbb{E}}[Y|X, z]. \tag{3}$$

For the graphs in Figures 2c and 2d it yields a different value,

$$\text{NDE}_Y(z', z) = \underset{X \sim P(X)}{\mathbb{E}}[Y|X, z'] - \underset{X \sim P(X)}{\mathbb{E}}[Y|X, z], \tag{4}$$

because in this case X is not causally affected by Z and, hence, here the expectations are over $P(X)$ instead of $P(X|z)$. These expectations come from the nested interventions, i.e., X_z. We argue that the direct effect of Z on Y shall not depend on the direction of the causal link between X and Z. Furthermore, the choice to use X_z as a reference value in the definitions of natural direct effects is arbitrary, e.g., one could use $X_{z'}$ instead. To address these two issues, we introduce a *marginal direct effect* as

$$\text{MDE}_Y(z', z) = \mathbb{E}[Y_{X'', z'} - Y_{X'', z}] = \underset{X \sim P(X)}{\mathbb{E}} \text{CDE}_Y(z', z|X) = \underset{X \sim P(X)}{\mathbb{E}}[Y|X, z'] - \underset{X \sim P(X)}{\mathbb{E}}[Y|X, z], \tag{5}$$

which takes an expectation over the probabilistic interventions on X, setting its value to random samples of X'', where X'' is a variable independent from all other variables, but has the same marginal distribution as X. This measure yields the same value for all graphs in Figure 2. Then, to preserve influence of non-protected attributes we can minimize the following loss

$$L_{\text{MDE}}(X) = \underset{X'', X}{\mathbb{E}} \ell(\text{MDE}_Y(X, X''), \text{MDE}_{\hat{Y}}(X, X'')). \tag{6}$$

or its feature-specific version, which computes the loss separately for each component of X,

$$L_{\text{MDE}}^{\text{IND}}(X) = \sum_i L_{\text{MDE}}(X_i) = \sum_i \underset{X_i'', X_i}{\mathbb{E}} \ell(\text{MDE}_Y(X_i, X_i''), \text{MDE}_{\hat{Y}}(X_i, X_i'')). \tag{7}$$

A similar loss could be constructed based on the comparison between $\text{CDE}_Y(X, X''|Z)$ and $\text{CDE}_{\hat{Y}}(X, X''|Z)$. In this paper we focus on losses based on MDE, or SHAP input influence measure described next.

3.3 Problem formulation based on input influence measures

Alternatively, influence can be measured on the grounds of input influence measures introduced to explain black-box AI models. For the purpose of this section, we introduce a concatenation of variables X and Z as $W = XZ$, i.e., samples of W are tuples $w = (x, z)$ and $w \in \mathcal{X} \times \mathcal{Z} = \mathcal{W}$. Components of W are indexed, e.g., W_i is the i-th feature among the set $\mathcal{F}$ of all protected and non-protected features, i.e., $i \in \mathcal{F}$. To measure the influence of a certain feature W_i prior works suggest to make a probabilistic intervention on that variable by replacing it with an independent random variable [45, 49, 50]. In particular, let primed variables have the same joint distribution as the non-primed variables, $\forall_{w \in \mathcal{W}} P(W' = w) = P(W = w)$, while being independent from them, $W' \perp W$. Let double primed variables have the same marginal distributions as the non-primed variables, $\forall_{i \in \mathcal{F}} \forall_{w \in \mathcal{W}_i} P(W_i'' = w) = P(W_i = w)$, and be independent from each other and the non-primed variables, i.e., $\forall_{i \in \mathcal{F}} \forall_{j \neq i} W_i'' \perp W_j''$, $W'' \perp W'$ and $W'' \perp W$.

For any subset of features T that does not contain i, prior works define a marginal influence (MI) using W' as a random baseline [45, 49],

$$\mathrm{MI}_Y(W_i|w, T) = \underset{W'}{\mathbb{E}} \left[Y_{w_{T \cup \{i\}} W'_{-(T \cup \{i\})}} - Y_{w_T W'_{-T}} \right],$$

where the random variable $W_T W'_{-T}$ represents a concatenation of random variables W_T and $W'_{-T} = W'_{\mathcal{F} \setminus T}$, which amounts to a modified W with its components W_i, for each $i \in \mathcal{F} \setminus T$, replaced by the respective components of W'; likewise $w_T W'_{-T}$ is a concatenation of sample w_T and random variable W'_{-T}.

A popular measure of the influence of input w_i is based on the Shapley value (SHAP), which averages the marginal influence over all possible subsets T of all features except for i [49, 50],

$$\mathrm{SHAP}_Y(w_i|w) = \sum_{T \subseteq \mathcal{F} \setminus \{i\}} \frac{\mathrm{MI}_Y(W_i|w, T)}{|\mathcal{F}| \binom{|\mathcal{F}|-1}{|T|}}. \tag{8}$$

For instance, for the case of two variables,

$$\mathrm{SHAP}_Y(x|x, z) = \underset{X', Z'}{\mathbb{E}} [(Y_{x,z} - Y_{X',z} + Y_{x,Z'} - Y_{X',Z'})/2]. \tag{9}$$

Then, to preserve influence of non-protected attributes we can minimize the respective loss,

$$L_{\mathrm{SHAP}}(X) = \underset{X}{\mathbb{E}} \ell(\underset{Z''}{\mathbb{E}} \mathrm{SHAP}_Y(X|XZ''), \underset{Z''}{\mathbb{E}} \mathrm{SHAP}_{\hat{Y}}(X|XZ'')), \tag{10}$$

or its feature-specific version,

$$L_{\mathrm{SHAP}}^{\mathrm{IND}}(X) = \sum_i L_{\mathrm{SHAP}}(X_i) = \sum_i \underset{X}{\mathbb{E}} \ell(\underset{Z''}{\mathbb{E}} \mathrm{SHAP}_Y(X_i|XZ''), \underset{Z''}{\mathbb{E}} \mathrm{SHAP}_{\hat{Y}}(X_i|XZ'')). \tag{11}$$

While here we have constructed loss functions based on SHAP, other input influence measures, such as PFI or SAGE, can be used as well [51–55]. We leave the exploration of other losses for future works.

4 LEARNING FAIR AND EXPLAINABLE MODELS

We seek models $\hat{Y}$ that remove the influence of the protected attributes Z, while preserving the influence of non-protected attributes X by minimizing $L_{\mathrm{MDE}}(X)$ or $L_{\mathrm{SHAP}}(X)$, which lead to a simple closed-form solution, or their feature-specific versions, i.e., $L_{\mathrm{MDE}}^{\mathrm{IND}}(X)$ or $L_{\mathrm{SHAP}}^{\mathrm{IND}}(X)$, which we solve via transfer learning. Either of these approaches can be used to remove direct or indirect discrimination (see example in Subsection 4.3).

4.1 Minimizing $L_{\mathbf{MDE}}(X)$ or $L_{\mathbf{SHAP}}(X)$

Definition 6. Interventional mixture of a model $y(x, z)$ w.r.t. attribute Z is a model $\hat{y}_\pi(x) = \mathbb{E}_{\tilde{Z}} \, \hat{y}(x, \tilde{Z})$, where $\tilde{Z}$ is a random variable independent from all other variables, has the same support as Z, and a distribution $\pi(\tilde{Z})$.

Marginal interventional mixture (MIM) is $\hat{y}_{\mathrm{MIM}}(x) = \mathbb{E}_{Z'} \, \hat{y}(x, Z')$.

Proposition 1. For variable Y, the objective $L_{\mathrm{MDE}}(X)$ is minimized by the MIM.

PROOF. $L_{\mathrm{MDE}}(X) = \mathbb{E}_{X'', X} \, \ell(\mathbb{E}_{Z'}[Y_{X, Z'} - Y_{X'', Z'}], \mathbb{E}_{Z'}[\hat{Y}_{X, Z'} - \hat{y}_{X'', Z'}])$, so for $\hat{y}_{\mathrm{MIM}}(x) = \mathbb{E}_Z \, y(x, Z)$ it is zero. □

Proposition 2. For a real-valued and analytic $y(x, z)$, the MIM is an interventional mixture that minimizes the objective $L_{\mathrm{SHAP}}(X)$.

PROOF SKETCH. Without loss of generality, for simplicity let us consider the case of two variables X and Z. Let us expand $y(x, z)$ into a Taylor series around the point $x = 0, z = 0$. The series is a sum of components $Cx^k z^l$, where C is a constant and k and l are integers from 1 to ∞. Then, we replace Y in the definition of L_{SHAP} with the Taylor series and make a proof by induction. Minimizing this objective gives a potentially infinite set of conditions $\mathbb{E}[\tilde{Z}^l] = \mathbb{E}[Z^l]$ for the respective moments of $\tilde{Z}$. Since l can be any positive integer, these conditions are met if $P(\tilde{Z} = z) = P(Z = z)$. The full proof is in Appendix A. □

Example. We continue the example of a model for loan interest rate. The full model is $y(x, z) = \beta_0 - x_1 - z$. The MIM is $\hat{y}_{\mathrm{MIM}} = \beta_0 - x_1 - \mathbb{E}_Z \, Z$.

4.2 Minimizing $L_{\mathbf{MDE}}^{\mathbf{IND}}(X)$ and $L_{\mathbf{SHAP}}^{\mathbf{IND}}(X)$ via transfer learning

The minimization of the feature-specific losses, $L_{\mathrm{MDE}}^{\mathrm{IND}}(X)$ and $L_{\mathrm{SHAP}}^{\mathrm{IND}}(X)$, does not result in closed-form solutions, so we apply a respective gradient descent. First, we drop the protected attribute(s) Z from the data. We then obtain the "Trad. w/o Z" model by minimizing the cross entropy loss, $H(\hat{y}, y) = -\sum_i y_i \log \hat{y}_i$. Next, we optimize for either $L_{\mathrm{MDE}}^{\mathrm{IND}}(X)$ or $L_{\mathrm{SHAP}}^{\mathrm{IND}}(X)$. For both objectives we use ℓ_2 loss. We refer to these two-stage optimization-based methods as OPT-MDE and OPT-SHAP, respectively. The training is done using momentum based gradient optimizer ADAM [56] via batch gradient descent. We fine-tune two hyper-parameters: learning rate (α) and number of epochs (N). During fine-tuning we pick the values for which we get the best performance on the validation set. In our datasets, α is $1e - 3$ to $1e - 2$ and N is from 20 to 100. Our implementations of the methods will be released publicly as a Python library.

4.3 Removal of indirect discrimination via nested use of proposed methods

Potentially any feature that is predictive of Y and different than Z could fulfill business necessity, as we pointed in Subsection 3.1.1. However, a feature X_i can be *unfairly influenced* and, hence, illegal. Then, we have two options: i) not include feature X_i in the model of Y or, ii) create a model of X_i, remove from it the impact of Z, and then use the corrected $\hat{X}_i$ in the model of Y and drop the impact of Z, while using either of the proposed methods for removing the impact of Z from the models of X_i and Y. In the next section, we exemplify the latter option using MIM, while comparing it with counterfactual fairness.

4.3.1 Comparison with path-specific counterfactual fairness. In contrast to our proposed methods, path-specific counterfactual fairness (PSCF) requires the knowledge of a full causal graph. Hence, we study an exemplary linear model

introduced in the PSCF paper [14]. We maintain the original notation:

$$M = \theta^m + \theta_z^m Z + \theta_c^m C + \epsilon_m, \tag{12}$$

$$L = \theta^l + \theta_z^l Z + \theta_c^l C + \theta_l^l M + \epsilon_l, \tag{13}$$

$$Y = \theta^y + \theta_z^y Z + \theta_c^y C + \theta_y^y M + \theta_l^y L + \epsilon_y, \tag{14}$$

where C, M, L are components of $\boldsymbol{X}$, while ϵ_c, ϵ_m, ϵ_l are exogenous noise variables. The causal influence of Z on decisions Y and the mediator M is assumed unfair and all other influences are fair. In other words, Y is affected by direct discrimination via Z and indirect discrimination via M. This means that the MIM needs to be applied first to M and then to Y. Same as PSCF, the MIM corrects "the decision through a correction on all the variables that are descendants of the sensitive attribute along unfair pathways". Thus, we first apply the MIM to get a non-discriminatory $\hat{m}_{\text{MIM}}$, then we propagate $\hat{m}_{\text{MIM}}$ to its descendants, and finally apply MIM to get $\hat{y}_{\text{MIM}}$,

$$\hat{m}_{\text{MIM}} = \theta^m + \theta_z^m \overline{z} + \theta_c^m c = m - \theta_z^m (a - \overline{z}), \tag{15}$$

$$\hat{l}_{\text{MIM}} = \theta^l + \theta_z^l z + \theta_c^l c + \theta_m^l \hat{m}_{\text{MIM}}, \tag{16}$$

$$\hat{y}_{\text{MIM}} = \theta^y + \theta_z^y \overline{z} + \theta_c^y c + \theta_m^y \hat{m}_{\text{MIM}} + \theta_l^y \hat{l}_{\text{MIM}}, \tag{17}$$

where $\overline{z}$ stands for the mean of Z.

A comparison with PSCF reveals that $\hat{y}_{\text{MIM}} = \hat{y}_{\text{PSCF}} + \Delta$, where $\Delta = \overline{z}(\theta_z^y + \theta_m^y \theta_z^m + \theta_l^y \theta_m^l \theta_z^m)$. In fact, the mean squared error w.r.t. Y is larger for PSCF than for MIM by the the square of the difference, i.e., $\mathbb{E}(Y - \hat{Y}_{\text{PSCF}})^2 = \mathbb{E}(Y - \hat{Y}_{\text{MIM}})^2 + \Delta^2$. PSCF is based on NDE (Equation 2), it was introduced for binary Z, and relies on a choice of reference value, z', also known as baseline, which is assumed $z' = 0$ in the above example. However, this choice is arbitrary and it is not clear what baseline should be for non-binary Z. By contrast, the MIM introduces a distribution $\pi(z')$ over the reference intervention, which mimics "probabilistic interventions" from explainability literature [45, 49]. This difference between PSCF and MIM mirrors the difference between NDE and MDE, respectively, and it leads to Δ. Thanks to this, the MIM can be applied to continuous Z and it results in more accurate models. The above result that MIM is at least as accurate as PSCF is true for any linear model and any choice of the reference value z'.

5 RESULTS OF EXPERIMENTS

We examine the performance of our method and other supervised learning methods addressing discrimination in binary classification on synthetic and real-world datasets. We measure $\mathbb{E}_{\boldsymbol{X}, Z'} |\text{SHAP}_Y(X_i | \boldsymbol{X}, \boldsymbol{Z}')|$, following the measure of global feature influence proposed by Lundberg and Lee [50], and $\mathbb{E}_{X_i, X_i'} |\text{MDE}_Y(X_i, X_i')|$, both of which are evaluated using outcome probabilities. Note that these measures are different than our loss functions, which make the results less obvious. To reduce computational costs, we use sub-sampling to compute the measures. In addition, we measure accuracy, demographic disparity ($|P(\hat{y} = 1|z = 0) - P(\hat{y} = 1|z = 1)|$), and equal opportunity difference ($|P(\hat{y} = 1|y = 1, z = 0) - P(\hat{y} = 1|y = 1, z = 1)|$). The datasets are partitioned into 20:80 test and train sets and all results, including model accuracy, are computed on the test set.

5.1 Evaluated learning methods

Several methods have been proposed to train machine learning models that prevent a combination of disparate treatment and impact [16–18]. Such methods, however, can induce a discriminatory bias in model parameters [33]. Other studies propose novel mathematical notions of fairness, such as equalized opportunity, $P(\hat{y} = 1|y = 1, z = 0) = P(\hat{y} = 1|y =$

$1, z = 1$), and equalized odds, $P(\hat{y} = 1|y = b, z = 0) = P(\hat{y} = 1|y = b, z = 1)$, $b \in \{0, 1\}$ [21, 23–25], or parity mistreatment, i.e., $P(\hat{y} \neq y|z = 0) = P(\hat{y} \neq y|z = 1)$ [20]. Recent works expose the *impossibility* of simultaneously satisfying multiple non-discriminatory objectives, such as equalized opportunity and parity mistreatment [57–59]. Thus, there exist multiple supervised learning methods for addressing discrimination, but they are often mutually exclusive. We therefore evaluate four of such methods addressing different non-discriminatory objectives at each of the stages of a machine learning pipeline where discrimination can be addressed: pre-processing, in-processing, and post-processing.

Pre-processing: Reweighing approach from Kamiran and Calders [60]. Before training a given model, this approach modifies the weights of features with the goal of removing discrimination, defined as demographic disparity, by the protected feature.

In-processing: Reductions model from Agarwal et al. [61] yields a randomized classifier with the lowest empirical error subject to a given fairness constraint. We evaluate four variations of reductions constraining on demographic parity, equalized odds, equal opportunity, and error ratio (represented as "DP", "EO", "TPR", and "ER"). (2) Adversarial debiasing from Zhang et al. [62] jointly maximizes accuracy while minimizing a generative adversarial network's ability to determine the protected attribute from the prediction.

Post-processing: Calibrated equalized odds approach from Pleiss et al. [24] that extends Hardt et al. [21]. Building upon the prior work, calibrated equalized odds maintains calibrated probability estimates, i.e., estimates are independent of the protected attribute, while matching an equal cost constraint between the predictions of two groups. In our evaluation the constraint is a weighted combination between the false-negative and false-positive rates between the two groups in the protected attribute.

In all cases, we use the implementations of these algorithms as provided in the AI Fairness 360 (AIF360) open-source library [63]. Each of the models requires access to protected attribute during training time. The post-processing approach, calibrated equalized odds, also needs access to the protected attribute during test time. The baseline "traditional" model is a result of standard supervised learning. Underlying classifier for all the evaluated models is logistic regression. We also evaluate a logistic regression model that drops the protected attribute, Z, before training. In the figures these models are abbreviated as "Trad": standard supervised learning, "Adv Deb": adversarial debiasing, "Exp Grad": reductions model, and "Cal Eq Odds": calibrated equalized odds.

5.2 Synthetic datasets

To generate the synthetic dataset we draw samples from a multivariate normal distribution with standard normal marginals and given correlations. We then convert a column of our matrix into binary values, set that as Z, and set the rest as X. We compare the learning methods while increasing the correlation $r(X_1, Z)$ from 0 to 1. We first use a simple scenario (A), $Y = \sigma(X_1 + X_2 + Z + 1)$, where σ is the logistic function and the correlations between both (X_1, X_2) and (X_2, Z) are zero. Then, we have the scenario (B) of $Y = \sigma(0 * X_1 + X_2 + 0 * Z + 1)$ where the correlation between (X_2, Z) is 0.5.

5.2.1 Comparison of introduced methods. As both the MIM and OPT methods minimize loss functions based on the preservation of the influence of non-protected attributes, the models perform comparably. Across all metrics for the synthetic scenarios, both OPT methods perform similarly while the MIM has minute deviations that are often within the 95% confidence interval of at least one of the OPT methods (Figure 3). The MIM achieves $\mathbb{E}|\text{SHAP}_Y(X_2|X, Z)|$ that is larger than that of standard model, likely because this measure is slightly different than the measure in the loss that MIM minimizes (Equation 6). Given this similarity and for readability, in the remaining sections we compare only the
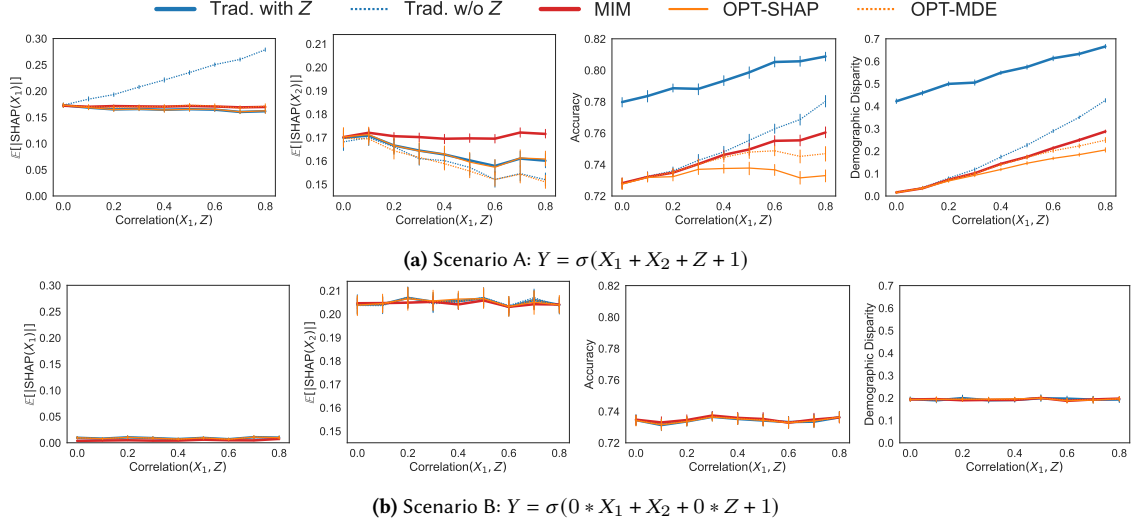
(a) Scenario A: $Y = \sigma(X_1 + X_2 + Z + 1)$

(b) Scenario B: $Y = \sigma(0 * X_1 + X_2 + 0 * Z + 1)$

Fig. 3. Average absolute input influence, measured via MDE or SHAP, model accuracy, and demographic disparity, versus the Pearson correlation between X_1 and Z, under the two scenarios.

MIM with the traditional and state-of-the-art methods. Results for the OPT methods are qualitatively similar (Appendix B).

5.2.2   Comparison vs. state-of-the-art methods. The MIM preserve X_1's influence with respect to the full standard model as $r(X_1, Z)$ increases (red and solid blue lines in Figure 4 & 5). As expected in scenario A, the influence of X_1 increases with correlation for the traditional method that simply drops Z, i.e., it induces indirect discrimination via X_1 (dotted blue line in Figure 4). Interestingly, even though the MIM does not optimize for either fairness measure, it performs better for both fairness measures than the traditional method dropping Z in scenario A.

Other methods addressing discrimination either change the influence of X_1 with the growing correlation $r(X_1, Z)$ ("Exp Grad" in Figure 4 & 5) or use the protected attribute Z and thus discriminate directly (see "Cal Eq Odds", "Reweighing" in Figure 4 & 5). For instance, the method optimizing for parity of impact decreases the impact of X_1 in scenario A, because it aims to remove the correlation between $\hat{Y}$ and Z (brown line in Figure 4). In scenario B, the changes in the influence of X_1 with the growing correlation are especially noteworthy. The affected methods here are increasingly impacted by X_1 as it gets more associated with the protected attribute even though it had no impact in the data generating process. Results for MDE are qualitatively the same as for SHAP (Appendix C) and results for the reductions approach constrained on equal opportunity and error ratio can be found in Appendix B.

5.3   Real-world datasets

We train and test (80:20 random split) the evaluated methods on the COMPAS criminal recidivism dataset [64], German Credit, and Adult Census Income [65] datasets popular in machine learning fairness research.

- **COMPAS.** Here the model predicts the recidivism of an individual based on their demographics and criminal history with race being the protected attribute. We use the binary outcomes as provided by Bellamy et al. [63]. To make the presentation more clear, we exacerbate the racial bias by removing 500 samples of positive outcomes
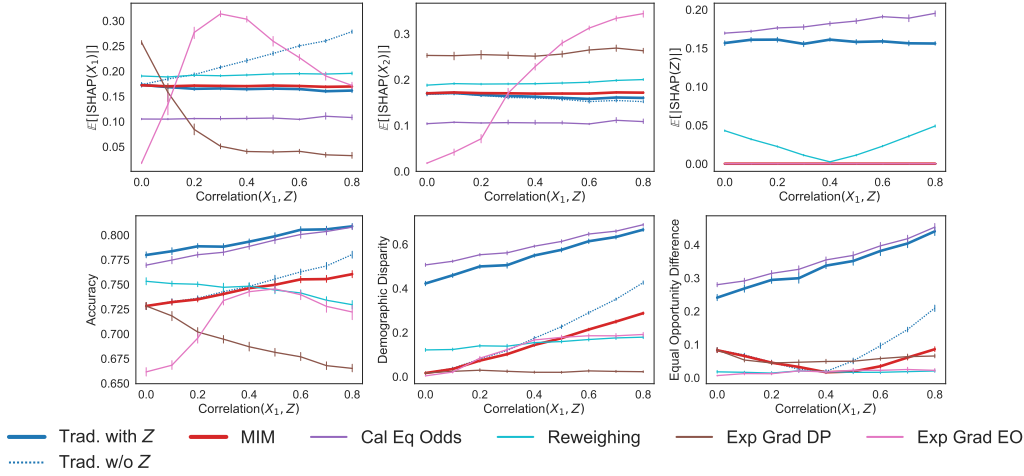
Fig. 4. Scenario A: SHAP influence of X_1, model accuracy, and two popular fairness measures as we increase the correlation $r(X_1, Z)$ for $Y = \sigma(X_1 + X_2 + Z + 1)$. Error bars show 95% confidence intervals based on 30 samples.
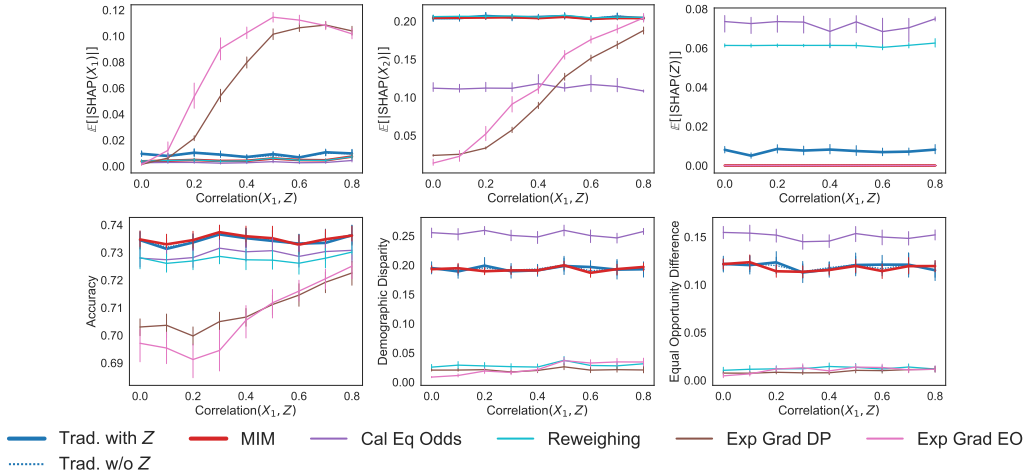

Fig. 5. Scenario B: SHAP influence of X_1, model accuracy, and two popular fairness measures as we increase the correlation $r(X_1, Z)$ for $Y = \sigma(0 * X_1 + X_2 + 0 * Z + 1)$. Error bars show 95% confidence intervals based on 30 samples.

(no recidivism) for African-Americans. The two attributes most correlated with race are age and number of prior counts.

- **German Credit.** A financial dataset with the task being to determine if a loan applicant's credit risk is "good" or "bad" using sex as the protected attribute. We drop non-numeric attributes leaving information about the loan applicant's job, household, and the sought loan. The two attributes most correlated with a applicant's sex are their age and number of dependents.
- **Adult Census Income.** The task for this dataset is to determine if someone's annual income is more than $50k with sex being the protected attribute. Other attributes give information about a person's education, job,

relationships, and demographics. The two attributes most correlated with a person's sex are if they are a husband and if they have a spouse in the armed forces. Note that due to the number of features of this dataset and its effect on computation time for input influence, we omit the results of the OPT methods.

Data loading and pre-processing functions from the AIF360 library are used for these real-world datasets [63]. We train and test all the evaluated models over 30 trials for the COMPAS and German Credit datasets and 10 trials for the Adult Census Income dataset.

In line with the synthetic results, the MIM (and OPT methods) is not influenced by the protected attribute (leftmost column in Figure 6) and, with respect to the traditional model, preserves the influence for the two attributes most correlated with the protected attribute in these real-world scenarios (blue and red bars in the two middle columns of Figure 6). While most of the evaluated models outperform the MIM in terms of demographic disparity (the rightmost column in Figure 6), they are either influenced by the protected attribute (the leftmost column in Figure 6) or do not preserve the influence of at least one of the most correlated attributes (the two middle columns in Figure 6) and have significantly lower accuracy (Figure 7), e.g., "Exp Grad" for COMPAS (Figures 6a & 7a). As with the synthetic results, the changes in influence for the features most correlated with the protected attribute indicate that these methods induce indirect discrimination during training, despite having better performance for certain fairness measures. Results for additional fairness metrics can be found in Appendix B.

6 LIMITATIONS AND FUTURE WORK

This manuscript focuses on two influence measures, MDE and SHAP, and corresponding loss functions for influence preservation. Prior studies show that input influence measures like SHAP can be fooled into stating that a protected attribute has no influence on a model [66]. With this, someone may be able to trick our approach into believing a model was fair by our definition, even though in reality it was not. In such adversarial scenarios, our approach may experience the limitations of other discrimination preventing methods where satisfying a specified fairness objective still leads to discrimination. There exist many other influence measures than the two studied here, and other loss functions could be constructed based on these and other influence measures. We hope to explore these research directions in future works.

While our theoretical guarantees for the preservation of MDE or SHAP hold for wide classes of models, our experiments compare simple logistic models. It would be interesting to test the proposed methods on more complex non-linear models in various important real-world application scenarios across domains. Given that the number of fairness objectives is already high and that we propose new fairness objectives, there is a need for evaluating learning algorithms addressing fairness. A potential approach could rely on realistic simulations of discrimination and test whether a given learning method is able to retrieve the non-discriminatory data-generating process. We develop this approach in a parallel submission (citation hidden to preserve double-blind review).

Most importantly, any fairness objective can be misused by people to justify their systems as fair, since we have only a limited understanding of causal processes happening in real-world, particularly in human decision-making. In such contexts, we emphasize that understanding the causal processes relevant to the decision-making at hand in collaboration with domain experts is of the highest priority, since it can lead to more accurate and more fair models.

7 CONCLUSIONS

The presented results shed a new light on the problem of discrimination prevention in supervised learning. First, we propose a formal definition of induced discrimination, inspired by discrimination via association [12]. We measure influence of features to capture induced discrimination. Second, we show that state-of-the-art methods addressing
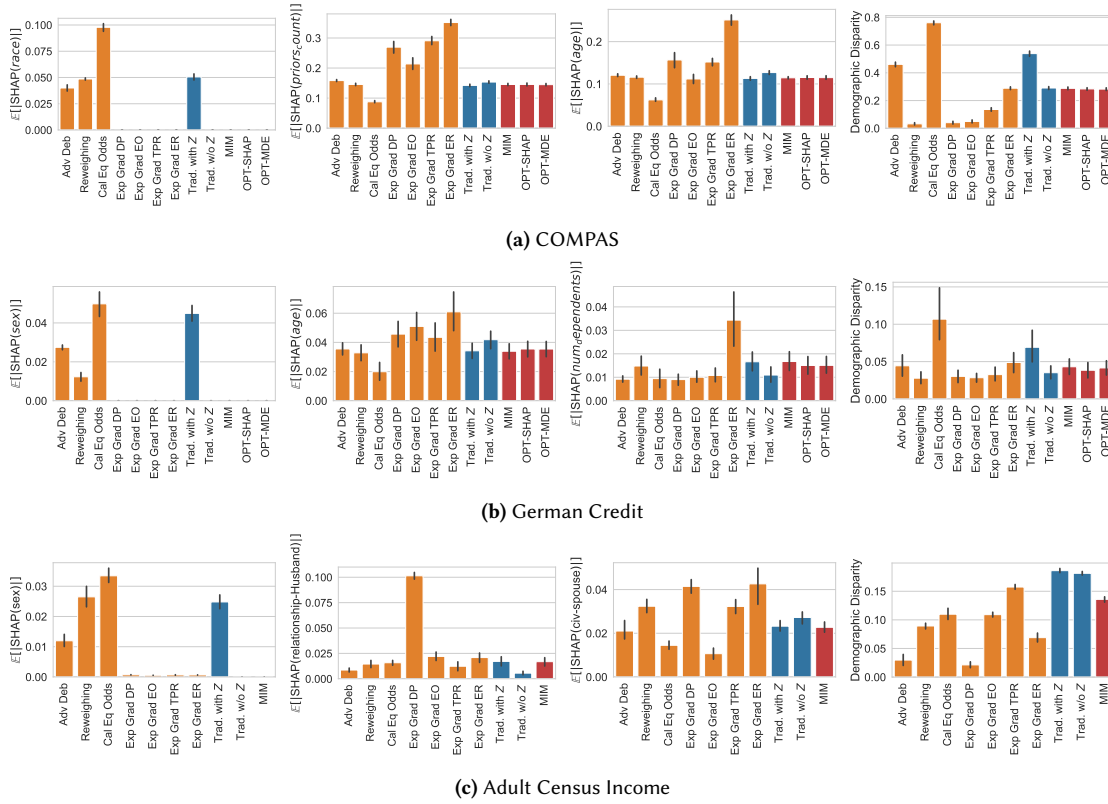
(a) COMPAS

(b) German Credit

(c) Adult Census Income

Fig. 6. Averaged absolute SHAP for the protected attribute and two features most correlated with it and demographic disparity on the COMPAS, German Credit, and Adult Census Income datasets. Error bars show 95% confidence intervals.

(a) COMPAS   **(b)** German Credit   **(c)** Adult Census Income

Fig. 7. Accuracy for the evaluated models on the COMPAS, German Credit, and Adult Census Income datasets. Error bars show 95% confidence intervals.

discrimination often return biased models when they are trained on datasets that are or are *not* affected by discrimination. Third, for discrimination prevention we propose to use a marginal interventional mixture of full models, which prevents the induction of discrimination via association. In the scenarios where discrimination does not affect the training data, the proposed learning algorithm falls back to a traditional learning, which ensures that the method does not bias the

model needlessly. These results provide support for the use of the marginal interventional mixture in the circumstances where discrimination could have affected the training dataset.

REFERENCES

[1] Title VII of the Civil Rights Act, 1964. 7, 42 U.S.C., 2000e et seq.

[2] The Fair Housing Act, 1968. 42 U.S.C.A., 3601-3631.

[3] Kwame Ture, Charles V Hamilton, and Stokely Carmichael. *Black power: The politics of liberation in America: With new afterwords by the authors.* Vintage Books, 1968.

[4] Andrew Altman. Discrimination. In Edward N Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2016 edition, 2016.

[5] Kasper Lippert-Rasmussen. The Badness of Discrimination. 9(2):167–185, 2012. doi: 10.1007/sl0677-006-9014-x.

[6] Yves Zenou and Nicolas Boccard. Racial discrimination and redlining in cities. *Journal of Urban economics*, 48(2):260–285, 2000.

[7] Jesus Hernandez. Redlining revisited: mortgage lending patterns in Sacramento 1930–2004. *International Journal of Urban and Regional Research*, 33 (2):291–313, 2009.

[8] Margery Austin Turner and Felicity Skidmore. Mortgage Lending Discrimination : A Review of Existing Evidence Lending Discrimination : A Review of existing Evidence. In *The Urban Institute*, pages 1–176. 1999.

[9] Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace. Consumer-Lending Discrimination in the FinTech Era. Technical report, National Bureau of Economic Research, Cambridge, MA, jun 2019. URL http://www.nber.org/papers/w25943.pdf.

[10] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019.

[11] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna M. Wallach. Assessing the fairness of AI systems: AI practitioners' processes, challenges, and needs for support. *CoRR*, abs/2112.05675, 2021. URL https://arxiv.org/abs/2112.05675.

[12] Sandra Wachter. Affinity Profiling and Discrimination by Association in Online Behavioural Advertising. *SSRN Electronic Journal*, pages 1–74, 2019. ISSN 1556-5068. doi: 10.2139/ssrn.3388639.

[13] Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. Learning Optimal Fair Policies. In *Proceedings of the 36th International Conference on Machine Learning*, pages PMLR 97:4674–4682, sep 2019. URL http://arxiv.org/abs/1809.02244.

[14] Silvia Chiappa. Path-Specific Counterfactual Fairness. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:7801–7808, jul 2019. ISSN 2374-3468. doi: 10.1609/aaai.v33i01.33017801. URL https://aaai.org/ojs/index.php/AAAI/article/view/4777.

[15] Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. PC-Fairness: A unified framework for measuring causality-based fairness. *Advances in Neural Information Processing Systems*, 32(NeurIPS), 2019. ISSN 10495258.

[16] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, page 560, New York, New York, USA, 2008. ACM Press. ISBN 9781605581934. doi: 10.1145/1401890.1401959. URL http://dl.acm.org/citation.cfm?doid=1401890.1401959.

[17] Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. pages 259–268, 2014. doi: 10.1145/2783258.2783311. URL http://arxiv.org/abs/1412.3756.

[18] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness Constraints: Mechanisms for Fair Classification. *Fairness, Accountability, and Transparency in Machine Learning*, jul 2015. URL http://arxiv.org/abs/1507.05259.

[19] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web - WWW '17*, pages 1171–1180, New York, New York, USA, 2017. ACM Press. ISBN 9781450349130. doi: 10.1145/3038912.3052660. URL http://arxiv.org/abs/1610.08452{%}0Ahttp://dx.doi.org/10.1145/3038912.3052660http://dl.acm.org/citation.cfm?doid=3038912.3052660.

[20] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness Constraints: Mechanisms for Fair Classification. *Artificial Intelligence and Statistics*, 54, 2017. URL https://arxiv.org/abs/1507.05259.

[21] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of Opportunity in Supervised Learning. In D D Lee, M Sugiyama, U V Luxburg, I Guyon, and R Garnett, editors, *Advances in Neural Information Processing Systems*, pages 3315–3323. Curran Associates, Inc., oct 2016. ISBN 9781509037612. doi: 10.1109/ICCV.2015.169. URL http://arxiv.org/abs/1610.02413http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf.

[22] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi, and Adrian Weller. From Parity to Preference-based Notions of Fairness in Classification. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 229–239. Curran Associates, Inc., 2017. URL http://arxiv.org/abs/1707.00010http://papers.nips.cc/paper/6627-from-parity-to-preference-based-notions-of-fairness-in-classification.pdf.

[23] Blake Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning Non-Discriminatory Predictors. (1), 2017. URL http://arxiv.org/abs/1702.06081.

[24] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. On Fairness and Calibration. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5680–5689. Curran Associates, Inc., 2017. URL http://arxiv.org/abs/1709.02012http://papers.nips.cc/paper/7151-on-fairness-and-calibration.pdf.

[25] Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. *Advances in Neural Information Processing Systems*, 2018-Decem(NeurIPS):2791–2801, 2018. ISSN 10495258.

[26] Anil Aswani and Matt Olfat. Optimization Hierarchy for Fair Statistical Decision Problems. 2019. URL http://arxiv.org/abs/1910.08520.

[27] Luca Oneto, Michele Donini, and Massimiliano Pontil. General Fair Empirical Risk Minimization. pages 1–8, 2020. doi: 10.1109/ijcnn48605.2020. 9206819.

[28] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding Discrimination through Causal Reasoning. In *Advances in Neural Information Processing Systems 30*, pages 656–666. Curran Associates, Inc., jun 2017. URL http://arxiv.org/abs/1706.02744http://papers.nips.cc/paper/6668-avoiding-discrimination-through-causal-reasoning.pdf.

[29] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual Fairness. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4066–4076. Curran Associates, Inc., 2017. URL http://arxiv.org/abs/1703.06856http://papers.nips.cc/paper/6995-counterfactual-fairness.pdf.

[30] Junzhe Zhang and Elias Bareinboim. Fairness in Decision-Making – The Causal Explanation Formula. *AAAI*, pages 2037–2045, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16949.

[31] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Capuchin: Causal Database Repair for Algorithmic Fairness. feb 2019. URL http://arxiv.org/abs/1902.08283.

[32] Zachary C. Lipton and Jacob Steinhardt. Troubling trends in machine-learning scholarship. *Queue*, 17(1):1–15, 2019. ISSN 15427749. doi: 10.1145/3317287.3328534.

[33] Zachary C. Lipton, Alexandra Chouldechova, and Julian McAuley. Does mitigating ML's impact disparity require treatment disparity? *Advances in Neural Information Processing Systems*, 2018-Decem(ML):8125–8135, 2018. ISSN 10495258.

[34] Lu Zhang, Yongkai Wu, and Xintao Wu. A causal framework for discovering and removing direct and indirect discrimination. *IJCAI International Joint Conference on Artificial Intelligence*, 0:3929–3935, 2017. ISSN 10450823. doi: 10.24963/ijcai.2017/549.

[35] Ricci v. DeStefano 557 U.S. 557, Docket No. 07-1428, 2009. Supreme Court of the United States.

[36] P. J. Bickel, E. A. Hammel, and J. W. O'Connell. Sex Bias in Graduate Admissions: Data from Berkeley. *Science*, 187(4175):398–404, feb 1975. ISSN 0036-8075. doi: 10.1126/science.187.4175.398. URL https://www.science.org/doi/10.1126/science.187.4175.398.

[37] Nina Grgic-Hlaca, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. Human Perceptions of Fairness in Algorithmic Decision Making. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, pages 903–912, New York, New York, USA, feb 2018. ACM Press. ISBN 9781450356398. doi: 10.1145/3178876.3186138. URL http://arxiv.org/abs/1802.09548http://dl.acm.org/citation.cfm?doid=3178876.3186138.

[38] Glenn Shafer. Causality and Responsibility. In *Cardozo Law Review*, number 22, pages 101–123. ACM Press, New York, New York, USA, 2001. ISBN 9781450322638. doi: 10.1007/978-3-7908-1792-8_23. URL http://dl.acm.org/citation.cfm?doid=2505515.2505754http://link.springer.com/10.1007/978-3-7908-1792-8{_}23.

[39] Hana Chockler and Joseph Y. Halpern. Responsibility and blame: A structural-model approach. *IJCAI International Joint Conference on Artificial Intelligence*, 22:147–153, 2003. ISSN 10450823. doi: 10.1613/jair.1391.

[40] Malcolm Forster and Elliott Sober. How to Tell when Simpler, More Unified, or Less Ad Hoc Theories will Provide More Accurate Predictions. *British Journal of Philosophy of Science*, 45(January):1–35, 1994. ISSN 0007-0882. doi: 10.1093/bjps/45.1.1.

[41] Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics*, 3(1), 2016. ISSN 2196-0089. doi: 10.1186/s40535-016-0018-x.

[42] Anthony C. Constantinou, Yang Liu, Kiattikun Chobtham, Zhigao Guo, and Neville K. Kitson. Large-scale empirical validation of bayesian network structure learning algorithms with noisy data, 2020.

[43] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A Benchmark of in-the-Wild Distribution Shifts. pages 1–87, 2020. ISSN 23318422. URL http://arxiv.org/abs/2012.07421.

[44] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An Investigation of Why Overparameterization Exacerbates Spurious Correlations. In *ICML'20*, may 2020. URL http://arxiv.org/abs/2005.04345.

[45] Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable AI: A causal problem. (2015), oct 2019. URL http://arxiv.org/abs/1910.13413.

[46] Judea Pearl. *Causality: Models, Reasoning and Inference.* Cambridge University Press, 2nd edition, 2009. ISBN 0521773628.

[47] Ilya Shpitser. Segregated graphs and marginals of chain graph models. *Advances in Neural Information Processing Systems*, 2015-Janua:1720–1728, 2015. ISSN 10495258.

[48] A Philip Dawid. Beware of the DAG! *JMLR: Workshop and Conference Proceedings*, 6:59–86, 2008.

[49] Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. *Proceedings - 2016 IEEE Symposium on Security and Privacy, SP 2016*, pages 598–617, 2016. doi: 10.1109/SP.2016.42.

[50] Scott M. Lundberg and Su In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 2017-Decem(Section 2):4766–4775, 2017. ISSN 10495258.

[51] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: A corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010. ISSN 13674803. doi: 10.1093/bioinformatics/btq134.

[52] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-Augu:1135–1144, 2016. doi: 10.1145/2939672.2939778.

[53] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *34th International Conference on Machine Learning, ICML 2017*, 7:5109–5118, 2017.

[54] Charles T. Marx, Richard Lanas Phillips, Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. Disentangling influence: Using disentangled representations to audit model predictions. *Advances in Neural Information Processing Systems*, 32, 2019. ISSN 10495258.

[55] Ian Covert, Scott Lundberg, and Su-In Lee. Understanding Global Feature Contributions With Additive Importance Measures. In *NeurIPS*, apr 2020. URL http://arxiv.org/abs/2004.00668.

[56] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[57] Alexandra Chouldechova. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2):153–163, jun 2017. ISSN 2167-6461. doi: 10.1089/big.2016.0047. URL http://arxiv.org/abs/1703.00056http://online.liebertpub.com/doi/10.1089/big.2016.0047.

[58] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *Proceedings of Innovations in Theoretical Computer Science (ITCS)*, 2017. doi: 10.1111/j.1740-9713.2017.01012.x. URL http://arxiv.org/abs/1609.05807.

[59] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im)possibility of fairness. 2016. URL http://arxiv.org/abs/1609.07236.

[60] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33 (1):1–33, Oct 2012. ISSN 0219-3116. doi: 10.1007/s10115-011-0463-8. URL https://doi.org/10.1007/s10115-011-0463-8.

[61] Alekh Agarwal, Aliiia Beygelzimer, Miroslav Dudfk, John Langford, and Wallach Hanna. A reductions approach to fair classification. *35th International Conference on Machine Learning, ICML 2018*, 1:102–119, 2018.

[62] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 335–340, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278779. URL https://doi.org/10.1145/3278721.3278779.

[63] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. October 2018. URL https://arxiv.org/abs/1810.01943.

[64] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How We Analyzed the COMPAS Recidivism Algorithm. *Pro Publica*, 2016. URL https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[65] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

[66] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2020.

APPENDIX A: PROOF OF PROPOSITION 2

Without loss of generality, we consider the case of two variables x and z. From the definition of $L_{\text{SHAP}}(X)$ and $\text{SHAP}_Y(x|xz)$, under ℓ_2 loss:

$$L_{\text{SHAP}}(X) = \mathbb{E}_X(\mathbb{E}_{Z''}\text{SHAP}_Y(X|XZ'') - \mathbb{E}_{Z''}\text{SHAP}_{\hat{Y}}(X|XZ''))^2 =$$

$$= \mathbb{E}_X\left(\mathbb{E}_{X',Z',Z''}\left[\frac{Y_{X,Z''}-Y_{X',Z''}+Y_{X,Z'}-Y_{X',Z'}}{2} - \frac{\hat{Y}_{X,Z''}-\hat{Y}_{X',Z''}+\hat{Y}_{X,Z'}-\hat{Y}_{X',Z'}}{2}\right]\right)^2$$

For an interventional mixture $\hat{y}_\pi(x) = \mathbb{E}_{\tilde{Z}}\hat{y}(x,\tilde{Z})$,

$$L_{\text{SHAP}}(X) = \mathbb{E}_X\left(\mathbb{E}_{X',Z',Z''}\left[\frac{Y_{X,Z''}-Y_{X',Z''}+Y_{X,Z'}-Y_{X',Z'}}{2}\right] - \mathbb{E}_{X',\tilde{Z}}\left[\hat{Y}_{X,\tilde{Z}}-\hat{Y}_{X',\tilde{Z}}\right]\right)^2.$$

Assuming that the function $y(x,z)$ is analytic, we can expand $y(x,z)$ into a Taylor series around the point $(x=0,z=0)$, which is a series of components Cx^kz^l, where C is a real-valued constant and k and l are integers from 0 to ∞. Let us consider a related series, $y_n(x,z) = \sum_{m=1}^n \alpha_m x^{i(m)}z^{j(m)}$, that can represent a subset of n components of a Taylor series, where α_i are some real-valued constants and $i(m)$ and $j(m)$ are functions returning unique pairs of non-negative integers ordered by m such that $\forall_{m_1\neq m_2}(i(m_1),j(m_1))\neq(i(m_2),j(m_2))$.

Our proof strategy is to first show that the MIM is an interventional mixture that minimizes $L_{\text{SHAP}}(X)$ for the case of $n = 1$, i.e., $y_1(x, z)$. Then, we prove by induction that the MIM is an interventional mixture that minimizes $L_{\text{SHAP}}(X)$ for any n. Since $y_\infty(x, z)$ includes the full Taylor series of $y(x, z)$, so this step ends the proof.

First, we show that for $y_1(x, z) = \alpha_1 x^k z^l$, where $k = i(1)$ and $l = j(1)$, the optimal mixing distribution $\pi^*(\tilde{Z})$ is the marginal distribution, i.e., $\pi^*(\tilde{Z} = z) = P(Z = z)$. Note that the expectation in the interventional mixture can be written as $\mathbb{E}_{\tilde{Z}} x^k \tilde{Z}^l = \beta x^k$, where $\beta = \alpha_1 \mathbb{E}_{\tilde{Z}} \tilde{Z}^l$. Then,

$$L_{\text{SHAP}}(X) = \mathbb{E}_{X} \left(\mathbb{E}_{X', Z', Z''} \left[\frac{X^k Z''^l - X'^k Z''^l + X^k Z'^l - X'^k Z'^l}{2\alpha_1^{-1}} - \beta X^k + \beta X'^k \right] \right)^2 \tag{18}$$

and the minimization of this objective reduces to finding the optimal scalar β. The necessary condition for the minimum of $L_{\text{SHAP}}(X)$ is that its first derivative is zero. Since X', Z', and Z'' have the same means as X and Z, respectively, so the first derivative can be simplified as follows,

$$\frac{\partial L_{\text{SHAP}}(X)}{\partial \beta} =$$
$$= - \mathbb{E}_{X, X', Z', Z''} \left[\left(\alpha_1 (X^k Z''^l - X'^k Z''^l + X^k Z'^l - X'^k Z'^l) - 2\beta \left(X^k - X'^k \right) \right) (X^k - X'^k) \right] =$$
$$= - \mathbb{E}_{X} \left[\left(\alpha_1 (X^k \overline{Z^l} - \overline{X^k} \overline{Z^l} + X^k \overline{Z^l} - \overline{X^k Z^l}) - 2\beta \left(X^k - \overline{X^k} \right) \right) \left(X^k - \overline{X^k} \right) \right],$$

where $\overline{X}$ is the mean of X. After performing a few basic algebraic operations (note that $\mathbb{E}_X[\overline{X^k Z^l}(X^k - \overline{X^k})] = \mathbb{E}_X[\overline{X^k Z^l}(X^k - \overline{X^k})] = 0$), the necessary condition reads

$$\frac{\partial L_{\text{SHAP}}(X)}{\partial \beta} = - \mathbb{E}_{X} \left[\left(2\alpha_1 X^k \overline{Z^l} - 2\beta \left(X^k - \overline{X^k} \right) \right) \left(X^k - \overline{X^k} \right) \right] =$$
$$= -2 \left(\alpha_1 \mathbb{V}[X^k] \overline{Z^l} - \beta \mathbb{V}[X^k] \right) = 0,$$

where $\mathbb{V}[X]$ is the variance of X. This condition is fulfilled if $\beta = \alpha_1 \overline{Z^l}$, which requires that $\mathbb{E} \tilde{Z}^l = \overline{Z^l}$. The condition is met if the distributions of Z and $\tilde{Z}$ are the same. The extremum is actually a minimum, since the second derivative, $\mathbb{V}[X^k]$, is positive for any random variable X with non-zero variance and positive k. For the special case of $k = 0$ or $l = 0$ the MIM minimizes the objective globally, by achieving $L_{\text{SHAP}}(X) = 0$.

Next, let us assume that for a certain $y_{n-1}(x, z) = \sum_{m=1}^{n-1} \alpha_m x^{i(m)} z^{j(m)}$, where $n > 1$, the MIM minimizes $L_{\text{SHAP}}(X)$. Given this assumption, we aim to prove that the MIM is the optimal mixture also for $y_n(x, z) = y_{n-1}(x, z) + \alpha_n x^{i(n)} z^{j(n)}$. In this case, the objective can be written as

$$L_{\text{SHAP}}(X) = \mathbb{E}_{X} \left[(A_{n-1}(X) + a_n(X))^2 \right],$$

where

$$a_n(x) = \alpha_n \left(x^{i(n)} \overline{Z^{j(n)}} - \overline{X^{i(n)}} \overline{Z^{j(n)}} + x^{i(n)} \overline{Z^{j(n)}} - \overline{X^{i(n)} Z^{j(n)}} \right) / 2 - \beta_n \left(x^{i(n)} - \overline{X^{j(n)}} \right)$$

and $A_{n-1}(x) = \sum_{m=1}^{n-1} a_m(x)$. To simplify the notation, we write $\boldsymbol{\beta} = (\beta_1, ..., \beta_n)$. Next, we expand the objective,

$$L_{\text{SHAP}}(X) = \mathbb{E}_{X} \left[(A_{n-1}^2(X) + a_n^2(X) + 2A_{n-1}(X)a_n(X)) \right].$$

From the assumption we know that the MIM minimizes $\mathbb{E}_X \left[A_{n-1}^2(X) \right]$. Also, $\mathbb{E}_X \left[a_n^2(X) \right]$ is equivalent to the objective defined in Equation 18, which is minimized by the MIM as well, as we showed above. Hence, we shall focus now on the

remaining term, i.e., $l(\boldsymbol{\beta}) = \mathbb{E}_X\left[(A_{n-1}(X)a_n(X))\right]$. The necessary conditions for an extremum are

$$\frac{\partial l}{\partial \beta_n} = -2\,\mathbb{E}_X\left[A_{n-1}(X)\left(X^{i(n)} - \overline{X^{j(n)}}\right)\right] = 0,$$

$$\forall_{m<n}\frac{\partial l}{\partial \beta_m} = -2\,\mathbb{E}_X\left[a_n(X)\left(X^{i(m)} - \overline{X^{j(m)}}\right)\right] = 0.$$

These conditions are satisfied when $\forall_{m=1}^n \beta_m = \alpha_m\overline{Z^{i(m)}}$, which are met if $\forall_{m=1}^n \mathbb{E}\tilde{Z}^{i(m)} = \overline{Z^{i(m)}}$. Hence, the necessary conditions for extremum are met if the distribution of $\tilde{Z}$ is the same as the marginal distribution of Z. The corresponding Hessian matrix is positive semi-definite, so the extremum is a minimum.

We have shown that the MIM is an interventional mixture that minimizes $L_{\text{SHAP}}(X)$ for $y_1(x, z)$. In addition, it minimizes it also for $y_n(x, z)$, assuming that it does so for $y_{n-1}(x, z)$. Thus, by induction, the MIM is an interventional mixture that minimizes $L_{\text{SHAP}}(X)$ for any n and any function $y(x, z)$ that has a Taylor expansion.

APPENDIX B: ADDITIONAL RESULTS

Synthetic Scenarios. Here, we present the SHAP influence of X_2 and Z, disparate impact ($|P(\hat{y} = 1|z = 0)/P(\hat{y} = 1|z = 1)|$), and a relaxed version of equalized odds ($((|FPR_{z=0} - FPR_{z=1}| + |TPR_{z=0} - TPR_{z=1}|)/2)$ of the scenarios in the main text: A) $Y = \sigma(X_1 + X_2 + Z + 1)$ and B) $Y = \sigma(0 * X_1 + X_2 + 0 * Z + 1)$. We also provide results for the OPT methods, the reductions method from Agarwal et al. [61] subject to equal opportunity and error ratios fairness constraints("TPR", and "ER"). As expected, the evaluated models that use Z by design are influenced by it ("Cal Eq Odds", Reweighing in Figure 8 & 9).

Real-world datasets. In Figures 10, 11, & 12 we provide results for the disparate impact, equalized odds, and equal opportunity fairness metrics and for the OPT methods on the COMPAS, German Credit, and Adult Census Income datasets.
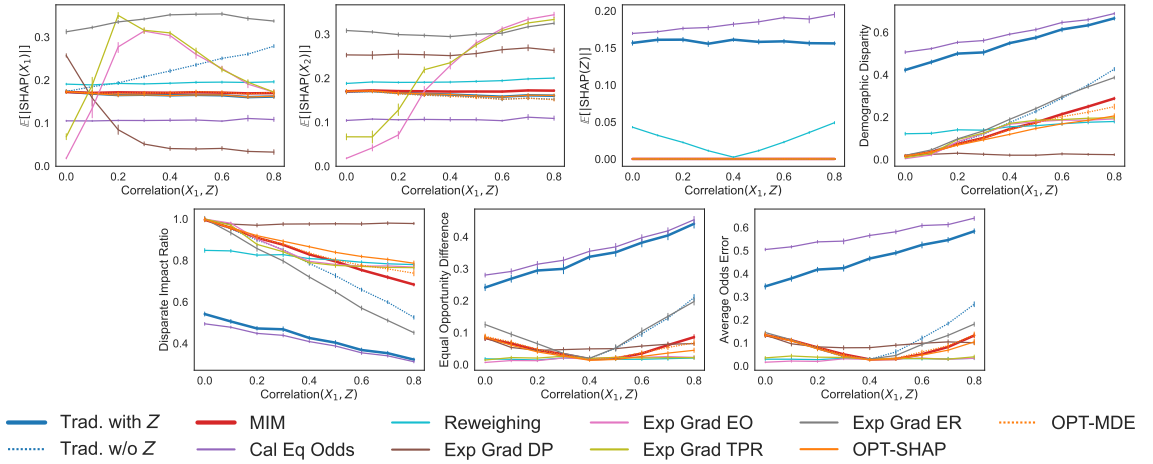

Fig. 8. Scenario A: Averaged absolute SHAP of X_1, X_2, and Z and four fairness measures as we increase the correlation $r(X_1, Z)$. Error bars show 95% confidence intervals.
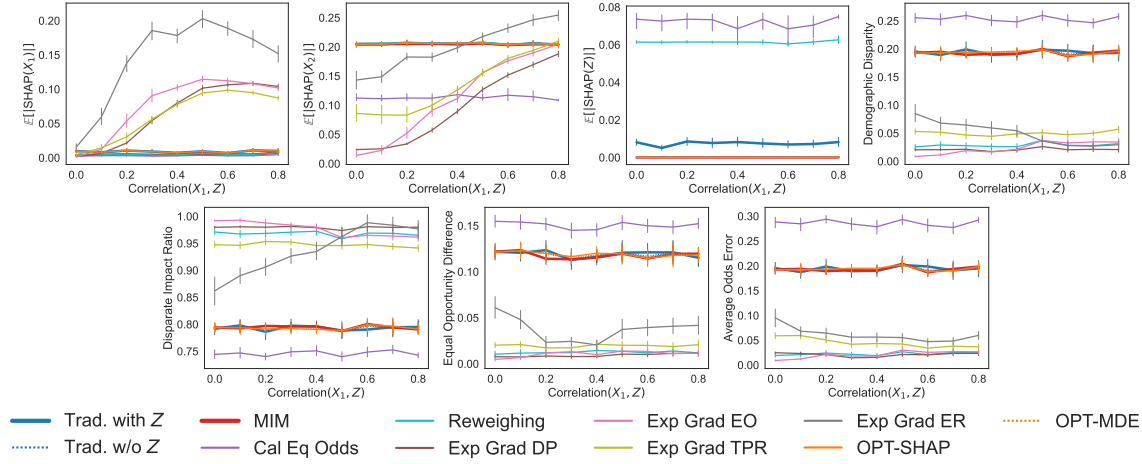
Fig. 9. Scenario B: Averaged absolute SHAP of X_1, X_2, and Z and four fairness measures as we increase the correlation $r(X_1, Z)$. Error bars show 95% confidence intervals.
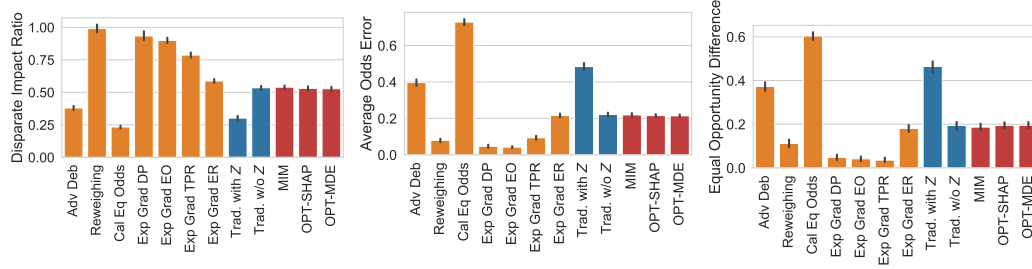

Fig. 10. Disparate impact, equalized odds, and equal opportunity measures for the evaluated models on the COMPAS dataset. Error bars show 95% confidence intervals.
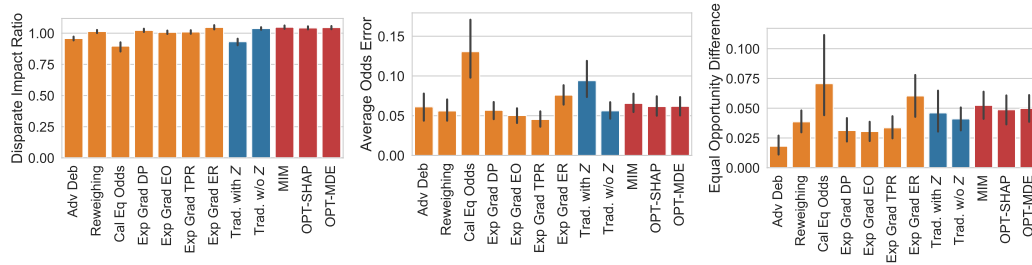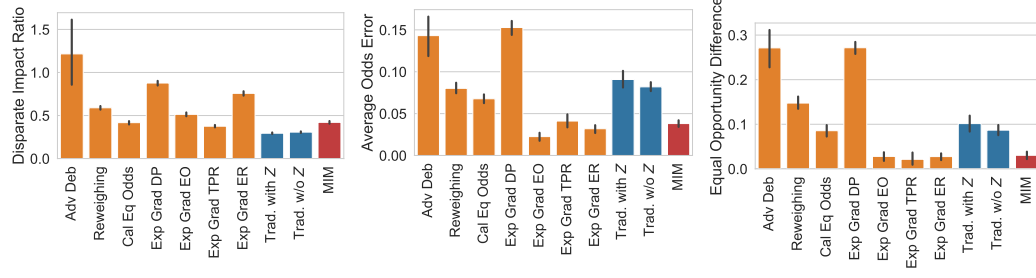

Fig. 11. Disparate impact, equalized odds, and equal opportunity measures for the evaluated models on the German Credits dataset. Error bars show 95% confidence intervals.

APPENDIX C: MARGINAL DIRECT EFFECT (MDE)

In addition to the SHAP input influence measure, we measure the MDE ($\mathbb{E}_{X_i, X_i''} |\text{MDE}_Y(X_i, X_i'')|$) on the same features for all datasets in Figures 13-17. All results for MDE are qualitative equivalent to that of SHAP for each dataset.

Fig. 12. Disparate impact, equalized odds, and equal opportunity measures for the evaluated models on the Adult Census Income dataset. Error bars show 95% confidence intervals.
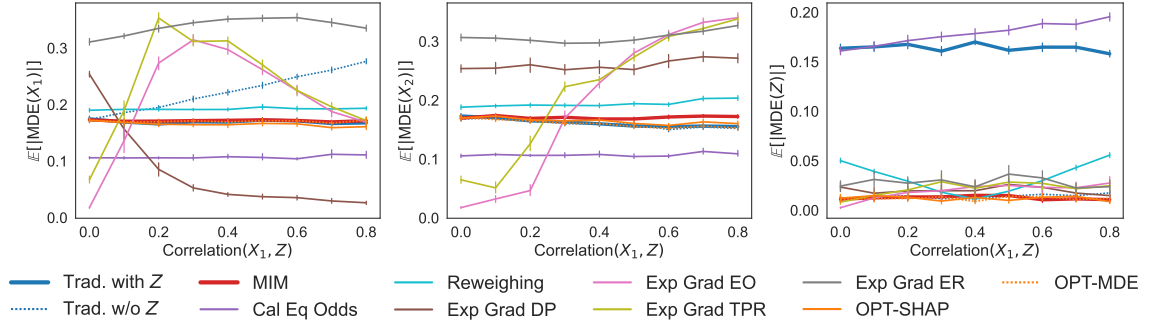

Fig. 13. Scenario A: MDE as we increase the correlation $r(X_1, Z)$. Error bars show 95% confidence intervals.
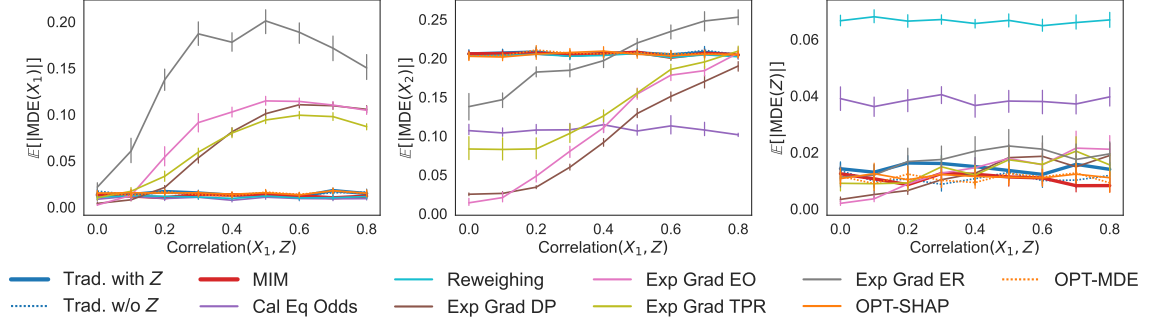

Fig. 14. Scenario B: MDE as we increase the correlation $r(X_1, Z)$. Error bars show 95% confidence intervals.
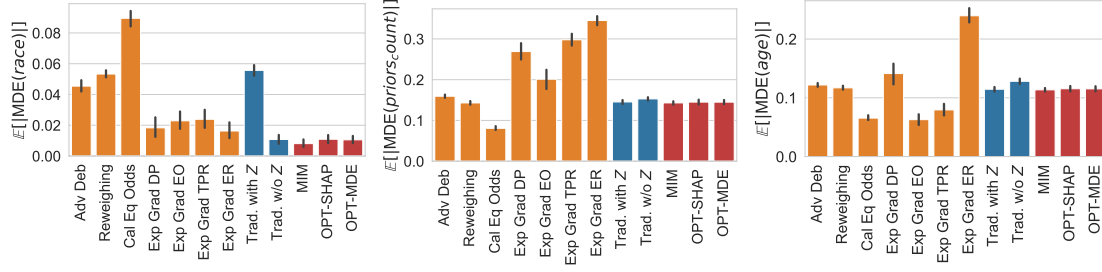
Fig. 15. MDE for the protected attribute and two features most correlated with it for the evaluated models on the COMPAS dataset. Error bars show 95% confidence intervals.
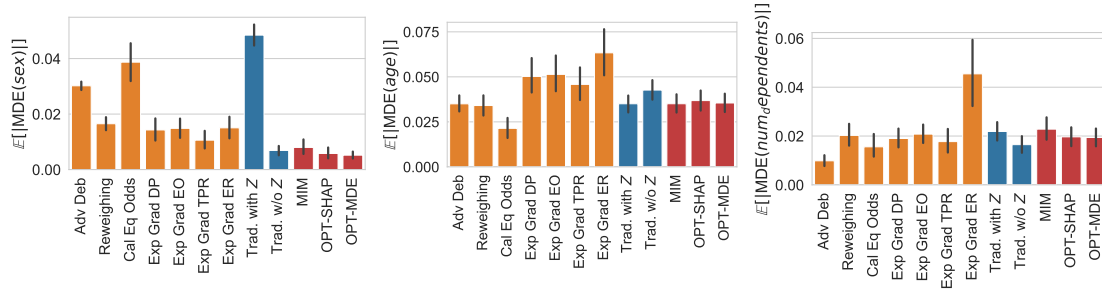

Fig. 16. MDE for the protected attribute and two features most correlated with it for the evaluated models on the German Credit dataset. Error bars show 95% confidence intervals.
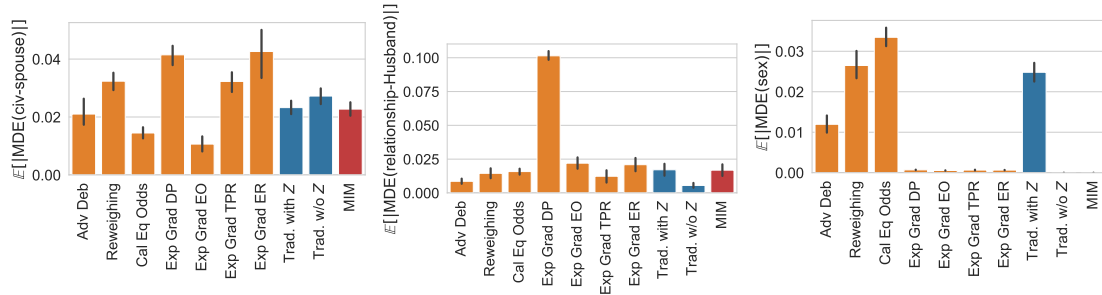

Fig. 17. MDE for the protected attribute and two features most correlated with it for the evaluated models on the Adult Census Income dataset. Error bars show 95% confidence intervals.