Marrying Fairness and Explainability in Supervised Learning

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

ABSTRACT

Machine learning algorithms that aid human decision-making may inadvertently discriminate against certain protected groups. Therefore, 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.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning algorithms**; **Supervised learning**; • **Applied computing** → Law, social and behavioral sciences.

KEYWORDS

machine learning, explainability, algorithmic fairness, discrimination, supervised learning

ACM Reference Format:
Przemyslaw Grabowicz, Nicholas Perello, and Aarshee Mishra. 2022. Marrying Fairness and Explainability in Supervised Learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3531146.3533236

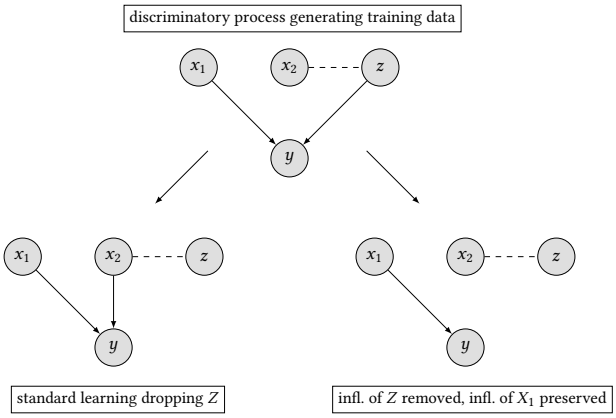


Figure 1: An illustration of the graphical models that result from applying different learning methods to the example scenario: standard learning dropping Z (bottom left), the removal of influence of Z while preserving the influence of X (bottom right). The directed edges correspond to causal relations,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

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 [52, 53] and social science [2, 31, 54] 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, $\boldsymbol{X}$, and should not depend on the protected attribute, Z, i.e., $P(y|\boldsymbol{x}, z) = P(y|\boldsymbol{x}, z')$ 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 [22, 63], 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 [2, 31], that is $P(y|z) = P(y|z')$, unless there is a "justified reason" or "business necessity" for this *disparate impact* [52, 53].

Example. The following example runs through the manuscript. Consider a hypothetical linear model of loan interest rate, Y. Prior

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

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533236

works suggest that interest rates differ by race, Z [5, 55]. Some loan-granting clerks may produce non-discriminatory decisions, $y = \beta_0 - x_1$, while other clerks may discriminate directly, $y_{\text{dir}} = \beta_0 - x_1 - z$ (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 discriminating clerk may induce discrimination in the interest rate, by using a proxy for race, $y_{\text{ind}} = \beta_0 - x_1 - x_2$, 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 $D = \{(x_1, x_2, z, y_{\text{dir}})\}$ 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 May 2022) [49], and business necessity is often neglected, which may be slowing down the widespread adoption of fair machine learning methods [35]. 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 data-rich machine learning 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 [56]. Causality-based research offers so-called *path-specific counterfactual fairness* that enables designation of fair paths for business necessity [8, 37, 58], 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 is 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 to 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, $\boldsymbol{Z}$, while maintaining the influence of the remaining features, $\boldsymbol{X}$, 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 $\boldsymbol{Z}$ while minimizing the inducement of discrimination through $\boldsymbol{X}$. We show that this method keeps influence of $\boldsymbol{X}$ and $\boldsymbol{Z}$ close to the target values and, in addition, decreases popular disparity measures, while keeping high model accuracy. Our methods are released publicly via an easy-to-use `FaX-AI` Python library (https://github.com/social-info-lab/FaX-AI).

2 RELATED WORKS

In machine learning, discrimination is typically defined based on statistical independence [4, 15, 17, 21, 38, 40, 41, 57, 59–62] or causal relations [25, 29, 45, 64]. Well-known fairness objectives, such as parity of impact and equalized odds, correspond or are related to the statistical independence between $\boldsymbol{Z}$ and Y [4]. However, legal systems allow for exceptions from this independence through the business necessity clause, which permits usage of an attribute $\boldsymbol{X}$ associated with $\boldsymbol{Z}$ and results in the decisions Y depending on $\boldsymbol{Z}$ through $\boldsymbol{X}$ 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 [33], which results in shortcomings. For instance, the algorithms that put constraints on the aforementioned disparities in treatment and impact [17, 40, 60] could negatively affect females with short hair and/or programming skills, because of those features' (fair or unfair) association with males [32].

A relevant line of research proposes to define direct and indirect discrimination as direct and indirect causal influence of $\boldsymbol{Z}$ on Y, respectively [64, 65]. 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. For instance, the majority's view in the Supreme Court case of Ricci v. DeStefano [43] argued that the defendants could not argue that the disputed promotion examinations results were inconsistent with business necessity. *Path-specific* notions of causal fairness address this issue to a limited extent [8, 37, 58]. 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 [8, 25, 29, 37, 45, 58, 64], 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 $\boldsymbol{X}$ and protected variables $\boldsymbol{Z}$, where $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{z} \in \mathcal{Z}$, $y \in \mathcal{Y}$, 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 $y = y(\boldsymbol{x}, \boldsymbol{z}, \epsilon)$, where ϵ is an exogenous noise variable. Since the exogenous noise is unpredictable, we focus on the de-noised function $y(\boldsymbol{x}, \boldsymbol{z}) = \mathrm{E}_\epsilon\, y(\boldsymbol{x}, \boldsymbol{z}, \epsilon)$ 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 $\boldsymbol{x}$ and their race $\boldsymbol{z}$, or ii) credit score assignments for a customer, given their financial record $\boldsymbol{x}$ 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., $\mathrm{E}[\ell(Y, \hat{y}(\boldsymbol{X}, \boldsymbol{Z}))]$, where the expectation is over the set of training samples $(\boldsymbol{x}, \boldsymbol{z}, y)$ 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 $\boldsymbol{Z}$ from the model without inducing discrimination, that is, without increasing the impact of attributes $\boldsymbol{X}$ associated with $\boldsymbol{Z}$ in an unjustified and discriminatory way.

We propose that a non-discriminatory model shall remove the influence of the protected features $\boldsymbol{Z}$ on Y, while preserving the influence of the remaining attributes $\boldsymbol{X}$ 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 $\boldsymbol{Z}$ 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) $\boldsymbol{Z}$ 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) [52] 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) [53] determines that hiring decisions shall not be influenced by race, color, religion, sex, and national origin.

In the context of making decisions Y using features $\boldsymbol{X}$, some of the features may be associated with, or affected by, the protected attribute $\boldsymbol{Z}$. 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) $\boldsymbol{Z}$ with non-protected feature(s) $\boldsymbol{X}$ 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 [7]. 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 [20], 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 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 $\boldsymbol{Z}$ 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 $\boldsymbol{Z}$ from the unfairly influenced $\boldsymbol{X}$ 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

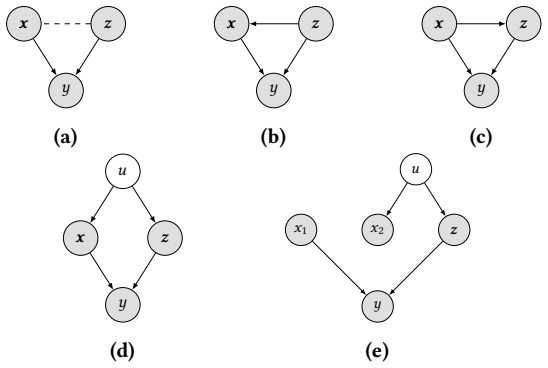


Figure 2: The considered setting. We make no assumptions about the relations between x and z (marked with a dashed edge), nor their components. Hence, the graph (a) includes all exemplary cases (b-e). The graph (e) depicts the data-generating process from Example and shows no relationship between the components x_1 and x_2 of x. The random variable u is an exogenous noise, i.e., an unmeasured independent random variable.

decisions, legal systems consider the epistemic state of decision-makers [9, 46], e.g., whether an employer knew about discrimination in company's hiring process, and their intentions [2], 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 [11, 18, 50], 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 [28, 44], i.e., they constitute an inaccurate epistemic state. Unfortunately, in practice it may be impossible to test causal validity of model $y(\boldsymbol{x}, \boldsymbol{z})$, 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 $\boldsymbol{X}$ and, possibly, $\boldsymbol{Z}$, as detailed in the causal explainability literature [23]. That model can suffer the effects of training on discriminatory data. In the remainder 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) [39]. Other frameworks, such as segregated graphs [47] and extended conditional independence [14] 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 $\boldsymbol{X}$ and $\boldsymbol{Z}$ 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) [11, 50]. 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 (w.r.t. the causal graph) influences the model of Y. We make no assumptions about the relations between $\boldsymbol{X}$ and $\boldsymbol{Z}$ 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 $do(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z})$ is written as $Y_{\boldsymbol{x},\boldsymbol{z}}$, which is the outcome we would have observed had the variables $\boldsymbol{X}$ and $\boldsymbol{Z}$ been set to the values $\boldsymbol{x}$ and $\boldsymbol{z}$ via an intervention. The causal *controlled direct effect* on Y of changing the value of $\boldsymbol{Z}$ from a reference value $\boldsymbol{z}$ to $\boldsymbol{z'}$ given that $\boldsymbol{X}$ is set to $\boldsymbol{x}$ [39] is

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

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) $\boldsymbol{Z}$ on the decisions Y and $\exists_{z,z'}\exists_{\boldsymbol{x}}\mathrm{CDE}_Y(\boldsymbol{z}, \boldsymbol{z'}|\boldsymbol{x}) \neq 0$.

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

To remove direct discrimination, one can construct a model $\hat{Y}$ that does not use $\boldsymbol{Z}$. However, the removal of direct discrimination may induce discrimination via the attributes X_i associated with the protected attributes $\boldsymbol{Z}$, even if there is no causal link from $\boldsymbol{Z}$ 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 $\boldsymbol{Z}$ between the processes Y and $\hat{Y}$ in the sense that $\exists_{\boldsymbol{z}}\exists_{\boldsymbol{x},\boldsymbol{x'}}\mathrm{CDE}_Y(\boldsymbol{x}, \boldsymbol{x'}|\boldsymbol{z}) \neq \mathrm{CDE}_{\hat{Y}}(\boldsymbol{x}, \boldsymbol{x'}|\boldsymbol{z})$ given that $P(\boldsymbol{x}|\boldsymbol{z}) \neq P(\boldsymbol{x})$ or $P(\boldsymbol{x'}|\boldsymbol{z}) \neq P(\boldsymbol{x'})$.

Example. Consider the aforementioned linear models of loan interest rate, $\hat{y}_1$ and $\hat{y}_2$. Note that $CDE_{\hat{y}_1}(\boldsymbol{x}, \boldsymbol{x}'|\boldsymbol{z}) - CDE_{\hat{y}_2}(\boldsymbol{x}, \boldsymbol{x}'|\boldsymbol{z}) = r * (x_2 - x_2')$, since $\boldsymbol{x}$ has two components $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ 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 [39], to measure the extent of induced discrimination, we introduce natural direct and indirect effects using nested counterfactuals, $Y_{\boldsymbol{X}_{\boldsymbol{z}}, \boldsymbol{z}'}$, denoting the outcome that would have been observed if $\boldsymbol{Z}$ were set to $\boldsymbol{z}'$ and $\boldsymbol{X}$ were set to the value it would have taken if $\boldsymbol{Z}$ were set to $\boldsymbol{z}$. *Natural direct effect* of changing the value of $\boldsymbol{Z}$ from a reference value $\boldsymbol{z}$ to $\boldsymbol{z}'$ is

$$\mathrm{NDE}_Y(\boldsymbol{z}', \boldsymbol{z}) = \mathrm{E}[Y_{\boldsymbol{X}_{\boldsymbol{z}}, \boldsymbol{z}'} - Y_{\boldsymbol{X}_{\boldsymbol{z}}, \boldsymbol{z}}]. \tag{2}$$

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* [39] yields

$$\begin{aligned}\mathrm{NDE}_Y(\boldsymbol{z}', \boldsymbol{z}) &= \mathrm{E}[Y_{\boldsymbol{X}_{\boldsymbol{z}}, \boldsymbol{z}'} - Y_{\boldsymbol{X}_{\boldsymbol{z}}, \boldsymbol{z}}] \\ &= \mathrm{E}_{\boldsymbol{X} \sim P(\boldsymbol{X}|\boldsymbol{z})}[Y|\boldsymbol{X}, \boldsymbol{z}'] - \mathrm{E}_{\boldsymbol{X} \sim P(\boldsymbol{X}|\boldsymbol{z})}[Y|\boldsymbol{X}, \boldsymbol{z}].\end{aligned} \tag{3}$$

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

$$\mathrm{NDE}_Y(\boldsymbol{z}', \boldsymbol{z}) = \mathrm{E}_{\boldsymbol{X} \sim P(\boldsymbol{X})}[Y|\boldsymbol{X}, \boldsymbol{z}'] - \mathrm{E}_{\boldsymbol{X} \sim P(\boldsymbol{X})}[Y|\boldsymbol{X}, \boldsymbol{z}], \tag{4}$$

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

$$\begin{aligned}\mathrm{MDE}_Y(\boldsymbol{z}', \boldsymbol{z}) &= \mathrm{E}[Y_{\boldsymbol{X}'', \boldsymbol{z}'} - Y_{\boldsymbol{X}'', \boldsymbol{z}}] = \mathrm{E}_{\boldsymbol{X} \sim P(\boldsymbol{X})} \mathrm{CDE}_Y(\boldsymbol{z}', \boldsymbol{z}|\boldsymbol{X}) \\ &= \mathrm{E}_{\boldsymbol{X} \sim P(\boldsymbol{X})}[Y|\boldsymbol{X}, \boldsymbol{z}'] - \mathrm{E}_{\boldsymbol{X} \sim P(\boldsymbol{X})}[Y|\boldsymbol{X}, \boldsymbol{z}],\end{aligned} \tag{5}$$

which takes an expectation over the probabilistic interventions on $\boldsymbol{X}$, setting its value to random samples of $\boldsymbol{X}''$, where $\boldsymbol{X}''$ is a variable independent from all other variables, but has the same marginal distribution as $\boldsymbol{X}$. 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

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

or its feature-specific version, which computes the loss separately for each component of $\boldsymbol{X}$,

$$\begin{aligned}L^{\mathrm{IND}}_{\mathrm{MDE}}(\boldsymbol{X}) &= \sum_i L_{\mathrm{MDE}}(X_i) \\ &= \sum_i \mathrm{E}_{X_i'', X_i}\, \ell(\mathrm{MDE}_Y(X_i, X_i''), \mathrm{MDE}_{\hat{Y}}(X_i, X_i'')).\end{aligned} \tag{7}$$

A similar loss could be constructed based on the comparison between $\mathrm{CDE}_Y(\boldsymbol{X}, \boldsymbol{X}''|\boldsymbol{Z})$ and $\mathrm{CDE}_{\hat{Y}}(\boldsymbol{X}, \boldsymbol{X}''|\boldsymbol{Z})$. In this paper we focus on losses based on MDE or the 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 $\boldsymbol{X}$ and $\boldsymbol{Z}$ as $\boldsymbol{W} = \boldsymbol{XZ}$, i.e., samples of $\boldsymbol{W}$ are tuples $\boldsymbol{w} = (\boldsymbol{x}, \boldsymbol{z})$ and $\boldsymbol{w} \in \mathcal{X} \times \mathcal{Z} = \mathcal{W}$. Components of $\boldsymbol{W}$ 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 [13, 23, 34]. In particular, let primed variables have the same joint distribution as the non-primed variables, $\forall_{\boldsymbol{w} \in \mathcal{W}} P(\boldsymbol{W}' = \boldsymbol{w}) = P(\boldsymbol{W} = \boldsymbol{w})$, while being independent from them, $\boldsymbol{W}' \perp \boldsymbol{W}$. 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., $\forall_{i \in \mathcal{F}} \forall_{j \neq i} W_i'' \perp W_j''$, $\boldsymbol{W}'' \perp \boldsymbol{W}'$ and $\boldsymbol{W}'' \perp \boldsymbol{W}$.

For any subset of features T that does not contain i, prior works define a marginal influence (MI) using $\boldsymbol{W}'$ as a random baseline [13, 23],

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

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

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 [13, 34],

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

For instance, for the case of two variables,

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

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

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

or its feature-specific version,

$$\begin{aligned}L^{\mathrm{IND}}_{\mathrm{SHAP}}(\boldsymbol{X}) &= \sum_i L_{\mathrm{SHAP}}(X_i) \\ &= \sum_i \mathrm{E}_{\boldsymbol{X}}\, \ell(\mathrm{E}_{\boldsymbol{Z}''} \mathrm{SHAP}_Y(X_i|\boldsymbol{XZ}''), \mathrm{E}_{\boldsymbol{Z}''} \mathrm{SHAP}_{\hat{Y}}(X_i|\boldsymbol{XZ}'')).\end{aligned} \tag{11}$$

While here we have constructed loss functions based on SHAP, other input influence measures, such as PFI or SAGE, can be used as well [3, 12, 36, 42, 51]. 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 $\boldsymbol{Z}$, while preserving the influence of non-protected attributes $\boldsymbol{X}$ by minimizing $L_{\text{MDE}}(\boldsymbol{X})$ or $L_{\text{SHAP}}(\boldsymbol{X})$, which lead to a simple closed-form solution, or their feature-specific versions, i.e., $L^{\text{IND}}_{\text{MDE}}(\boldsymbol{X})$ or $L^{\text{IND}}_{\text{SHAP}}(\boldsymbol{X})$, 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 $L_{\text{MDE}}(\boldsymbol{X})$ or $L_{\text{SHAP}}(\boldsymbol{X})$

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

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

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

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

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

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 $\text{E}[\tilde{Z}^l] = \text{E}[Z^l]$ 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. In the loan interest rate example, the full model is $y(\boldsymbol{x}, z) = \beta_0 - x_1 - z$. The MIM is $\hat{y}_{\text{MIM}} = \beta_0 - x_1 - \text{E}_Z\, Z$.

4.2 Minimizing $L^{\text{IND}}_{\text{MDE}}(\boldsymbol{X})$ and $L^{\text{IND}}_{\text{SHAP}}(\boldsymbol{X})$ via transfer learning

The minimization of the feature-specific losses, $L^{\text{IND}}_{\text{MDE}}(\boldsymbol{X})$ and $L^{\text{IND}}_{\text{SHAP}}(\boldsymbol{X})$, 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 $L^{\text{IND}}_{\text{MDE}}(\boldsymbol{X})$ or $L^{\text{IND}}_{\text{SHAP}}(\boldsymbol{X})$. 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 [26] 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 from 10^{-3} to 10^{-2} and N is from 20 to 100. Our implementations of the methods are released publicly via `FaX-AI` 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 $\boldsymbol{Z}$ could fulfill business necessity, as we pointed in Subsection 3.1.1. However, a feature X_i can be unfairly and illegally influenced by $\boldsymbol{Z}$. If decisions Y used such X_i, then Y would be indirectly discriminatory. We have two options to prevent that: 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 $\boldsymbol{Z}$, then use the corrected $\hat{X}_i$ in the model of Y, and finally drop the impact of $\boldsymbol{Z}$ on $\hat{Y}$, while using either of the proposed methods for removing the impact of $\boldsymbol{Z}$ 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.

Example. In the loan example, the annual salary x_1 of a loan applicant could have been affected by discrimination, e.g., $x_1 = s + z$, where s stands for job-related skills. In such case, a bank shall first debias the salary, either by developing a model of X_1 using available information about S and applying our methods, or by retrieving a debiased $\hat{x}_1$ from another source, e.g., the applicant's employer, who is better positioned (and is obliged by law) to debias the salary. In this case, $\hat{x}_{1,\text{MIM}} = s + \bar{z}$ and $\hat{y}_{\text{MIM}} = \beta_0 - \hat{x}_{1,\text{MIM}} - \bar{z} = \beta_0 - s - 2\bar{z}$, where $\bar{z}$ is the mean of Z, so skills determine interest rate.

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 [8]. We maintain the original notation:

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

$$L = \theta^l + \theta^l_z Z + \theta^l_c C + \theta^l_l M + \epsilon_l, \tag{13}$$

$$Y = \theta^y + \theta^y_z Z + \theta^y_c C + \theta^y_y M + \theta^y_l L + \epsilon_y, \tag{14}$$

where C, M, L are components of $\boldsymbol{X}$, while $\epsilon_c, \epsilon_m, \epsilon_l$ 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 $\hat{y}_{\text{MIM}}$,

$$\hat{m}_{\text{MIM}} = \theta^m + \theta^m_z \bar{z} + \theta^m_c c = m - \theta^m_z (a - \bar{z}), \tag{15}$$

$$\hat{l}_{\text{MIM}} = \theta^l + \theta^l_z z + \theta^l_c c + \theta^l_m \hat{m}_{\text{MIM}}, \tag{16}$$

$$\hat{y}_{\text{MIM}} = \theta^y + \theta^y_z \bar{z} + \theta^y_c c + \theta^y_m \hat{m}_{\text{MIM}} + \theta^y_l \hat{l}_{\text{MIM}}, \tag{17}$$

where $\bar{z}$ stands for the mean of Z. A comparison with PSCF reveals that $\hat{y}_{\text{MIM}} = \hat{y}_{\text{PSCF}} + \Delta$, where $\Delta = \bar{z}(\theta^y_z + \theta^y_m \theta^m_z + \theta^y_l \theta^l_m \theta^m_z)$. 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., $\text{E}(Y - \hat{Y}_{\text{PSCF}})^2 = \text{E}(Y - \hat{Y}_{\text{MIM}})^2 + \Delta^2$. 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,

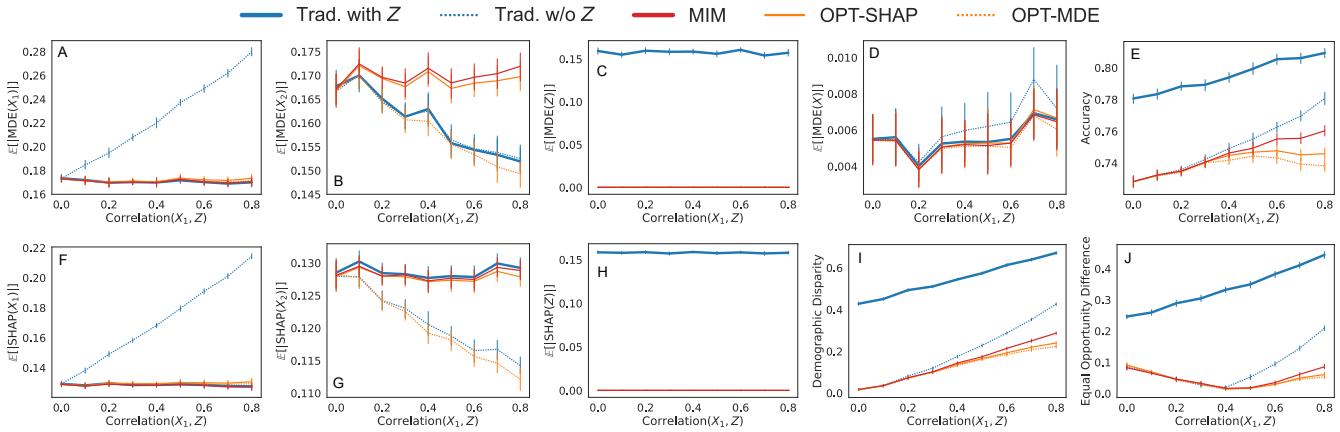


Figure 3: Average absolute input influence, measured via MDE (panels A-D) or SHAP (F-H), model accuracy (E), demographic disparity (I), and accuracy disparity (J) versus the Pearson correlation between X_1 and Z, under Scenario A, where $Y = \sigma(X_1 + X_2 + Z + 1)$.

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 [13, 23]. 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 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 $\mathrm{E}_{\boldsymbol{X},Z}\,|\mathrm{SHAP}_Y(X_i|\boldsymbol{X},Z)|$, following the measure of global feature influence proposed by Lundberg and Lee [34], and $\mathrm{E}_{X_i,X_i'}\,|\mathrm{MDE}_Y(X_i,X_i')|$, both of which are evaluated using outcome probabilities. Note that these measures are different than our loss functions, which make the results non-obvious, yet still intuitive. To reduce computational costs, we use sub-sampling to compute the measures. In addition, we measure accuracy and demographic disparity ($|\,P(\hat{y}=1|z=0) - P(\hat{y}=1|z=1)|$). Results for other measures, such as equalized odds and equal opportunity difference, can be found in Appendix B. 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 [17, 40, 60]. Such methods, however, can induce a discriminatory bias in model parameters [32]. 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, $P(\hat{y}=1|y=b,z=0) = P(\hat{y}=1|y=b,z=1)$, $b \in \{0,1\}$ [15, 21, 41, 57], or parity mistreatment, i.e., $P(\hat{y}\neq y|z=0) = P(\hat{y}\neq y|z=1)$ [61]. Recent works expose the *impossibility* of simultaneously satisfying multiple non-discriminatory objectives, such as equalized opportunity and parity mistreatment [10, 19, 27]. 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 [24]. 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. [1] 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)

Post-processing: Calibrated equalized odds approach from Pleiss et al. [41] 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 [6]. 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

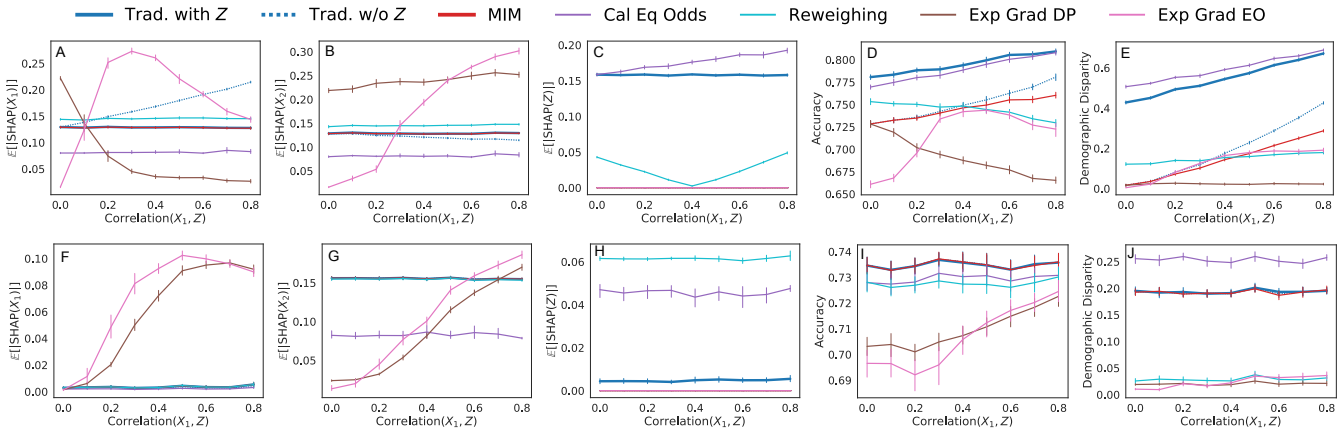


Figure 4: SHAP influence of X_1, X_2, and Z, model accuracy, and demographic disparity as we increase the correlation $r(X_1, Z)$ under the two synthetic scenarios: (top row) Scenario A, $Y = \sigma(X_1 + X_2 + Z + 1)$, and (bottom row) Scenario B, $Y = \sigma(0 * X_1 + X_2 + 0 * Z + 1)$. Error bars show 95% confidence intervals based on 30 samples.

the protected attribute, Z, before training. In the figures these models are abbreviated as "Trad": standard supervised learning, "Exp Grad": reductions model, and "Cal Eq Odds": calibrated equalized odds.

5.2 Synthetic datasets

To generate the datasets 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 $\boldsymbol{X}$. We compare the learning methods while increasing the correlation $r(X_1, Z)$ from 0 to 1. We first introduce and study 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 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 the MIM and the two OPT methods minimize loss functions based on the preservation of the influence of non-protected attributes, the resulting models perform comparably (red and two orange lines in Figure 3). All introduced methods achieve their objectives (compare them against the blue lines in Figure 3), i.e., they all remove the influence of $\boldsymbol{Z}$ (Figures 3C, 3H), the MIM preserves the influence of pooled $\boldsymbol{X}$ (Figure 3D), the OPT-MDE preserves the MDE of individual $\boldsymbol{X}_i$ (Figures 3A, 3B), and the OPT-SHAP preserves the SHAP of individual $\boldsymbol{X}_i$ (Figures 3F, 3G). Interestingly, the MIM performs nearly the same as the OPT-SHAP across all measures, despite not being designed to achieve the feature-specific loss of OPT-SHAP (Equation 11). Since the MIM is guaranteed to preserve the SHAP of the pooled $\boldsymbol{X}$, and SHAP meets the completeness axiom (a.k.a. additivity axiom) [13, 23], which says that the sum of influence of individual features equals to the influence of all features pooled together, hence the MIM can achieve both the pooled and individual objectives, as in this case. Note, however, that the MIM is slightly more accurate than the OPT-SHAP (Figure 3E) at the cost of minimally higher demographic disparity (Figure 3I) and equal opportunity difference, i.e., accuracy disparity (Figure 3J).

5.2.2 Comparison vs. state-of-the-art methods. Given the similarity of the MIM to the OPT methods, its computational efficiency, and for readability, here we compare only the MIM with the traditional and state-of-the-art methods (figures including OPT methods are in Appendix B). The MIM preserve X_1's influence with respect to the standard full model as $r(X_1, Z)$ increases (red and solid blue lines in Figures 4A, 4B, 4F, 4G). 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 4A). In the remainder of the paper we report results for the SHAP influence, since the results for MDE are qualitatively the same (Appendix C). Importantly, even though the MIM does not optimize for any fairness measure, it performs better in demographic disparity (Figure 4E) and all other fairness measures (Appendix B) than the traditional method dropping Z.

Other methods addressing discrimination either change the influence of X_1 with the growing correlation $r(X_1, Z)$ ("Exp Grad" methods in Figure 4) or use the protected attribute Z and thus discriminate directly ("Cal Eq Odds" and "Reweighing" methods in Figure 4). On the one hand, the method optimizing for parity of impact ("Exp Grad DP") in Scenario A unnecessarily decreases the influence of $\boldsymbol{X}_1$ (brown line in Figure 4A), which leads to an accuracy loss (Figure 4D), because its goal is to remove the correlation between $\hat{Y}$ and Z. In Scenario B, the changes in the influence of X_1 with the growing correlation are especially noteworthy. The affected methods ("Exp Grad") are increasingly influenced by X_1 as it gets more associated with the protected attribute (Figure 4F), despite X_1 not having impact on Y, because this enables them to increasingly utilize X_2 in their model of Y (Figure 4G) and improve accuracy (Figure 4I) under a respective fairness constraint. Other reductions approaches, constrained on equal opportunity and error ratio, yield similar outcomes (Appendix B). On the other hand, the methods allowing the influence of Z perform relative well in Scenario A, because they counteract discrimination by using Z directly

(violet and teal lines in Figures 4C, 4H) to maintain stable influence of X_1 and X_2 on $\hat{Y}$ (Figures 4A, 4B, 4F, 4G) and a high model accuracy (Figures 4D, 4I), independently of $r(X_1, Z)$. However, in Scenario B, where there is no discrimination, these methods introduce reverse discrimination to counteract the correlation between X_2 and Z, without considering the possibility that this correlation is a *fair relationship*, and achieve lower accuracy than the MIM (Figure 4I).

5.3 Real-world datasets

We train and test (80:20 random split) the evaluated methods on the COMPAS criminal recidivism dataset [30], German Credit, and Adult Census Income [16] 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. [6]. To make the presentation more clear, we exacerbate the racial bias by removing 500 samples of positive outcomes (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 [6]. 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 5) 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 5). While most of the evaluated models outperform the MIM in terms of demographic disparity (the rightmost column in Figure 5), they are either influenced by the protected attribute (the leftmost column in Figure 5) or do not preserve the influence of at least one of the most correlated attributes (the two middle columns in Figure 5) and have significantly lower accuracy (Figure 6), e.g., "Exp Grad" for COMPAS (Figures 5a & 6a). 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.

6 LIMITATIONS AND FUTURE WORK

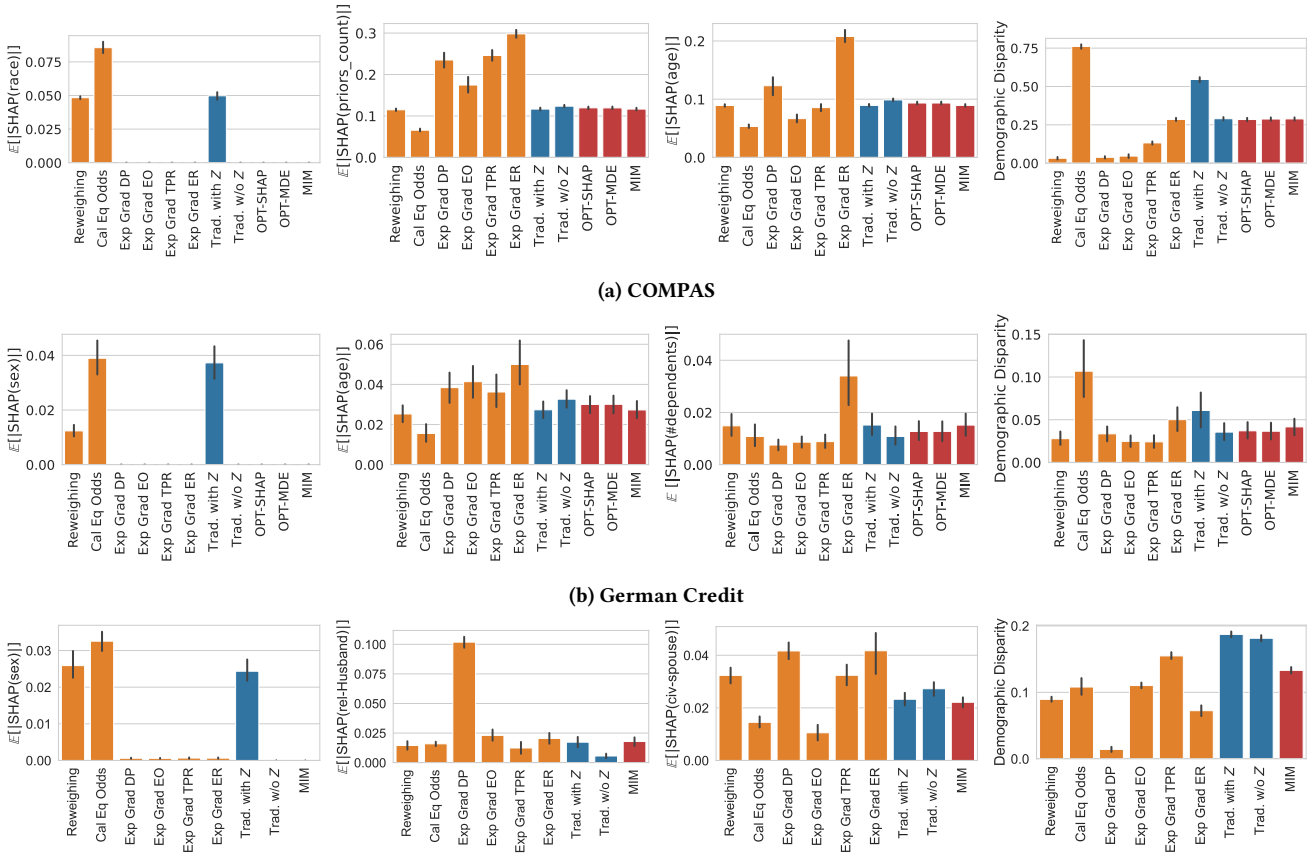
This manuscript focuses on two influence measures, MDE and SHAP, and corresponding loss functions for influence preservation. Prior studies show that input influence measures like SHAP can be fooled into stating that a protected attribute has no influence on a model [48]. With this, someone may be able 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.

Most importantly, any fairness objective can be misused by people to justify their systems as fair, especially if our limited understanding of causal processes happening in real-world decision-making adds up to the confusion. For instance, if a company develops a model of Y using $\boldsymbol{X}$ and some X_i is unfairly influenced, then first they shall apply our method to a model of X_i and second to a model of Y. An omission of the first step, whether intentional or unintentional, would result in indirect discrimination. In such contexts, we emphasize that understanding the causal processes relevant to the decision-making at hand in collaboration with domain experts and goodwill are 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 [56]. We measure influence of features to capture induced discrimination. Second, we show that state-of-the-art methods addressing 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

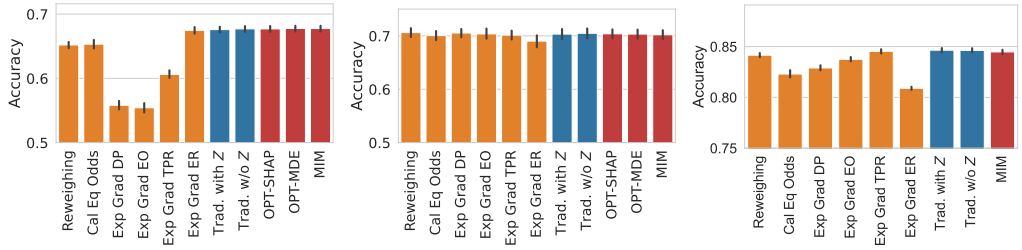


(a) COMPAS

(b) German Credit

(c) Adult Census Income

Figure 5: 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**

Figure 6: Accuracy for the evaluated models on the COMPAS, German Credit, and Adult Census Income datasets. Error bars show 95% confidence intervals.

in the circumstances where discrimination could have affected the training dataset.

ACKNOWLEDGMENTS

We thank Luis F. Lafuerza for his feedback and multiple rounds of comments and Isabel Valera, Muhammad Bilal Zafar, and Krishna Gummadi for discussions on early versions of this work. P.A.G. acknowledges support from Volkswagen Foundation (Ref. 92136) and DARPA and ARO (Cooperative Agreement No. W911NF-20-2-0005). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the DARPA or ARO, or the U.S. Government.

REFERENCES

[1] Alekh Agarwal, Aliiia Beygelzimer, Miroslav Dudfk, John Langford, and Wallach Hanna. 2018. A reductions approach to fair classification. *35th International Conference on Machine Learning, ICML 2018* 1 (2018), 102–119. arXiv:1803.02453v3

[2] Andrew Altman. 2016. Discrimination. In *The Stanford Encyclopedia of Philosophy* (2016 ed.), Edward N Zalta (Ed.). Metaphysics Research Lab, Stanford University.

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

[4] Anil Aswani and Matt Olfat. 2019. Optimization Hierarchy for Fair Statistical Decision Problems. (2019). arXiv:1910.08520 http://arxiv.org/abs/1910.08520

[5] Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace. 2019. *Consumer-Lending Discrimination in the FinTech Era.* Technical Report. National Bureau of Economic Research, Cambridge, MA. 1–51 pages. https://doi.org/10.3386/w25943

[6] 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. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. (Oct. 2018). https://arxiv.org/abs/1810.01943

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

[8] Silvia Chiappa. 2019. Path-Specific Counterfactual Fairness. *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (jul 2019), 7801–7808. https://doi.org/10.1609/aaai.v33i01.33017801

[9] Hana Chockler and Joseph Y. Halpern. 2003. Responsibility and blame: A structural-model approach. *IJCAI International Joint Conference on Artificial Intelligence* 22 (2003), 147–153. https://doi.org/10.1613/jair.1391 arXiv:0312038 [cs]

[10] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (jun 2017), 153–163. https://doi.org/10.1089/big.2016.0047 arXiv:1703.00056

[11] Anthony C. Constantinou, Yang Liu, Kiattikun Chobtham, Zhigao Guo, and Neville K. Kitson. 2020. Large-scale empirical validation of Bayesian Network structure learning algorithms with noisy data. arXiv:2005.09020 [cs.LG]

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

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

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

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

[16] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[17] Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2014. Certifying and removing disparate impact. (2014), 259–268. https://doi.org/10.1145/2783258.2783311 arXiv:1412.3756

[18] Malcolm Forster and Elliott Sober. 1994. 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 (1994), 1–35. https://doi.org/10.1093/bjps/45.1.1

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

[20] Nina Grgic-Hlaca, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human Perceptions of Fairness in Algorithmic Decision Making. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*. ACM Press, New York, New York, USA, 903–912. https://doi.org/10.1145/3178876.3186138 arXiv:1802.09548

[21] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, D D Lee, M Sugiyama, U V Luxburg, I Guyon, and R Garnett (Eds.). Curran Associates, Inc., 3315–3323. https://doi.org/10.1109/ICCV.2015.169 arXiv:1610.02413

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

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

[24] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (01 Oct 2012), 1–33. https://doi.org/10.1007/s10115-011-0463-8

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

[26] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG]

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

[28] 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. 2020. WILDS: A Benchmark of in-the-Wild Distribution Shifts. (2020), 1–87. arXiv:2012.07421 http://arxiv.org/abs/2012.07421

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

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

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

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

[33] Zachary C. Lipton and Jacob Steinhardt. 2019. Troubling trends in machine-learning scholarship. *Queue* 17, 1 (2019), 1–15. https://doi.org/10.1145/3317287.3328534 arXiv:1807.03341

[34] Scott M. Lundberg and Su In Lee. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* 2017-Decem, Section 2 (2017), 4766–4775. arXiv:1705.07874

[35] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna M. Wallach. 2021. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *CoRR* abs/2112.05675 (2021). arXiv:2112.05675 https://arxiv.org/abs/2112.05675

[36] Charles T. Marx, Richard Lanas Phillips, Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2019. Disentangling influence: Using disentangled representations to audit model predictions. *Advances in Neural Information Processing Systems* 32 (2019). arXiv:1906.08652

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

[38] Luca Oneto, Michele Donini, and Massimiliano Pontil. 2020. General Fair Empirical Risk Minimization. (2020), 1–8. https://doi.org/10.1109/ijcnn48605.2020.9206819 arXiv:1901.10080

[39] Judea Pearl. 2009. *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge University Press.

[40] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*. ACM Press, New York, New York, USA, 560. https://doi.org/10.1145/1401890.1401959

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

[42] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "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 (2016), 1135–1144. https://doi.org/10.1145/2939672.2939778 arXiv:1602.04938

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

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

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

[46] Glenn Shafer. 2001. Causality and Responsibility. In *Cardozo Law Review*. Number 22. ACM Press, New York, New York, USA, 101–123. https://doi.org/10.1007/978-3-7908-1792-8_23

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

[48] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*.

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

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

[51] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. *34th International Conference on Machine Learning, ICML 2017* 7 (2017), 5109–5118. arXiv:1703.01365

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

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

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

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

[56] Sandra Wachter. 2019. Affinity Profiling and Discrimination by Association in Online Behavioural Advertising. *SSRN Electronic Journal* (2019), 1–74. https://doi.org/10.2139/ssrn.3388639

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

[58] Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. 2019. PC-Fairness: A unified framework for measuring causality-based fairness. *Advances in Neural Information Processing Systems* 32, NeurIPS (2019). arXiv:1910.12586

[59] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web - WWW '17*. ACM Press, New York, New York, USA, 1171–1180. https://doi.org/10.1145/3038912.3052660 arXiv:1610.08452

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

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

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

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

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

[65] Lu Zhang, Yongkai Wu, and Xintao Wu. 2017. A causal framework for discovering and removing direct and indirect discrimination. *IJCAI International Joint Conference on Artificial Intelligence* 0 (2017), 3929–3935. https://doi.org/10.24963/ijcai.2017/549 arXiv:1611.07509