INFORMATION-THEORETIC QUANTIFICATION OF INHERENT DISCRIMINATION BIAS IN TRAINING DATA FOR SUPERVISED LEARNING

Sokrat Aldarmini and Mohamed Nafea
Missouri University of Science & Technology
{sokrataldarmini,mnafea}@mst.edu

ABSTRACT

Algorithmic fairness research has mainly focused on adapting learning models to mitigate discrimination based on protected attributes, yet understanding inherent biases in training data remains largely unexplored. Quantifying these biases is crucial for informed data engineering, as data mining and model development often occur separately. We address this by developing an information-theoretic framework to quantify the marginal impacts of dataset features on the discrimination bias of downstream predictors. We postulate a set of desired properties for candidate discrimination measures and derive measures that (partially) satisfy them. Distinct sets of these properties align with distinct fairness criteria like demographic parity or equalized odds, which we show can be in disagreement and not simultaneously satisfied by a single measure. We use the Shapley value to determine individual features' contributions to overall discrimination, and prove its effectiveness in eliminating redundancy. We validate our measures through a comprehensive empirical study on numerous real-world and synthetic datasets. For synthetic data, we use a parametric linear structural causal model to generate diverse data correlation structures. Our analysis provides empirically validated guidelines for selecting discrimination measures based on data conditions and fairness criteria, establishing a robust framework for quantifying inherent discrimination bias in data.

1 INTRODUCTION

The integration of data-driven learning models in various aspects of human life, e.g., criminal justice (Kirchner et al., 2016), finance (Diwate et al., 2021), and healthcare (Kinyanjui et al., 2019), raised concerns about discriminatory practices based on legally protected attributes such as race or gender (Munoz et al., 2016; Mökander et al., 2022). Algorithmic fairness examines how models may inherit discriminatory biases from training data, defines fairness criteria (Calders et al., 2009; Dwork et al., 2012; Hardt et al., 2016), and develops algorithmic methods to achieve them (Kamishima et al., 2011; Zemel et al., 2013). Fairness criteria include individual fairness (similar individuals are treated similarly) (Dwork et al., 2012); demographic parity (equal positive rates across demographic groups) (Calders et al., 2009); equalized odds and equality of opportunity (equal error rates across demographic groups) (Hardt et al., 2016); and predictive parity (equal precision across demographic groups) (Kleinberg et al., 2017). Methods to achieve fairness criteria include pre-processing (Kamiran & Calders, 2012; Calmon et al., 2017), in-processing (Kamishima et al., 2011; Zafar et al., 2017), and post-processing (Hardt et al., 2016; Petersen et al., 2021; Xian et al., 2023).

Despite considerable progress in algorithmic fairness, little research addresses quantifying discrimination bias *inherent* in the training data. While input feature impact on model accuracy is well-studied for reasons like performance improvement (AlSagri & Ykhlef, 2020), dimensionality reduction (Kira & Rendell, 1992), data representation (Yang & Moody, 1999), and interpretability (Scott et al., 2017), quantifying feature impact on discrimination bias is more challenging. This is due to the complex interactions among dataset features, group memberships, target labels, and model predictions, as well as the reliance of fairness criteria on these variables. For instance, features with high discriminatory impact may also be strongly relevant for accurate predictions (Grgic-Hlaca et al., 2016). Further, most existing work assumes access to model predictions to evaluate fairness criteria, which is impractical when data engineering and model development are conducted separately (Zemel et al., 2013). We address these challenges by developing a framework for quantifying marginal discriminatory impacts of features on downstream models without requiring access to predictions.

Our quantification relies on constructing novel information-theoretic measures, defined for sets of features, and deducing marginal contributions of individual features using the Shapley value function (Shapley, 1953; Scott et al., 2017). We adopt an axiomatic approach, based on partial decomposition of information (Bertschinger et al., 2014) and causal reasoning (Peters et al., 2017), advocating for a set of desired properties and constructing measures that satisfy them. *A novel aspect in our framework is to relate these properties to various fairness criteria, and study the tensions associated with a single measure simultaneously achieving them*. We conduct an extensive ablation study on numerous real-world and synthetic datasets, yielding a constructive guideline for which discrimination measure shall be used given certain fairness criteria and dataset conditions.

1.1 Related work

Several studies explored methods to quantify feature impact on accuracy of downstream learning models (Kira & Rendell, 1992; Scott et al., 2017). However, only a handful of works studied both accuracy and discrimination impacts (Khodadadian et al., 2021; Grgic-Hlaca et al., 2016; Dutta et al., 2022). Fewer still addressed the problem without assuming access to model's predictions (Khodadadian et al., 2021; Pelegrina et al., 2024). Our work adopts a similar axiomatic framework as in (Khodadadian et al., 2021), however, our construction of discrimination measures is different. *We extensively revisit the desired properties proposed in (Khodadadian et al., 2021). We relate our revised properties to existing fairness criteria to study the tension in constructing measures that simultaneously achieve them; a critical analysis missing from (Khodadadian et al., 2021).* Another line of work adopts a framework similar to (Khodadadian et al., 2021) to quantify feature contribution to the discrimination of a predefined model (Dutta et al., 2022). Dutta et al. (2020; 2021) follow an axiomatic approach to quantify non-exempt discrimination for a predefined model, i.e., the part of discrimination that cannot be accounted for by features critical for accurate predictions.

A framework similar to ours, in avoiding dependence on a predefined model, employs Shapley values to quantify feature importance and identify disparity-prone features which result in disparate outcomes (Pelegrina et al., 2024). Unlike (Pelegrina et al., 2024), we focus on quantifying inherent discrimination bias with respect to predefined sensitive attributes, e.g., race or gender, rather than identifying potential sensitive attributes. Pelegrina et al. (2024) utilized a statistical measure evaluated on sets of features (an empirical estimate of the normalized cross-covariance (Fukumizu et al., 2007)) which has proven effective in a non-coalition framework (Pelegrina et al., 2023). *In contrast, our work proposes distributional information-theoretic measures based on partial decomposition of information, facilitating more control over the aspects a measure can capture in the data.*

The Shapley value function (Shapley, 1953), a concept from game theory used to determine players' marginal contributions to a game's overall utility, has gained recent attention for interpreting learning models (Scott et al., 2017; Ghorbani & Zou, 2020). Researchers applied Shapley value for interpretability in various ways to (i) quantify feature contributions to model accuracy (Janzing et al., 2020); (ii) assess the impact of sensitive attributes on model predictions (Mase et al., 2021), and (iii) evaluate individual neuron contributions in deep neural networks (Ghorbani & Zou, 2020). Shapley value was also used for feature selection by identifying features most relevant for prediction or model accuracy (Cohen et al., 2005). *This work employs Shapley value to quantify the marginal impacts of dataset features on the discrimination bias of any downstream learning model*. Importantly, our theoretical results (Thm. 3.4) highlight the advantage of using Shapley value for quantifying *marginal* discriminatory impacts, particularly when there is substantial feature redundancy. This result supports the rather empirical conclusion of Pelegrina et al. (2024), which justified, via a synthetic data experiment, the use of Shapley value for assessing feature importance over non-coalition methods (Pelegrina et al., 2023) when redundancy among input features is present.

The scarcity of benchmark datasets remains a significant challenge in algorithmic fairness (Ding et al., 2021). Several studies proposed methods to induce biases in existing real-world datasets (Jiang et al., 2024; Wen et al., 2021; Barbierato et al., 2022). Fewer works explored generating *purely* synthetic biased data (Barbierato et al., 2022; Baumann et al., 2023). Barbierato et al. (2022) present a methodology for controlling bias in synthetic data generation with categorical features based on a Gaussian probabilistic network and biased discretization. Baumann et al. (2023) explore fundamental bias types as well as encapsulating them within a single feature. *In contrast, we introduce a simple yet effective model for generating purely synthetic biased datasets using a parametric linear structural causal model (SCM). Our generating model yields a large number of biased synthetic datasets through a small number of tunable parameters.*

1.2 PROBLEM DESCRIPTION AND OUR CONTRIBUTIONS

We aim to quantify inherent data biases that could be replicated or amplified by downstream models. The research question we pose here is *how to quantify the impact of dataset features on discrimination bias of any downstream model without assuming access to model predictions, relying only on dataset variables?* This is motivated by three key reasons. First, data engineering and model development are often separate tasks conducted by separate entities, necessitating model-agnostic quantification of the discrimination bias inherent in the data. Second, we seek to disentangle this bias from model-specific amplifications. Third, similar to Pelegrina et al. (2024), we avoid the computational complexity associated with using trained models in a coalition-based framework, like Shapley value. Our formulation gives rise to several challenges, which we address through the following contributions:

- We construct several discrimination measures (for sets of features) via an axiomatic framework and apply Shapley value to deduce marginal discriminatory impacts. Compared to (Khodadadian et al., 2021), we extensively revisit their desired properties for a discrimination measure, relax some constraints we deem unnecessary, and postulate two *undesired properties*. Further, unlike Khodadadian et al. (2021), which proposes a single discrimination measure, we advocate for using diverse measures aligned with distinct fairness criteria and data conditions. We correlate the notions of demographic parity (DP) and equalized odds (EO) with two distinct sets of desired properties, demonstrate the tensions for a single measure simultaneously satisfying both property sets (Lemma 3.1 and Example 3.3), and empirically validate these tensions (Sections 4 and 5).
- We provide a theoretical result demonstrating that using Shapley values for determining marginal discriminatory impacts of features moderates the effect of redundancy among coalitions of features (Thm. 3.4). As argued by Pelegrina et al. (2024), this redundancy can be over-quantified when using methods that do not account for interactions among features (non-coalitional methods). Our theoretical result supports the rather empirical finding of Pelegrina et al. (2024).
- We propose a parameterized linear SCM for synthetic data generation of a large number of biased datasets, and examine various parameter configurations of the generating model (Section 4).
- We conduct a comprehensive ablation study on numerous real-world and synthetic datasets to empirically demonstrate the efficacy of our measures in capturing marginal discriminatory impacts of features. Our empirical evaluation supports our theoretical framework and yields a principled guideline for which discrimination measures to use under given data conditions and fairness criteria.

2 BACKGROUND AND PRELIMINARIES

We begin with a background about common machine learning (ML) fairness notions and the theoretical tools we use to quantify marginal discriminatory impacts, namely, *partial information decomposition* (PID) and the *Shapley value function*. We then highlight some existing discrimination measures. Let us first highlight the notations we use throughout the paper.

Notation. Consider a dataset (X^n, A, Y), with n general features $X^n = \{X_1, \cdots, X_n\}$, a sensitive attribute A, and true target label Y. $[n]$ denotes the sequence $\{1, \ldots, n\}$. For $S \subseteq [n]$, $X_S \triangleq \{X_i : i \in S\}$ is the subset of X^n indexed by S. $S^c \triangleq [n] \backslash S$. $|S|$ denotes the cardinality of S. $\emptyset$ is the empty set and $2^{[n]}$ is the power set of $[n]$. For a random variable A, $\mathcal{A}$ and p_A denote its sample space and probability distribution. $\Delta_{\mathcal{A}}$ is the probability simplex over $\mathcal{A}$. For two random variables A, B, $p_{A,B}$ and $p_{A|B}$ denote their joint and conditional distributions. $\mathbb{E}_{a \sim p_A}[f(a)]$ is the expected value of $f(a)$ when a is sampled from p_A. $I(A; B)$ and $I(A; B|C)$ denote the mutual information between A and B, and their conditional mutual information given C. Consider the predictor h which predicts Y based on X^n, i.e., $\hat{Y} \triangleq h(X^n)$. For a subset X_S, $S \subseteq [n]$; $\hat{Y}^{(S)} \triangleq h|_{X_S}(X^n)$ refers to the restriction of h to X_S, i.e., input features in X_{S^c} are set as constants to their mean values.

2.1 FAIRNESS NOTIONS IN MACHINE LEARNING

Existing fairness notions require access to predictions to enforce statistical or counterfactual parity of decisions across pairs/groups of individuals. These are categorized into *individual* and *group* notions. *Individual fairness* advocates for treating similar individuals similarly via a Lipschitz constraint on the outcomes of a randomized predictor (Dwork et al., 2012). *Group fairness* notions consider groups of individuals defined by their group memberships A. For example, **Demographic Parity (DP)** requires independence of A from the prediction $\hat{Y} = h(X^n)$ (Calders et al., 2009). While widely used, DP cripples accuracy when base rates across demographic groups, i.e., $p_{Y|A=a}(1)$ for $a \in \mathcal{A}$,

are unequal (Hardt et al., 2016). To circumvent this, Hardt et al. (2016) proposed **Equalized Odds (EO)**, which requires conditional independence of $\hat{Y}$ and A given the true label Y.

2.2 Partial Information Decomposition (PID)

Consider three random variables A, B, R and let $p_{A,B,R}$ be their joint distribution. PID decomposes the mutual information (MI) between the pair (A, B) and a reference R, i.e., $I(R;(A,B))$, into four distinct non-negative components as follows (Bertschinger et al., 2014):

$$I(R;(A,B)) = UI(R;A\backslash B) + UI(R;B\backslash A) + SI(R;A,B) + CI(R;A,B). \tag{1}$$

$UI(R;A\backslash B)$ is the unique information about R contained in A but not in B. $UI(R;B\backslash A)$ is defined similarly. $SI(R;A,B)$ is the shared information that each of A and B have, individually, about R. $CI(R;A,B)$ is the complementary information about R that requires both A and B together, rather than individually, to obtain. Based on Figure 1, the following identities hold:

$$I(R;A) = UI(R;A\backslash B) + SI(R;A,B) \text{ and } I(R;B) = UI(R;B\backslash A) + SI(R;A,B). \tag{2}$$

We adopt the unique information characterization proposed by Bertschinger et al. (2014)[1]. Specifically, let $I_q(R;A|B)$ be the conditional mutual information (CMI) w.r.t. the joint distribution $q \in \Delta^*_{\mathcal{A},\mathcal{B},\mathcal{R}} \triangleq \{q \in \Delta_{\mathcal{A},\mathcal{B},\mathcal{R}} : q_{R,A} = p_{R,A} \text{ and } q_{R,B} = p_{R,B}\}$. $\Delta^*_{\mathcal{A},\mathcal{B},\mathcal{R}}$ is the simplex of all joint probability distributions for which their marginals $q_{R,A}$ and $q_{R,B}$ are equal to the true marginals $p_{R,A}$ and $p_{R,B}$. Then,

$$UI(R;A|B) \triangleq \min_{q\in\Delta^*_{\mathcal{A},\mathcal{B},\mathcal{R}}} I_q(R;A|B). \tag{3}$$

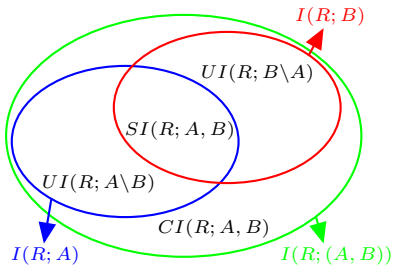


Figure 1: Partial Information Decomposition of $I(R;(A,B))$.

Conditional PID. Consider a fourth random variable Z and the joint distribution $p_{A,B,R,Z}$. We define the conditional unique information about R in A but not in B, given Z, as $CUI_Z(R;A\backslash B) \triangleq \mathbb{E}_{z\sim p_Z}[UI(R_z;A_z\backslash B_z)]$: R_z, A_z, B_z are distributed according to $p_{R,A,B|Z=z}$. The conditional shared and complementary information, $CSI_Z(R;A,B)$ and $CCI_Z(R;A,B)$, are defined similarly. The decomposition in (1), (2) holds for conditional PID components.

Lemma 2.1. *Monotonicity properties of PID (Rauh et al., 2014). $UI(R;A\backslash B)$ is non-decreasing in R and A, and non-increasing in B. $SI(R;A,B)$ is non-decreasing in both A and B.*

The same properties hold for conditional PID components due to the linearity of expectation.

2.3 The Shapley value function

We use Shapley value to compute the marginal discriminatory impacts of features. A cooperative game is defined by a set of n players and a pay-off function $v(S) : 2^{[n]} \to \mathbb{R}$, $S \subseteq [n]$, and $v(\emptyset) = 0$. Shapley value estimates the marginal contribution of player i to the pay-off as (Shapley, 1953):

$$\phi_i(v) = \sum\nolimits_{S\subseteq[n]\backslash\{i\}} \frac{1}{n!}\left((n-1-|S|)!|S|!\right)\left(v(S\cup\{i\}) - v(S)\right), \qquad i \in [n]. \tag{4}$$

The choice of the weights in (4) makes it the unique function satisfying (Young, 1985): (1) *Symmetry*: If $\forall S \subseteq [n]\backslash\{i,j\}, i \neq j$, $v(S\cup\{i\}) = v(S\cup\{j\})$, then $\phi_i(v) = \phi_j(v)$. (2) *Efficiency*: $\sum_{i\in[n]}\phi_i(v) = v([n])$. (3) *Dummy player*: If $\forall S \subseteq [n]\backslash\{i\}, v(S\cup\{i\}) = v(S)$, then $\phi_i(v) = 0$. (4) *Linearity*: For pay-off functions v_1, v_2 and $\forall\alpha_1,\alpha_2 \in \mathbb{R}$, $\phi_i(\alpha_1 v_1 + \alpha_2 v_2) = \alpha_1\phi_i(v_1) + \alpha_2\phi_i(v_2)$.

2.4 Existing measures

Next, we revisit some existing discrimination measures, which we relate to our framework in Section 3. Khodadadian et al. (2021) proposed the following measure,

$$v^D(X_S) = I(A;X_S)I(A;X_S|Y)SI(Y;X_S,A), \tag{5}$$

which does not require access to predictions and satisfies (1) *non-negative and non-decreasing*, (2) *A-independence*: If $X_S \perp A$, then $v(X_S) = 0$, (3) *AY-independence*: If $X_S \perp A|Y$, then $v(X_S) = 0$, and (4) *Y-independence*: If $X_S \perp Y$, then $v(X_S) = 0$. In Section 3, we revisit these desired properties

[1]Note that $I(R;(A,B))$, $I(R;A)$, and $I(R;B)$ are computed directly using the joint distribution $P_{A,B,R}$. Therefore, using (1) and (2), explicitly characterizing the four components on the RHS of (1) requires explicit characterization of either the unique, shared, or complementary information.

for a discrimination measure. In contrast, the following measures require access to either a white-box predictor h or the joint distribution of the predictions $\hat{Y}$ and the data (X^n, A, Y). Dutta et al. (2022) introduced two measures to quantify the discrimination impact of X_S on a predefined predictor $\hat{Y} = h(X^n)$. The first uses $\hat{Y}^{(S)} = h|_{X_S}(X^n)$ as

$$v^D(X_S) = I(A; \hat{Y}^{(S)}): \quad \text{Mutual information between } A \text{ and } h \text{ restricted to } X_S. \tag{6}$$

The second measure can be computed using the joint distribution $p_{X,A,\hat{Y}}$ as

$$v^D(X_S) = SI(A; \hat{Y}, X_S): \quad \text{Shared information between } \hat{Y} \text{ and } X_S \text{ about } A. \tag{7}$$

Dutta et al. (2020) introduced two measures to quantify *non-exempt discrimination*, i.e., cannot be explained by features critical for prediction, X_C, where $C \subseteq [n]$. These are defined as follows:

$$v^D(X_C) = UI(A; \hat{Y} \backslash X_C): \quad \text{Unique Information about } A \text{ in } \hat{Y} \text{ but not in } X_C. \tag{8}$$

$$v^D(X_C) = I(A; \hat{Y} | X_C): \quad \text{Conditional mutual information between } A \text{ and } \hat{Y} \text{ given } X_C. \tag{9}$$

3 Proposed measures

This section outlines our *axiomatic framework* for developing *predictor-free discrimination measures*. We begin by postulating a set of *desired properties*. We then derive a set of candidate measures based on these properties. Our approach differs from that in (Khodadadian et al., 2021) in a few key aspects: (1) We do not mandate our measures to adhere to all desired properties since for a given data distribution and fairness criteria, desired properties for a discrimination measure could exhibit inherent conflict. (2) For the data-generating causal diagram (Figure 2), we allow for a direct influence from the sensitive attribute A to the target label Y, besides influencing Y through the remaining features X^n. Further, similar to Kusner et al. (2017), we assume ancestral closure of A, i.e., A is a root variable, since a parent of a sensitive attribute itself is a sensitive attribute. For simplicity, we assume that Y is a sink node, i.e., there are no feedback loops from the target label Y to data features $\{A, X^n\}$. Finally, we assume that no single observed feature is a deterministic function of other features.

3.1 Desired properties

We derive discrimination measures for subsets of features X_S, $S \subseteq [n]$. A discrimination measure $v^D(X_S)$ should satisfy

Property 1 (Non-negativity). $v^D(X_S) \geq 0$, *with equality if* $S = \emptyset$.

Property 2 (Monotonicity). $v^D(X_{S_1}) \leq v^D(X_{S_2})$, *for any* $S_1, S_2 \subseteq [n]$ *such that* $S_1 \subseteq S_2$.

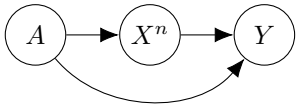


Figure 2: Causal diagram for data variables.

All our proposed measures satisfy Property 1 (*non-negativity*). In contrast, depending on the data, Property 2 (*monotonicity*) may not be essential since including certain features could resolve the discriminatory impact of others. For instance, including features correlated with Y yet independent of A could reduce the discrimination impact of X_S for a target accuracy. Notably, marginal discriminatory impacts deduced using the Shapley value for a monotonic measure are non-negative; but could be negative if the measure is not monotonic. Removal of features with negative marginal discriminatory impacts may result in increasing the discrimination of a downstream predictor. The following properties are inspired by demographic parity (DP) and equalized odds (EO) criteria.

Property 3 (DP-independence). $X_S \perp A \Leftrightarrow v^D(X_S) = 0$.

Property 4 (EO-independence). $X_S \perp A | Y \Leftrightarrow v^D(X_S) = 0$.

Consider restricting Property 3 to a specific predictor h by replacing X_S with $\hat{Y}^{(S)} = h|_{X_S}(X^n)$. A measure v^D satisfying Property 3 equals zero iff $\hat{Y}^{(S)}$ satisfies DP. A similar argument holds for Property 4. The following Lemma demonstrates the inherent conflict between Properties 3 and 4:

Lemma 3.1. *There is no measure that satisfies both Property 3 and Property 4, simultaneously.*

Proof. Suppose that a measure v^D satisfies Properties 3 and 4. For X_S s.t. $X_S \perp A$ but $X_S \not\perp A | Y$, we have that $v^D(X_S) = 0$ by Property 3, but $v^D(X_S) \neq 0$ by Property 4; a contradiction. □

Remark 1. *Lemma 3.1 resembles the result in (Kleinberg et al., 2017) which states that no predictor can satisfy DP and EO simultaneously except for the cases of a random predictor or when base rates across demographics are equal. Lemma 3.1 establishes a similar result for a discrimination measure v^D. For a general data distribution, v^D cannot satisfy DP- and EO-independence simultaneously, as this implies the equivalence $X_S \perp A \Leftrightarrow X_S \perp A | Y$, which does not hold in general.*

For a binary sensitive attribute and target label, the following lemma specifies the necessary conditions for which $X_S \perp A$ and $X_S \perp A | Y$ are equivalent.

Lemma 3.2. *For a subset of features X_S, binary target label Y and binary sensitive attribute A, such that $X_S \perp A$ and $X_S \perp A|Y$, it follows that either (1) $X_S \perp Y$, i.e., X_S is irrelevant to the prediction task, or (2) $A \perp Y$, i.e., the base rates are equal across demographic groups.*

All the proofs in this work are deferred to the appendix. We also consider relaxed versions of Properties 3 and 4 by replacing $\Leftrightarrow$ with $\Rightarrow$, as follows:

Definition 1. *A measure v^D satisfies **relaxed DP-independence** if $X_S \perp A \implies v^D(X_S) = 0$.*

Definition 2. *A measure v^D satisfies **relaxed EO-independence**, if $X_S \perp A|Y \implies v^D(X_S) = 0$.*

These relaxed versions are identical to "A-independence" and "AY-independence" in (Khodadadian et al., 2021), respectively (cf. Section 2.4). A measure satisfying both relaxed DP- and EO-independence (e.g., (5)) could misrepresent the marginal discriminatory impacts computed using the Shapley value function in (4). We illustrate this observation by the following example.

Example 3.3. *Consider features X_1, X_2 s.t. $X_1 \not\perp A$ but $X_1 \perp A|Y$ and $X_2 \perp A$ but $X_2 \not\perp A|Y$. Consider a predictor $\hat{Y} = h(X_1, X_2)$ and its restrictions $\hat{Y}^{(\{1\})} = h|_{X_{\{1\}}}$ and $\hat{Y}^{(\{2\})} = h|_{X_{\{2\}}}$. Suppose $\hat{Y} \not\perp A$, $\hat{Y} \not\perp A|Y$ (h violates DP and EO). We expect: (1) $\hat{Y}^{(\{2\})} \perp A$ but not necessarily $\hat{Y}^{(\{1\})} \perp A$, hence using X_1 as an input to h causes DP violation; (2) Similarly, $\hat{Y}^{(\{1\})} \perp A|Y$ but not necessarily $\hat{Y}^{(\{2\})} \perp A|Y$, hence X_2 causes EO violation. For non-negative measure v^D, the marginal discriminatory impact for X_i, $i = 1, 2$, is expressed using (4) as $\phi_i^D = (v^D(X_{\{1,2\}}) - (-1)^i v^D(X_{\{1\}}) + (-1)^i v^D(X_{\{2\}}))/2$. If v^D satisfies both relaxed DP- and EO-independence, then $v^D(X_{\{1\}}) = v^D(X_{\{2\}}) = 0$. Thus, $\phi_1^D = \phi_2^D = v^D(X_{\{1,2\}})/2$. This contradicts the expected contribution of X_1 and X_2 to the discrimination of $\hat{Y}$. When v^D satisfies only DP-independence or its relaxed version, $v^D(X_2) = 0$ and $v^D(X_1) > 0$ (or ≥ 0 for relaxed version), implying that $\phi_1^D > \phi_2^D$ (or $\geq$ for relaxed version). This aligns with the expected contributions of X_1 and X_2 to the discrimination of $\hat{Y}$ w.r.t. DP. A similar argument can be made for the EO counterpart.*

Example 3.3 highlights the limitation of measures satisfying both relaxed DP- and EO-independence. Thus, we advocate for measures that satisfy either of these properties but not both (depending on the desired fairness criteria). Finally, we propose the following properties, intended to prevent attributing discrimination bias to features X_S when the remaining features X_{S^c} could account for that bias.

Property 5 (DP-blocking). $X_S \perp A|X_{S^c} \implies v^D(X_S) = 0$.

Property 6 (EO-blocking). $X_S \perp A|\{X_{S^c}, Y\} \implies v^D(X_S) = 0$

Requirements similar to DP- and EO-Blocking were proposed in (Frye et al., 2020; Khodadadian et al., 2021) but for quantifying feature importance for prediction accuracy. Frye et al. (2020) incorporate causal knowledge through asymmetric Shapley values, requiring that if X_i is a deterministic causal ancestor of X_j, no importance should be attributed to X_j. Throughout the paper, we distinguish between properties aligned with DP and those aligned with EO, as highlighted in the following:

Definition 3. *DP- (resp. EO-) independence , relaxed DP- (resp. EO-) independence, and DP- (resp. EO-) blocking are aligned with demographic parity (resp. equalized odds), and collectively referred to as "DP-properties" (resp. "EO-properties").*

Next, we highlight the following *two undesired conditions*, met by some of our measures:

Condition 1. $X_S \perp Y \implies v^D(X_S) = 0$.

Condition 2. $A \perp Y \implies v^D(X_S) = 0$ *for all* $S \subseteq [n]$.

Condition 1 was proposed by Khodadadian et al. (2021) as a desired property, to be satisfied by the measure in (5). We argue that this condition is unnecessarily restrictive. To elaborate, $X_S \perp Y$ implies that X_S is unnecessary for prediction; but not that X_S has no discriminatory impact when used. For instance, a model developer may opt out of implementing a thorough feature-selection and include features irrelevant for prediction. Based on our framework, quantifying features' discriminatory impact on any downstream predictor should not enforce such features (those marginally independent of Y) to possess no discriminatory impact. Condition 2 limits the measure's ability to capture inherent discrimination bias for any features subset when base rates are equal across demographics. This is inspired from the result in (Locatello et al., 2019, Thm. 1), which shows that $A \perp Y$ does not guarantee DP of a downstream predictor even for the optimal Bayes predictor. Thus, we argue that Condition 2 is undesirable for a discrimination measure as it cripples the measure efficacy.

3.2 THE MEASURES

Next, we propose a set of candidate discrimination measures that partially satisfy the desired properties in Section 3.1. Based on Definition 3, we categorize these into: (i) *Measures aligned with DP*, satisfying only DP-properties, and (ii) *Measures aligned with EO*, satisfying only EO properties.

3.2.1 MEASURES ALIGNED WITH DP

A measure for the discriminatory impact of a subset X_S should quantify the amount of information X_S has about the sensitive attribute A, as described in our first candidate measure below:

Candidate measure 1: $v_1^D(X_S) = I(A; X_S)$. $v_1^D(X_S)$ is (i) *non-negative (Property 1)* due to non-negativity of KL-divergence; (ii) *monotonically non-decreasing (Property 2)* since $I(A; X_{S_2}) = I(A; X_{S_1}) + I(A; X_{S_2 \setminus S_1} | X_{S_1})$ for $S_1 \subseteq S_2 \subseteq [n]$ and $I(A; X_{S_2 \setminus S_1} | X_{S_1}) \geq 0$; (iii) *satisfies DP-independence (Property 3)* since $I(A; X_S) = 0$ iff $p_{X_S, A} = p_{X_S} p_A$ (Cover & Thomas, 2006). This measure is an adaptation of (6) or (7), by replacing $\hat{Y}$ with X_S where X_S can be viewed as a proxy for $\hat{Y}$ as we seek to quantify its discriminatory impact on the downstream predictor $\hat{Y} = h(X^n)$.

Note that $I(A; X_S) = UI(A; X_S \setminus X_{S^c}) + SI(A; X_S, X_{S^c})$. Thus, $v_1^D(X_S)$ attributes to X_S its unique information about A (not in X_{S^c}) and its shared information, with X_{S^c}, about A. This shared information is also attributed to X_{S^c} in $I(A; X_{S^c})$. We may wish to avoid this double attribution and use the measure $v_{1,u}^D(X_S) = UI(A; X_S \setminus X_{S^c})$. $v_{1,u}^D$ is (i) non-negative due to the non-negativity of PID components (cf. Section 2.2); (ii) monotonically non-decreasing (cf. Lemma 2.1); (iii) satisfies relaxed DP-independence (Definition 1) since $A \perp X_S$ implies $I(A; X_S) = UI(A; X_S \setminus X_{S^c}) + SI(A; X_S, X_{S^c}) = 0$; and (iv) satisfies DP-blocking (Property 5) since $A \perp X_S | X_{S^c}$ implies $I(A; X_S | X_{S^c}) = UI(A; X_S \setminus X_{S^c}) + CI(A; X_S, X_{S^c}) = 0$. Note that this measure is similar to (8), by replacing $\hat{Y}$ with X_S and X_C with X_{S^c}. We could also attribute to X_S the complementary information about A, provided jointly by X_S and X_{S^c}, since this information would be lost if X_S were removed. This results in the measure $v_{1,c}^D(X_S) = I(A; X_S | X_{S^c})$, which is also non-negative, monotonically non-decreasing (since $I(A; X_S | X_{S^c}) = I(A; X^n) - I(A; X_{S^c})$ and $I(A; X_{S^c})$ is non-increasing), and satisfies DP-blocking. $v_{1,c}^D$ is similar to (9), by replacing $\hat{Y}$ with X_S and X_C with X_{S^c}. Nevertheless, $v_1^D, v_{1,u}^D, v_{1,c}^D$ all result in identical marginal discriminatory impact, when Shapley value (4) is applied, as shown next.

Theorem 3.4. *Let $v_{SI}(X_S) = SI(A; X_S, X_{S^c})$, $v_{CI}(X_S) = CI(A; X_S, X_{S^c})$, $v_{CSI}(X_S) = CSI_Y(A; X_S, X_{S^c})$, $v_{CCI}(X_S) = CCI_Y(A; X_S, X_{S^c})$ be pay-off functions. The marginal contributions $\phi_i(v_{SI})$, $\phi_i(v_{CI})$, $\phi_i(v_{CSI})$, $\phi_i(v_{CCI})$, for all $i \in [n]$, using* (4)*, are equal to zero.*

Since $v_1^D(X_S) = v_{1,u}^D(X_S) + v_{SI}(X_S)$, $v_{1,c}^D(X_S) = v_{1,u}^D(X_S) + v_{CI}(X_S)$, Thm. 3.4 and linearity of Shapley value imply that $\phi_i(v_1^D) = \phi_i(v_{1,u}^D) = \phi_i(v_{1,c}^D)$. See Remark 2 for additional comments.

The measures $v_{1,u}^D, v_1^D, v_{1,c}^D$ are independent of Y, hence ignore Y's influence on the discrimination bias of downstream predictors. To include Y, consider the decomposition $I(A; X_S) = UI(A; X_S \setminus Y) + SI(A; X_S, Y)$. $UI(A; X_S \setminus Y)$ is the unique information about A in X_S but not in Y. Thus, it should be possible to find a representation of X_S that maintains its expressiveness w.r.t Y, yet discards $UI(A; X_S \setminus Y)$. The remaining component, $SI(A; X_S, Y)$, is discriminatory since it is a shared information between Y and X_S about A. This leads us to the following measure.

Candidate measure 2: $v_2^D(X_S) = SI(A; X_S, Y)$, is non-negative, monotonically non-decreasing, and satisfies relaxed DP-independence since $A \perp X_S$ implies $I(A; X_S) = UI(A; X_S \setminus Y) + SI(A; X_S, Y) = 0$. Yet, $v_2^D(X_S)$ meets undesired Condition 2 since $A \perp Y$ implies $I(A; Y) = UI(A; Y \setminus X_S) + SI(A; Y, X_S) = 0$. Thus, $v_2^D(X_S)$ could fail to capture the discriminatory impact of X_S when $A \perp Y$. Further, as shown by our experiments, when a simple predictor is used and the data exhibits very large $UI(A; X_S \setminus Y)$, $v_2^D(X_S)$ might not perform well in capturing the predictor's discrimination bias (on some datasets), since the predictor may overfit into this information.

Another way to incorporate Y into the construction of v^D is through the decomposition $I(X_S; A, Y) = UI(X_S; Y \setminus A) + UI(X_S; A \setminus Y) + SI(X_S; A, Y) + CI(X_S; A | Y)$. The unique information about X_S in Y but not in A, $UI(X_S; Y \setminus A)$, is not a discriminatory component. The remaining components however are discriminatory, as they quantify the information about X_S that is unique in A, shared between A and Y, or require A and Y jointly. We use these three discriminatory components to construct two measures, v_3^D, which is aligned with DP, and v_4^D, which is aligned with EO.

Candidate measure 3: $v_3^D(X_S) = SI(X_S; A, Y)$ is non-negative and satisfies relaxed DP-independence since $A \perp X_S$ implies $I(X_S; A) = UI(X_S; A \setminus Y) + SI(X_S; A, Y) = 0$. Yet, it meets undesired Condition 1, since $X_S \perp Y$ implies $I(X_S, Y) = UI(X_S; Y \setminus A) + SI(X_S; A, Y) = 0$. v_3^D

can be viewed as an adaptation of v_1^D by discarding $UI(X_S; A\backslash Y)$, the unique information about X_S in A but not in Y, which could be discriminatory. Our experiments demonstrate that when $UI(X_S; A\backslash Y)$ is large, the predictor could overfit into this component. In contrast, a predictor optimized to achieve DP could avoid overfitting into $UI(A; X_S\backslash Y)$ or $UI(X_S; A\backslash Y)$. This however assumes that data generation and model development are conducted by the same "trusted" party, while our framework aims to quantify discrimination bias in the data which could potentially be used by untrusted parties. v_3^D is the only measure we propose that is not non-decreasing, hence results in positive and negative marginal impacts (see Section 4).

3.2.2 Measures aligned with EO

We adapt the measures from Section 3.2.1 to be aligned with EO, by conditioning on Y.

Candidate measure 4: $v_4^D(X_S) = I(A; X_S|Y)$, is non-negative, monotonically non-decreasing, and satisfies EO-independence. Using a similar discussion as for candidate measure v_1^D, we obtain the measures: $v_{4,u}^D(X_S) = CUI_Y(A; X_S\backslash X_{S^c})$ which is non-negative, monotonically non-decreasing, and satisfies relaxed EO-independence and EO-blocking; as well as $v_{4,c}^D(X_S) = I(A; X_S|X_{S^c}, Y)$ which is non-negative, monotonically non-decreasing, and satisfies EO-blocking.

Note that v_4^D, $v_{4,u}^D$, $v_{4,c}^D$ are adapted version of v_1^D, $v_{1,u}^D$, $v_{1,c}^D$ by conditioning on Y. For measures v_2^D, v_3^D, conditioning on Y reduces their values to zero. Once again, using Thm. 3.4, it follows that the marginal impacts $\phi_i(v_4^D)$, $\phi_i(v_{4,u}^D)$, and $\phi_i(v_{4,c}^D)$, computed using Shapley, are identical.

3.2.3 Measures aligned with both DP and EO

Finally, we consider the measure in (5), and refer to it as $v_{\text{kh}}^D(X_S)$. Recall that $v_{\text{kh}}^D(X_S)$ satisfies both relaxed DP- and relaxed EO-independence. It also satisfies undesired Conditions 1, 2 since $Y \perp X_S$ implies $I(Y; X_S) = UI(Y; X_S\backslash A) + SI(Y; X_S, A) = 0$ and $A \perp Y$ implies $I(A; Y) = UI(Y; A\backslash X_S) + SI(Y; A, X_S) = 0$. We observe that by removing the quantity $SI(Y; A, X_S)$ (which vanishes when $Y \perp X_S$ or $A \perp Y$) from $v_{\text{kh}}^D(X_S)$, both undesired conditions could be avoided. This leads to our final candidate measure:

Candidate measure 5: $v_5^D(X_S) = I(A; X_S)I(A; X_S|Y)$ satisfies all desired properties met by v_{kh}^D, but violates Conditions 1, 2. We provide a summary of the measures' properties in Appendix 8.

Remark 2. *Through a synthetic-data experiment, Pelegrina et al. (2024) show that using Shapley value to quantify individual feature importance does not overestimate the shared information among features, compared to the non-coalition method from (Pelegrina et al., 2023). Thm. 3.4 theoretically validates this result by proving that Shapley aggregation eliminates the shared and complementary information, $SI(A; X_S, X_{S^c})$ and $CI(A; X_S, X_{S^c})$, from the deduced marginal impacts.*

Remark 3. *Marginal discriminatory impacts deduced using Shapley value applied to our measures do account for redundancy and synergy among features within X_S (or X_{S^c}). This is because our measures are defined for a subset X_S based on the joint distribution $p_{X_S, X_{S^c}, A, Y}$, yet without knowledge of interdependencies among features within X_S (nor X_{S^c}). For example, the measure may not differentiate between the presence or absence of duplicate instances of a particular feature within X_S. Kumar et al. (2020) demonstrated, through a toy example, that appending a single redundant feature to a 2-feature dataset alters deduced marginal feature importance scores. In Appendix 9, we modify this example to highlight the impact of redundancy on the correctness of deduced marginal discriminatory impacts, and further extend it to show that adding a large number of redundant features renders the marginal discriminatory impact dominated by redundant features.*

4 Experimental results: Synthetic data

We evaluate our discrimination measures using numerous synthetic datasets, generated via a parametric structural causal model (SCM) (Pearl, 2009), see Figure 2. Our SCM contains 6 observed variables (A, features $X_1 - X_4$, Y) and 2 latent confounders (Z_1, Z_2); $A \sim \text{Bern}(p_a)$, $Z_1 \sim \text{Gamma}(2, 1)$, and $Z_2 \sim \text{Uniform}([0, 2])$. In the corresponding causal diagram, A, Z_1, Z_2 are root nodes and Y is a sink node. Features $X_1 - X_4$, and soft label $\tilde{Y}$, are sampled using the structural equations:

$$X_i = \alpha_{a,i} w_{a,i} A + \sum\nolimits_{k \in [4]\backslash\{i\}} \alpha_{k,i} w_{k,i} X_k + \sum\nolimits_{j \in [2]} \alpha_{z_j,i} w_{z_j,i} Z_j + \varepsilon_i, \quad \text{for } i \in [4] \tag{10}$$

$$\tilde{Y} = \alpha_{a,y} w_{a,y} A + \sum\nolimits_{k \in [4]} \alpha_{k,y} w_{k,y} X_k + \sum\nolimits_{j \in [2]} \alpha_{z_j,y} w_{z_j,y} Z_j + \varepsilon_y, \tag{11}$$

where variables $\alpha \in \{0, 1\}$ are edge indicators, and variables $w \in [1, 2]$ are edge weights. $\varepsilon_i, \varepsilon_y \sim N(0, 1)$ are the exogenous Gaussian noise variables. We fix the causal order $\{1, 2, 3, 4\}$ among features in (10) to ensure the graph is acyclic. The binary target label Y is computed using the soft label $\tilde{Y}$ in (11) as $Y = \mathbb{1}((\tilde{Y} - \mathbb{E}[\tilde{Y}])/\sqrt{\text{Var}[\tilde{Y}]} > th_y)$ where $th_y \in \mathbb{R}$ is a threshold parameter.

We first sample numerous directed acyclic graphs (DAGs) according to the *parameteric* SCM in (10), (11), via sampling graph variables and setting their parameter values as follows. Edge indicators $\alpha_{\{a,k,z_j\},i} \sim \text{Bern}(p_d^{a,z,x})$, for $i, k \in [4]$, and $j \in [2]$, where $p_d^{a,z,x} \in \{0.2, 0.4, 0.6, 0.8\}$. $\alpha_{k,y} \sim \text{Bern}(p_d^{x,y})$, for $k \in [4]$, where $p_d^{x,y} \in \{0.5, 0.7, 0.9\}$. $\alpha_{z_j,y} \sim \text{Bern}(0.5)$, for $j \in [2]$. $\alpha_{a,y} \in \{0, 1\}$. $N_c \in \{0, 1, 2\}$ confounders: '0'= No confounders; '1'= only Z_1; and '2'= Z_1, Z_2. Demographic ratio $p_a \in \{0.1, 0.6\}$. Threshold $th_y \in \{0, 0.15, 0.3, 0.45, 0.6\}$. Edge weights $w \sim \text{Uniform}[1, 2]$, except $w_{a,y} = 2$. See Appendix 10 for more details about the selection of the parameters. For each possible combination of the parameters, we sampled three DAGs, resulting in a total of 2160 DAGs. For each realization of the DAG, we sample ten 100k-example datasets by sampling $A \sim \text{Bern}(p_a)$, $Z_1 \sim \text{Gamma}(2, 1)$, $Z_2 \sim \text{Uniform}([0, 2])$, $\varepsilon_i, \varepsilon_y \sim N(0, 1)$. Finally, we quantize the range of each feature X_i, for $i \in [4]$, into six bins of equal width.

For each DAG, we deduce the marginal discriminatory impacts, $\phi_i(v_j^D)$ for $i \in [4]$ and $j \in \{1, ..., 5, \text{kh}\}$, using the measures presented in Section 3.2, averaged over 3 out 10 datasets. We utilize the "dit Package" (James et al., 2018) for computing MI and the "Broja_2PID" Package (Makkeh et al., 2017; 2018) for computing PID components, which solves the optimization problem in (3) using ECOS solver (Domahidi et al., 2013). We upgrade the package with the MOSEK optimization suite in (ApS, 2019) to enhance calculation reliability. We also add a fault detection mechanism, see Appendix 11 for more details. We split each dataset into 67% train and 33% test datasets and train a 1-layer neural network (NN) with 3 hidden units. On the test set, we evaluate the discrimination bias of the predictor w.r.t to DP and EO as $b^{DP} \triangleq |p_{\hat{Y}|A=1}(1) - p_{\hat{Y}|A=0}(1)|$, and $b^{EO} \triangleq \mathbb{E}_{y \sim p_Y}[|p_{\hat{Y}|A=1,Y=y}(1) - p_{\hat{Y}|A=0,Y=y}(1)|]$.

We conduct an *ablation study* to validate the correctness of the deduced marginal discriminatory impacts based on our measures. For each dataset, we calculate the discrimination bias metrics with all features included, denoted by b_t^{DP} and b_t^{EO}, as well as when a feature i is removed (set to its mean value), denoted as b_i^{DP} and b_i^{EO}; where $i \in [n]$. We then compute the reduction of discrimination bias due to individual feature removal as: $d_i^{DP} \triangleq b_t^{DP} - b_i^{DP}$ and $d_i^{EO} \triangleq b_t^{EO} - b_i^{EO}$. To account for the varying scales of the marginal discriminatory impacts, $\phi_i(v_j^D)$, and the discrimination bias reductions, d_i^{DP} and d_i^{EO}, we compute a normalized version of each quantity as: $\bar{d}_i^{DP} \triangleq {}^{d_i^{DP}}/_{(max_i|d_i|)}$, $\bar{d}_i^{EO} \triangleq {}^{d_i^{EO}}/_{(\max_i|d_i|)}$, and $\bar{\phi}_i(v_j^D) \triangleq {}^{\phi_i(v_j^D)}/_{(\max_i|\phi_i(v_j^D)|)}$, where $\bar{d}_i^{DP}, \bar{d}_i^{EO}, \bar{\phi}_i(v_j^D) \in [-1, 1]$. We then compute the mean absolute error per feature for each DAG as: $e_j^{DP} = \frac{1}{8}\sum_i |\bar{\phi}_i(v_j^D) - \bar{d}_i^{DP}|$ and $e_j^{EO} = \frac{1}{8}\sum_i |\bar{\phi}_i(v_j^D) - \bar{d}_i^{EO}|$; $e_j^{DP}, e_j^{EO} \in [0, 1]$, averaged across 10 datasets. Similarly, we evaluate the correctness of the accuracy measure in (Khodadadian et al., 2021), and provide the results in Appendix 12.

For several DAGs (about 55%), feature removal affects the DP and EO discrimination bias of the predictor similarly. We utilize the following metric to assess a distinct behavior across the DAGs, w.r.t. DP and EO discrimination: $\Delta(DP, EO) = \frac{1}{8}\sum_i |\bar{d}_i^{DP} - \bar{d}_i^{EO}|$, averaged across 10 datasets. Subsequently, we divide the DAGs into two groups: (1) Distinct DP-EO DAGs, for which $\Delta(DP, EO) > 0.2$ (20% of full scale), with 964 out 2160 total DAGs. (2) Other DAGs, for which $\Delta(DP, EO) < 0.2$. Table 1 shows average measure errors for DP and EO across both DAG groups.

For distinct DP-EO DAGs, the DP-aligned measures, v_1^D, v_2^D, and v_3^D, outperform others in capturing the DP discrimination bias, with v_3^D having the lowest error. The EO-aligned measure v_4^D outperforms others in capturing the EO discrimination bias. The errors of measures v_5^D and v_{kh}^D, which satisfy both DP and EO properties, are higher than those of v_1^D, v_2^D, and v_3^D for DP, yet less than that of v_4^D. For EO, v_5^D and v_{kh}^D outperform v_1^D, v_2^D, v_3^D but not the EO-aligned measure v_4^D. For the other DAGs, v_3^D outperforms all other measures. Recall that v_3^D is the *only measure that is not monotonically non-decreasing* (violates Property 2), resulting in both positive and negative marginal discriminatory impacts. That is, v_3^D captures both increases and decreases in the predictor's discrimination bias due to feature removal. Across all datasets, the percentage of features that cause an increase in the discrimination bias when removed, i.e., replaced with their mean, is 32.1% for DP and 44.3% for EO. This supports our discussion in Section 3.1 that Property 2 is not essential. See Appendix 13 for the distribution of the measure errors for DP and EO; Appendix 14 for the measure errors for "equality of opportunity" fairness metric (Hardt et al., 2016); Appendix 15 for a comprehensive analysis of the relationship between the parameters of the data-generating model and the average measure errors, and Appendix 16 for guidelines for measure selection.

Measure	Distinct DP-EO DAGs		Other DAGs	
	DP	EO	DP	EO
$v_1^D(X_S)$	0.215556	0.379859	0.254301	0.302232
$v_2^D(X_S)$	0.211955	0.399146	0.276416	0.337613
$v_3^D(X_S)$	**0.189444**	0.373309	**0.231948**	**0.280861**
$v_4^D(X_S)$	0.242605	**0.362665**	0.254454	0.294174
$v_5^D(X_S)$	0.227579	0.367171	0.247409	0.286240
$v_{\text{kh}}^D(X_S)$	0.227591	0.369244	0.248619	0.288530

Table 1: Average measure errors in capturing marginal discriminatory impacts.

5 Experimental results: Real-world data

We evaluate our measures using 7 benchmark datasets. For each dataset, we specify the pair (Y, A). 3 datasets are derived from the U.S. Census American Community Survey (ACS) (Ding et al., 2021): ACS Income (income, race), ACS Public Coverage (public-health-coverage, race), and ACS Employment (employment-status, disability-status). The remaining datasets are: Adult (Kohavi & Becker, 1996) (income, gender), Census income (KDD) (Dua et al., 2017) (income, sex), ProPublica COMPAS (Larson et al., 2016) (recidivism-score, race), and Heritage health (Goldbloom & Hamner, 2011) (health-index, age). For each dataset, we select features correlated with A, prioritizing those with high MI with A and low conditional MI given Y (or vice versa); to differentiate feature impacts on discrimination bias w.r.t. DP and EO. Further, we quantize the range of each continuous feature into up to 10 bins of equal width. See Appendix 17 for more details about the datasets. For each dataset, we calculate the marginal discriminatory impacts of individual features using the measures in Section 3.2. Changes in discrimination bias of a downstream predictor are computed for DP and EO as in Section 4. We train 10 single-layer neural networks with different random seeds, optimizing hidden units (5–200) for accuracy. Average measure errors over the seeds for DP and EO are presented in Table 2. 5 datasets do not exhibit distinct DP-EO properties: ACS Income and Census Income are the only ones with clear distinctions. For ACS Income, DP-aligned measures have errors below 15% for DP, with v_3^D exhibiting the lowest error, despite high $UI(X_S; A\backslash Y)$ and $UI(A; X_S\backslash Y)$, accounting for approximately 60% of $I(A; X_S)$. No measures achieve errors below 40% for EO due to high redundancy, as $CSI_Y(A; X_{S_1}, X_{S_2})$ constitutes, on average, 49% of $I(A; X_{S_1}|Y)$ for $S_1, S_2 \subseteq [n]$, s.t., $S_1 \cap S_2 = \emptyset$. For Census Income (KDD), most measures show low errors for DP discrimination. Yet, the errors are larger than those in ACS Income, since $SI(A; X_{S_1}, X_{S_2})$ constitutes, on average, 40% of $I(A; X_{S_1})$ for $S_1, S_2 \subseteq [n]$, s.t., $S_1 \cap S_2 = \emptyset$. v_4^D has the lowest error for EO. Also, v_5^D and v_{kh}^D, which satisfy EO properties, have low errors. ASC Employment, Adult, and COMPAS datasets exhibited similar DP-EO properties. Notably, measure v_3^D demonstrates the lowest error on ACS Employment and COMPAS, capturing both increases and decreases in the predictor's discrimination bias, unlike the other measures. For Adult dataset, v_2^D exhibits the highest error for DP among the DP-aligned measures, since $UI(A; X_S\backslash Y) \approx I(A; X_S)$ for half the feature subsets. An analogous observation is noted for v_2^D on the COMPAS dataset. For ACS Public Coverage, the predictor achieves near-perfect DP and EO with $b_t^{DP} \approx b_t^{EO} \approx 0.5\%$; removal of two features causes an increase in the discrimination bias. Yet, none of the measures except v_3^D captures the increase in the discrimination bias. Notably, v_3^D has high error since it fails to capture the discriminatory impact of 'Age'. For most subsets that include this feature, $SI(X_S; A, Y)$ vanishes while $I(X_S; A) \approx UI(X_S; A\backslash Y)$. This significant unique information component could cause measure v_3^D to fail, as the predictor may inherit this information that is not captured by the measure. For the Health dataset, the predictor achieves marginal DP and EO violations, with $b_t^{DP} \approx b_t^{EO} \approx 4\%$. Non-decreasing measures fails to capture the slight decreases and the increases in discrimination bias, resulting in high measure errors. Measure v_3^D performs slightly better, but it is limited by the dataset's high redundancy. See Appendix 18 for additional results, and Appendix 19 for a conclusion of this work.

Measure	ACS Income		ACS Employment		ACS Public Coverage		Adult		Census Income (KDD)		COMPAS		Health	
	DP	EO	DP	EO	DP	EO	DP	EO	DP	EO	DP	EO	DP	EO
v_1^D	0.121	0.488	0.047	0.061	0.383	0.440	0.067	0.078	0.251	0.251	0.112	0.107	0.345	0.396
v_2^D	0.129	0.463	0.047	0.061	0.346	0.403	0.099	0.110	0.207	0.264	0.138	0.133	0.355	0.378
v_3^D	0.052	0.447	0.031	0.044	0.328	0.310	0.082	0.067	0.220	0.329	0.037	0.028	0.258	0.243
v_4^D	0.180	0.495	0.065	0.079	0.319	0.372	0.067	0.079	0.213	0.213	0.140	0.135	0.365	0.416
v_5^D	0.168	0.498	0.051	0.065	0.349	0.406	0.067	0.080	0.232	0.222	0.140	0.135	0.414	0.464
v_{kh}^D	0.164	0.492	0.044	0.058	0.333	0.391	0.067	0.080	0.230	0.220	0.140	0.135	0.418	0.469

Table 2: Average measure errors for DP and EO on real-world datasets.

REFERENCES

Hatoon AlSagri and Mourad Ykhlef. Quantifying feature importance for detecting depression using random forest. *the International Journal of Advanced Computer Science and Applications*, 11(5), 2020. Publisher: Science and Information (SAI) Organization Limited.

MOSEK ApS. Mosek optimization suite. 2019.

Pradeep Kr Banerjee, Johannes Rauh, and Guido Montúfar. Computing the unique information. In *the 2018 IEEE International Symposium on Information Theory*, pp. 141–145. IEEE, 2018.

Enrico Barbierato, Marco L Della Vedova, Daniele Tessera, Daniele Toti, and Nicola Vanoli. A methodology for controlling bias and fairness in synthetic data generation. *Applied Sciences*, 12 (9):4619, 2022. Publisher: MDPI.

Joachim Baumann, Alessandro Castelnovo, Riccardo Crupi, Nicole Inverardi, and Daniele Regoli. Bias on demand: A modelling framework that generates synthetic data with bias. In *the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1002–1013, 2023.

Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014. Publisher: Multidisciplinary Digital Publishing Institute.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *the 2009 IEEE International Conference on Data Mining Workshops*, pp. 13–18. IEEE, 2009.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. *the Annual Conference on Neural Information Processing Systems*, 30, 2017.

Shay B Cohen, Gideon Dror, and Eytan Ruppin. Feature selection based on the Shapley value. In *the International Joint Conference on Artificial Intelligence*, pp. 1–6, 2005.

T. M. Cover and Joy A. Thomas. *Elements of information theory.* Wiley-Interscience, 2006. ISBN 978-0-471-24195-9.

Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring Adult: New datasets for fair machine learning. In *the Annual Conference on Neural Information Processing Systems*, volume 34. Curran Associates, Inc., 2021.

Yash Diwate, Prashant Rana, and Pratik Chavan. Loan approval prediction using machine learning. *International Research Journal of Engineering and Technology*, 8(05):1741–1745, 2021.

Alexander Domahidi, Eric Chu, and Stephen Boyd. ECOS: An SOCP solver for embedded systems. In *the 2013 European Control Conference*, pp. 3071–3076. IEEE, 2013.

Dheeru Dua, Casey Graff, et al. UCI machine learning repository, 2017. *URL http://archive.ics.uci.edu/ml*, 7(1):62, 2017.

Sanghamitra Dutta, Praveen Venkatesh, Piotr Mardziel, Anupam Datta, and Pulkit Grover. An information-theoretic quantification of discrimination with exempt features. In *the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3825–3833, 2020.

Sanghamitra Dutta, Praveen Venkatesh, Piotr Mardziel, Anupam Datta, and Pulkit Grover. Fairness under feature exemptions: Counterfactual and observational measures. *IEEE Transactions on Information Theory*, 67(10):6675–6710, 2021. Publisher: IEEE.

Sanghamitra Dutta, Praveen Venkatesh, and Pulkit Grover. Quantifying feature contributions to overall disparity using information theory. *arXiv preprint arXiv:2206.08454*, 2022.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226, 2012.

Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric shapley values: Incorporating causal knowledge into model-agnostic explainability. In *the Annual Conference on Neural Information Processing Systems*, volume 33, 2020.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. *the Annual Conference on Neural Information Processing Systems*, 20, 2007.

Amirata Ghorbani and James Y. Zou. Neuron Shapley: Discovering the responsible neurons. In *the Annual Conference on Neural Information Processing Systems*, volume 33, pp. 5922–5932, 2020.

Anthony Goldbloom and Ben Hamner. Heritage health prize. `https://kaggle.com/competitions/hhp`, 2011. Kaggle.

Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In *the NIPS Symposium on Machine Learning and the Law*, volume 1, pp. 11. Barcelona, Spain, 2016. Issue: 2.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *the Annual Conference on Neural Information Processing Systems*, 29, 2016.

Ryan G James, Christopher J Ellison, and James P Crutchfield. "dit": A Python package for discrete information theory. *Journal of Open Source Software*, 3(25):738, 2018.

Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable AI: A causal problem. In *the International Conference on Artificial Intelligence and Statistics*, pp. 2907–2916. PMLR, 2020.

Lan Jiang, Clara Belitz, and Nigel Bosch. Synthetic dataset generation for fairer unfairness research. In *the 14th International Conference on Learning Analytics and Knowledge*, pp. 200–209, 2024.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012. Publisher: Springer.

Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *the 2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 643–650. IEEE, 2011.

Sajad Khodadadian, Mohamed Nafea, AmirEmad Ghassami, and Negar Kiyavash. Information theoretic measures for fairness-aware feature selection. *arXiv preprint arXiv:2106.00772*, 2021.

Newton Kinyanjui, Timothy Odonga, Celia Cintas, Noel Codella, Rameswar Panda, Prasanna Sattigeri, and Kush Varshney. Estimating skin tone and effects on classification performance in dermatology datasets. In *the Annual Conference on Neural Information Processing Systems*, 2019.

Kenji Kira and Larry A Rendell. The feature selection problem: Traditional methods and a new algorithm. In *the Tenth National Conference on Artificial Intelligence*, pp. 129–134, 1992.

Julia Kirchner, Surya Angwin, Jeff Mattu, and Lauren Larson. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. *Pro Publica: New York, NY, USA*, 2016.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *the 8th Innovations in Theoretical Computer Science Conference*. Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 2017.

Ronny Kohavi and Barry Becker. Adult data set. *UCI Machine Learning Repository*, 5:2093, 1996.

I Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with Shapley-value-based explanations as feature importance measures. In *the International Conference on Machine Learning*, pp. 5491–5500. PMLR, 2020.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *the Annual Conference on Neural Information Processing Systems*, 30, 2017.

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. ProPublica Compas Analysis—Data and Analysis for 'Machine Bias.'. *https://github. com/propublica/compas-analysis*, 2016.

Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. In *the Annual Conference on Neural Information Processing Systems*, volume 32, 2019.

Abdullah Makkeh, Dirk Oliver Theis, and Raul Vicente. Bivariate partial information decomposition: The optimization perspective. *Entropy*, 19(10):530, 2017. Publisher: MDPI.

Abdullah Makkeh, Dirk Oliver Theis, and Raul Vicente. BROJA-2PID: A robust estimator for bivariate partial information decomposition. *Entropy*, 20(4):271, 2018. Publisher: MDPI.

Masayoshi Mase, Art B Owen, and Benjamin B Seiler. Cohort Shapley value for algorithmic fairness. *arXiv preprint arXiv:2105.07168*, 2021.

Cecilia Munoz, Megan Smith, and DJ Patil. Big data: A report on algorithmic systems, opportunity, and civil rights. Executive Office of the President, USA, 2016.

Jakob Mökander, Prathm Juneja, David S Watson, and Luciano Floridi. The US Algorithmic Accountability Act of 2022 vs. The EU Artificial Intelligence Act: What can they learn from each other? *Minds and Machines*, 32(4):751–758, 2022. Publisher: Springer.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Guilherme Dean Pelegrina, Miguel Couceiro, and Leonardo Tomazeli Duarte. A statistical approach to detect disparity prone features in a group fairness setting. *AI and Ethics*, pp. 1–14, 2023. Publisher: Springer.

Guilherme Dean Pelegrina, Miguel Couceiro, and Leonardo Tomazeli Duarte. A preprocessing Shapley value-based approach to detect relevant and disparity prone features in machine learning. In *the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 279–289, 2024.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: Foundations and learning algorithms*. The MIT Press, 2017.

Felix Petersen, Debarghya Mukherjee, Yuekai Sun, and Mikhail Yurochkin. Post-processing for individual fairness. In *the Annual Conference on Neural Information Processing Systems*, volume 34, pp. 25944–25955, 2021.

Johannes Rauh, Nils Bertschinger, Eckehard Olbrich, and Jürgen Jost. Reconsidering unique information: Towards a multivariate information decomposition. In *the 2014 IEEE International Symposium on Information Theory*, pp. 2232–2236. IEEE, 2014.

M Scott, Su-In Lee, et al. A unified approach to interpreting model predictions. In *the Annual Conference on Neural Information Processing Systems*, volume 30, 2017.

Lloyd S Shapley. A value for n-person games. *Contribution to the Theory of Games*, 2, 1953.

Bingyang Wen, Luis Oliveros Colon, KP Subbalakshmi, and Rajarathnam Chandramouli. Causal-TGAN: Generating tabular data using causal generative adversarial networks. *arXiv preprint arXiv:2104.10680*, 2021.

Ruicheng Xian, Lang Yin, and Han Zhao. Fair and optimal classification via post-processing. In *the International Conference on Machine Learning*, pp. 37977–38012. PMLR, 2023.

Howard Yang and John Moody. Data visualization and feature selection: New algorithms for non-Gaussian data. In *the Annual Conference on Neural Information Processing Systems*, volume 12, 1999.

H Peyton Young. Monotonic solutions of cooperative games. *International Journal of Game Theory*, 14(2):65–72, 1985. Publisher: Springer.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *the 26th International Conference on World Wide Web*, pp. 1171–1180, 2017.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *the International Conference on Machine Learning*, 2013.

6 PROOF OF LEMMA 3.2

Lemma 3.2 states that: For a subset of features X_S, binary target label Y and binary sensitive attribute A, such that $X_S \perp A$ and $X_S \perp A|Y$, it follows that either (1) $X_S \perp Y$, i.e., X_S is irrelevant to the prediction task, or (2) $A \perp Y$, i.e., the base rates are equal across demographic groups.

Consider the conditional probability distributions $p_{X_S|A=a}(x)$ and $p_{Y|A=a}(y)$, where $a, y \in \{0, 1\}$ and $x \in \mathcal{X}_S$, with $\mathcal{X}_S$ representing the sample space of X_S. Using the law of total probability,

$$p_{X_S|A=a}(x) = p_{X_S|Y=1,A=a}(x)p_{Y|A=a}(1) + p_{X_S|Y=0,A=a}(x)p_{Y|A=a}(0).$$

Since $X_S \perp A|Y$, $p_{X_S|A=a}(x)$ can be written as

$$p_{X_S|A=a}(x) = p_{X_S|Y=1}(x)p_{Y|A=a}(1) + p_{X_S|Y=0}(x)p_{Y|A=a}(0),$$

On the other hand, given $X_S \perp A$, we have $p_{X_S|A=0}(x) = p_{X_S|A=1}(x)$. Hence, substituting in this equality the expression of $p_{X_S|A=a}(x)$, we get

$$p_{X_S|Y=1}(x)p_{Y|A=0}(1)+p_{X_S|Y=0}(x)p_{Y|A=0}(0) = p_{X_S|Y=1}(x)p_{Y|A=1}(1)+p_{X_S|Y=0}(x)p_{Y|A=1}(0).$$

Rearranging the equation, we get

$$p_{X_S|Y=1}(x)\left(p_{Y|A=0}(1) - p_{Y|A=1}(1)\right) + p_{X_S|Y=0}(x)\left(p_{Y|A=0}(0) - p_{Y|A=1}(0)\right) = 0.$$

Note that $p_{Y|A=0}(0) = 1 - p_{Y|A=0}(1)$ and $p_{Y|A=1}(0) = 1 - p_{Y|A=1}(1)$, hence we have

$$\left(p_{X_S|Y=1}(x) - p_{X_S|Y=0}(x)\right)\left(p_{Y|A=0}(1) - p_{Y|A=1}(1)\right) = 0,$$

which implies either $p_{X_S|Y=1}(x) = p_{X_S|Y=0}(x)$, i.e., $X_S \perp Y$, or $p_{Y|A=0}(1) = p_{Y|A=1}(1)$, i.e., $A \perp Y$.

7 PROOF OF THM. 3.4

Thm. 3.4 states that: Let $v_{SI}(X_S) = SI(A; X_S, X_{S^c})$, $v_{CI}(X_S) = CI(A; X_S, X_{S^c})$, $v_{CSI}(X_S) = CSI_Y(A; X_S, X_{S^c})$, $v_{CCI}(X_S) = CCI_Y(A; X_S, X_{S^c})$ be pay-off functions. The marginal contributions $\phi_i(v_{SI})$, $\phi_i(v_{CI})$, $\phi_i(v_{CSI})$, $\phi_i(v_{CCI})$, for all $i \in [n]$, using (4), are equal to zero.

We show that $\phi_i(v_{SI}) = 0$ for all $i \in [n]$. The proofs for the other pay-off functions follow similarly.

According to (4), $\phi_i(v_{SI})$ is given by

$$\phi_i(v_{SI}) = \sum_{S \subseteq [n]\setminus\{i\}} \omega(S)\left[v_{SI}(X_{S\cup\{i\}}) - v_{SI}(X_S)\right], \qquad \text{where} \quad \omega(S) \triangleq \frac{(n-1-|S|)!|S|!}{n!}. \tag{12}$$

Note that $n - 1 - |S| = |S^c\setminus\{i\}|$, hence we have $\omega(S) = \omega(S^c\setminus\{i\})$. Further, the sum index in (12) could be equivalently replaced by $S^c\setminus\{i\} \in [n]\setminus\{i\}$. Therefore, (12) could be rewritten as

$$\phi_i(v_{SI}) = \sum_{S^c\setminus\{i\} \subseteq [n]\setminus\{i\}} \omega(S^c)\left[v_{SI}(X_{S\cup\{i\}}) - v_{SI}(X_S)\right]. \tag{13}$$

Furthermore, we have

$$\begin{aligned} v_{SI}(X_{S\cup\{i\}}) - v_{SI}(X_S) &= -\left[SI(A; X_{S^c}, X_S) - SI(A; X_{S^c\setminus\{i\}}, X_{S\cup\{i\}})\right] \\ &= -\left[v_{SI}(X_{(S^c\setminus\{i\})\cup\{i\}}) - v_{SI}(X_{S^c\setminus\{i\}})\right]. \end{aligned}$$

Thus, substituting in (13), we can rewrite the Shapley value $\phi_i(v_{SI})$ as

$$\phi_i(v_{SI}) = -\sum_{S^c\setminus\{i\} \subseteq [n]\setminus\{i\}} \omega(S^c\setminus\{i\})\left[v_{SI}(X_{(S^c\setminus\{i\})\cup\{i\}}) - v_{SI}(X_{S^c\setminus\{i\}})\right] = -\phi_i(v_{SI}),$$

which implies that $\phi_i(v_{SI}) = 0$, and completes the proof.

8 SUMMARY OF THE MEASURES' PROPERTIES

Measure	Natural Properties		DP-Properties		EO-Properties		Conditions	
	Non. neg.	Non. dec.	DP indep.	DP blocking	EO indep.	EO blocking	1	2
v_1^D	Yes	Yes	Yes	No	No	No	No	No
$v_{1,u}^D$	Yes	Yes	Yes*	Yes	No	No	No	No
$v_{1,c}^D$	Yes	Yes	No	Yes	No	No	No	No
v_2^D	Yes	Yes	Yes*	No	No	No	No	Yes
v_3^D	Yes	No	Yes*	No	No	No	Yes	No
v_4^D	Yes	Yes	No	No	Yes	No	No	No
$v_{4,u}^D$	Yes	Yes	No	No	Yes*	Yes	No	No
$v_{4,c}^D$	Yes	Yes	No	No	No	Yes	No	No
v_5^D	Yes	Yes	Yes*	No	Yes*	No	No	No
v_{kh}^D	Yes	Yes	Yes*	No	Yes*	No	Yes	Yes

Table 3: Properties satisfied by our measures. (Yes*) indicates that the measure satisfies the relaxed version of the property.

9 EFFECT OF REDUNDANT FEATURES ON THE FEATURES' MARGINAL DISCRIMINATORY IMPACTS

Here, we present an example (adapted from (Kumar et al., 2020)) to demonstrate the effect of feature redundancy (within X_S) on the correctness of marginal discriminatory impacts of individual features, using Shapley value aggregation. To elaborate, marginal impacts are deduced using measures, defined for subsets X_S, based on the joint distribution $p_{X_S,X_{S^c},A,Y}$, yet without knowledge of interdependencies among features within X_S (nor X_{S^c}). Note that, the example in (Kumar et al., 2020) focuses on the effect of within-coalition feature redundancy on marginal features importance, computed using Shapley value. We adapt the example to focus on marginal discriminatory impacts. Additionally, we extend the analysis to examine the effect of adding a large number of redundant features.

Consider a dataset containing two features, X_1 and X_2, a target label, Y, and a sensitive attribute, A. Let $v^D(X_S)$ be a discrimination measure that is defined for a subset X_S based on the joint distribution $p_{X_S,X_{S^c},A,Y}$. Using (4), the marginal discriminatory impacts of individual features are given by

$$\phi_1(v^D) = \frac{1}{2}\left(v^D(X_{\{1,2\}}) - v^D(X_{\{2\}})\right) + \frac{1}{2}v^D(X_{\{1\}}),$$
$$\phi_2(v^D) = \frac{1}{2}\left(v^D(X_{\{1,2\}}) - v^D(X_{\{1\}})\right) + \frac{1}{2}v^D(X_{\{2\}}).$$

Consider adding a redundant feature X_3, a copy of X_2, to the dataset. Adding a copy of X_2 shall not change the characteristics of the learnt predictor. Further, the discriminatory impact of X_1 should not change since removing X_1 from the sets $X_{\{1,2\}}$ or $X_{\{1,2,3\}}$ yields identical information about A (provided by X_2). However, by adding X_3 to the original dataset and computing the marginal discriminatory impacts of X_1, X_2, and X_3, we get

$$\phi_1'(v^D) = \frac{2}{3}\left(v^D(X_{\{1,2\}}) - v^D(X_{\{2\}})\right) + \frac{1}{3}v^D(X_{\{1\}}),$$
$$\phi_2'(v^D) = \phi_3'(v^D) = \frac{1}{6}\left(v^D(X_{\{1,2\}}) - v^D(X_{\{1\}})\right) + \frac{1}{3}v^D(X_{\{2\}}),$$

since $v^D(X_{\{1,2,3\}}) = v^D(X_{\{1,2\}}) = v^D(X_{\{1,3\}})$, and $v^D(X_{\{2,3\}}) = v^D(X_{\{2\}}) = v^D(X_{\{3\}})$. Note that $\phi_1'(v^D)$ does not equal the marginal discriminatory impact of X_1 in the 2-features setting, i.e., $\phi_1(v^D)$. The redundant feature X_3 caused the marginal discriminatory impact of X_1 to be more reliant on the contribution of X_1 to the subset $X_{\{2\}}$, i.e., $\left(v^D(X_{\{1,2\}}) - v^D(X_{\{2\}})\right)$ rather than on $v_D(X_{\{1\}})$. Moreover, the marginal discriminatory impact of X_2 in the 2-features, i.e., $\phi_2(v^D)$, setting does not equal to $\phi_2'(v^D)$ nor the sum $\phi_2'(v^D) + \phi_3'(v^D)$. This demonstrates that the redundant

feature X_3 alters the deduced marginal discriminatory impacts of features X_1 and X_2, leading to results that contradict expectations.

Next, we extend the previous discussion to show that adding a large number of redundant features (copies of X_2) renders the marginal discriminatory impact of X_1 unrelated to $v^D(X_{\{1\}})$, while uniformly distributing $v(X_{\{2\}})$ over the copies of X_2, for their marginal discriminatory impacts.

Consider a dataset that contains n features $X_1, X_2, ..., X_n$, such that $X_2 = X_3 = \cdots = X_n$. The marginal discriminatory impact of X_1 is given by

$$\begin{aligned}
\phi_1(v^D) &= \sum_{S\subseteq[n]\backslash\{1\}} \frac{(n-1-|S|)!|S|!}{n!}\left[v^D(X_{S\cup\{1\}}) - v^D(X_S)\right] \\
&= \frac{1}{n}v^D(X_{\{1\}}) + \sum_{S\subseteq[n]\backslash\{1\},S\neq\emptyset} \frac{(n-1-|S|)!|S|!}{n!}\left[v^D(X_{S\cup\{1\}}) - v^D(X_S)\right] \\
&\overset{\text{(a)}}{=} \frac{1}{n}v^D(X_{\{1\}}) + \sum_{S\subseteq[n]\backslash\{1\},S\neq\emptyset} \frac{(n-1-|S|)!|S|!}{n!}\left[v^D(X_{\{1,2\}}) - v^D(X_{\{2\}})\right] \\
&= \frac{1}{n}v^D(X_{\{1\}}) + \left(v^D(X_{\{1,2\}}) - v^D(X_{\{2\}})\right) \sum_{S\subseteq[n]\backslash\{1\},S\neq\emptyset} \frac{(n-1-|S|)!|S|!}{n!} \\
&\overset{\text{(b)}}{=} \frac{1}{n}v^D(X_{\{1\}}) + \frac{n-1}{n}\left(v^D(X_{\{1,2\}}) - v^D(X_{\{2\}})\right) \\
&= \left(v^D(X_{\{1,2\}}) - v^D(X_{\{2\}})\right), \text{ as } n\to\infty
\end{aligned}$$

where (a) follows since the features in any non-empty subset $S \in [n]\backslash\{1\}$ are copies of X_2, and hence $v^D(X_S) = v^D(X_{\{2\}})$ and $v^D(X_{S\cup\{1\}}) = v^D(X_{\{1,2\}})$. (b) follows since

$$\sum_{S\subseteq[n]\backslash\{1\},S\neq\emptyset} \frac{(n-1-|S|)!|S|!}{n!} = \sum_{j=1}^{n-1} \sum_{S\subseteq[n]\backslash\{1\},|S|=j} \frac{(n-1-|S|)!|S|!}{n!} = \sum_{j=1}^{n-1} \frac{1}{n} = \frac{n-1}{n}.$$

The marginal discriminatory impact of the individual features X_i for $i \in [n]\backslash\{1\}$, are

$$\begin{aligned}
\phi_i(v^D) &= \sum_{S\subseteq[n]\backslash\{i\}} \frac{(n-1-|S|)!|S|!}{n!}\left[v^D(X_{S\cup\{i\}}) - v^D(X_S)\right] \\
&= \frac{1}{n}v^D(X_{\{i\}}) + \sum_{S\subseteq[n]\backslash\{i\},S\neq\emptyset} \frac{(n-1-|S|)!|S|!}{n!}\left[v^D(X_{S\cup\{i\}}) - v^D(X_S)\right] \\
&\overset{\text{(a)}}{=} \frac{1}{n}v^D(X_{\{2\}}) + \sum_{S=\{1\}} \frac{(n-1-|S|)!|S|!}{n!}\left[v^D(X_{\{1,2\}}) - v^D(X_{\{1\}})\right] \\
&= \frac{1}{n}v^D(X_{\{2\}}) + \frac{1}{n(n-1)}\left(v^D(X_{\{1,2\}}) - v^D(X_{\{1\}})\right)
\end{aligned}$$

where (a) follows because: (i) $v^D(X_{\{i\}}) = v^D(X_{\{2\}})$ since X_i is copy of X_2 (ii) For any feature X_j, such that $j \neq i$ and $j \neq 1$, we have $v^D(X_{\{i,j\}}) = v^D(X_{\{2\}})$ since X_i and X_j are copies of X_2; (iii) For any $S \subseteq [n]\backslash\{i\}$, such $|S| > 1$, we have either $1 \in S$, hence $v^D(X_{S\cup\{i\}}) = v^D(X_S) = v^D(X_{\{1,2\}})$, or $1 \notin S$, hence $v^D(X_{S\cup\{i\}}) = v^D(X_S) = v^D(X_{\{2\}})$. Further, the ratio of $\frac{1}{n(n-1)}\left(v^D(X_{1,2}) - v^D(X_1)\right)$ to $\frac{1}{n}v^D(X_2)$ approaches 0, when n approaches ∞, making the former negligible compared to the latter. Moreover, the ratio of $\frac{1}{n-1}v^D(X_{\{2\}})$ to $\frac{1}{n}v^D(X_{\{2\}})$ approaches 1, when n approaches ∞, making the former an approximation of the latter. Hence, $\phi_i(v^D)$ can be approximated by $\frac{1}{n-1}v^D(X_{\{2\}})$ for significantly large n.

To sum up, adding a large number of redundant features causes $\phi_1(v^D)$ to be equal to the contribution of X_1 to the subset $X_{\{2\}}$, i.e., $\left(v^D(X_{\{1,2\}}) - v^D(X_{\{2\}})\right)$, neglecting for the measure of singleton subset $X_{\{1\}}$. On the other hand, the marginal discriminatory impacts of X_2 or any of the redundant features (copies of X_2) can be approximated by distributing the measure of $X_{\{2\}}$ over all the redundant features including X_2 itself.

10 SELECTION OF THE DATA-GENERATING MODEL PARAMETERS

Here, we detail the parameters used to generate a causal directed graph defining the data-generating model for each synthetic dataset in Section 4. Each graph is generated by sampling graph variables and setting their parameter values as follows:

- The edge indicators $\alpha_{a,i}$, $\alpha_{k,i}$, and $\alpha_{z_j,i}$, for $i, k \in [4]$ and $j \in [2]$, are sampled from a Bern($p_d^{a,z,x}$) distribution, where $p_d^{a,z,x}$ represents the density of the causal connections from $\{A, Z_1, Z_2\}$ to X^n, and among the features in X^n. Here, $p_d^{a,z,x}$ is selected from $\{0.2, 0.4, 0.6, 0.8\}$, reflecting various levels of feature dependency on the sensitive attribute and unobserved variables (when exist), as well as the interdependencies among features.
- The edge indicators $\alpha_{k,y}$, for $k \in [4]$, are sampled from a Bern($p_d^{x,y}$) distribution, where $p_d^{x,y}$ represents the density of the causal connections from X^n to Y. We select $p_d^{x,y}$ from $\{0.5, 0.7, 0.9\}$, avoiding lower density values to ensure the target label maintains substantial correlation with dataset features.
- The edge indicators $\alpha_{z_j,y}$, for $j \in [2]$, are sampled from a Bern(0.5) distribution, reflecting equal likelihood of confounding, versus not confounding, the target label in the sampled graphs.
- $\alpha_{a,y}$ is set to either 0 or 1 to distinctively analyze scenarios with and without a direct causal link from A to Y.
- We consider a number of confounders $N_c \in \{0, 1, 2\}$, where '0' means 'No confounders', '1' means 'only Z_1', and '2' means 'both Z_1 and Z_2'. This introduces various levels of complexity of interdependencies among dataset variables due to the presence of hidden variables.
- Demographic ratio p_a is selected from $\{0.1, 0.6\}$ to simulate distinct demographic group distribution.
- Threshold th_y is selected from $\{0, 0.15, 0.3, 0.45, 0.6\}$ to simulate distinct prediction tasks.
- All edge weights w are sampled from a $\mathrm{Uniform}[1, 2]$ distribution, except for $w_{a,y}$, which is always set to 2.

11 CALCULATION OF THE PID COMPONENTS

This section outlines our approach to ensure accuracy and consistency of PID component calculation, particularly concerning the optimization problem defined in (3). We discuss several practical methods for solving this problem and provide brief insights into their applicability to our specific context. Next, we detail the solution we select and how we enhance its reliability. In the following, we highlight three solvers for the optimization problem in (3):

- "dit Package" (https://github.com/dit/dit): This package computes the unique information component by optimizing the conditional mutual information under marginal constraints using the global optimization technique "basin-hopping" (James et al., 2018). Initial attempts with this solver reveal impractical computation time when the number of random variables or their support sizes increase. This inefficiency is evident even when computing the discrimination measures for the synthetic datasets, which include only four features, each having a support cardinality of 6.
- "ComputeUI Package" (https://github.com/infodeco/computeUI): This package solves the optimization problem (3) using an alternating divergence minimization algorithm with guaranteed convergence (Banerjee et al., 2018). Although this package demonstrates slightly faster computation compared to the "dit Package", it is inadequate when the number of random variables or their cardinality increases.
- "BROJA_2PID Package" (https://github.com/Abzinger/BROJA_2PID): This package solves the optimization problem by solving an equivalent exponential cone program, that satisfies strong duality property (Makkeh et al., 2017; 2018). It demonstrates significantly fast computation of the PID components. The computation time for any of our discrimination measures that include PID components is only tens of seconds compared to more than

an hour using the "dit Package" or the "ComputeUI Package". We select this package to calculate the PID components.

The available "BROJA_2PID Package" utilizes a Conic Optimization software toolbox ECOS, which solves Exponential Cone Programs (Domahidi et al., 2013). As reported in (Makkeh et al., 2017, Section 4.2), ECOS solver encounters numerical errors in rare instances when solving the optimization problem in (3). To address this, we implement the solution proposed in (Makkeh et al., 2017, Section 5), which combines the ECOS solver with the MOSEK optimization suite (ApS, 2019). In this approach, ECOS serves as a backup solver when MOSEK fails to provide a solution, ensuring robust performance across various scenarios. Therefore, in our implementation, we integrate the MOSEK optimization suite in the "BROJA_2PID Package". This is implemented using the optimization modeling tool CVXPY (Diamond & Boyd, 2016), as an API for the MOSEK suite. To compare the solution of the two solvers, using synthetic datasets (cf. Section 4), we calculate the ratio of the absolute differences between the PID components, computed using the ECOS solver and the MOSEK suite, to the total mutual information (1). In some rare cases, the ratio reaches more than 5% due to the failure of one of the solvers by numerical issues or the sub-optimality of the ECOS solver solution. The failure of both solvers rarely happens. For the synthetic datasets experiment, we overcome this dual failure by regenerating the synthetic dataset with a new random seed number. The failure of both solvers does not happen in the real-world datasets experiment. Finally, to further ensure the reliability of our results, we implement a robust fault detection mechanism for the PID component computation, by verifying the monotonicity property (Property 2) of the discrimination measures (cf. Section 3.2). Since measure $v_3^D(X_S) = SI(X_S; A, Y)$ does not satisfy Property 2, we instead verbifies the monotonicity property of $UI(X_S; A \backslash Y)$. The violation of the monotonicity property rarely occurs, which is exclusive for subsets of features with relatively small measure values, where "relatively" is understood in comparison to the measured values of other subsets of features in the same dataset.

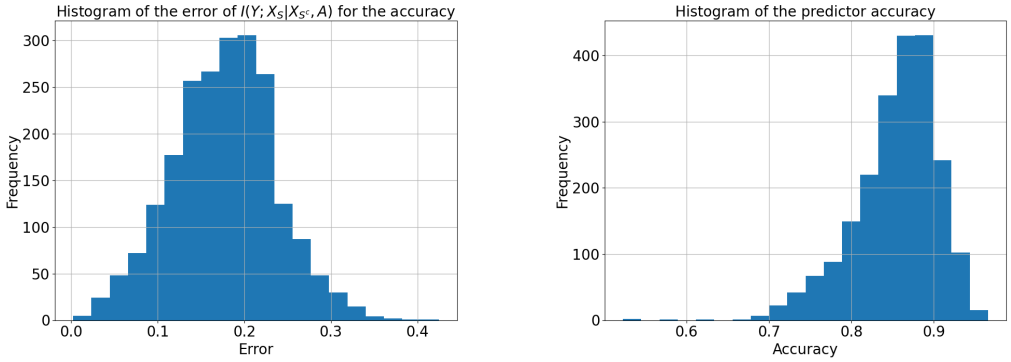
12 Experimental results of synthetic data for the accuracy measure

Here, we elaborate on the accuracy measure proposed by Khodadadian et al. (2021), revisiting its key properties, and presenting experimental results obtained using both synthetic and real datasets (cf. Sections 4 and 5). Khodadadian et al. (2021) proposed the following accuracy measure for a subset of feature X_S:

$$v^A(X_S) = I(Y; X_S|(A, X_{S^c}))$$

This measure satisfies the following desired properties. (1) *Non-negative*: $v^A(X_S) \geq 0$ with equality if $S = \emptyset$. (2) *Non-decreasing*: $v^A(X_{S_1}) \leq v^A(X_{S_2})$, for any $S_1, S_2 \subseteq [n]$, such that $S_1 \subseteq S_2$. (3) *Blocking:* if $Y \perp X_S|(A, X_{S^c})$, then $v^A(X_S) = 0$. Importantly, evaluating v^A does not require access to predictions $\hat{Y}$ of a predefined learning model, hence v^A is directly applicable to our framework.

The predictor accuracy is evaluated, based on the zero-one loss, as $acc = p_{X^n,A,Y}(Y = \hat{Y})$. Using the metric acc, we follow the same ablation study introduced for the discrimination measures in Section 4 to validate the correctness of deduced marginal accuracy impacts for measure v^A. The average accuracy measure error evaluated on the synthetic datasets (cf. Section 4) is 0.177. The distribution of the measure error is shown in Figure 3 (a). Additionally, we show the distribution of predictor accuracy for all the synthetic datasets in Figure 3 (b). In Table 4, we provide the accuracy measure error evaluated on the real-world datasets (cf. Section 5).



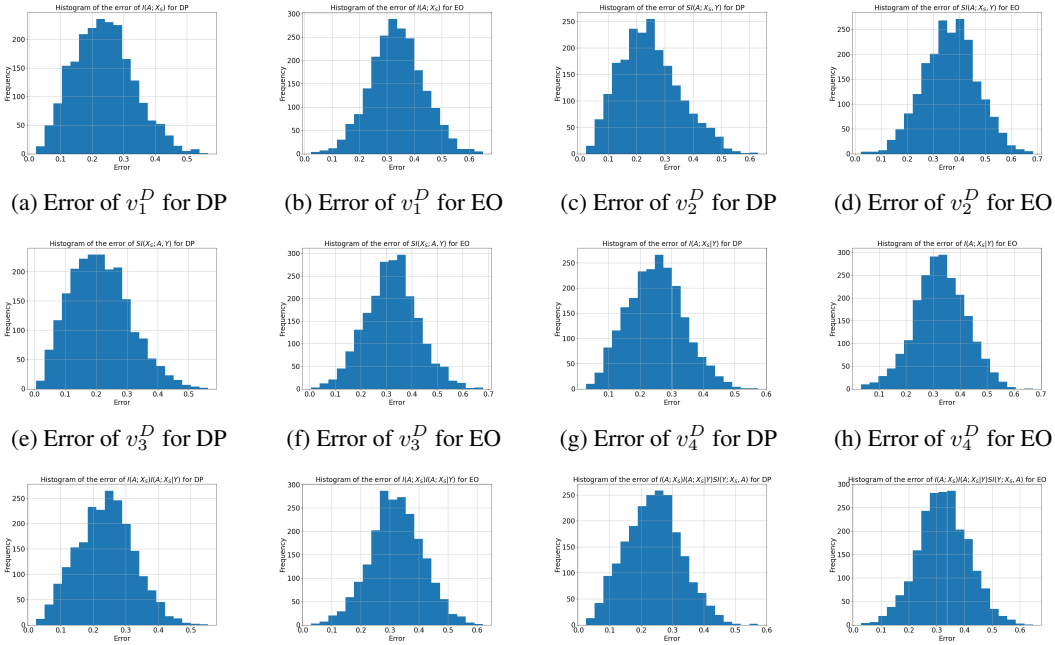
(a) (b)

Figure 3: (a) Distribution of the accuracy measure error. (b) Distribution of the predictor accuracy.

Measure	ACS Income	ACS Employment	ACS Public Coverage	Adult	Census income (KDD)	COMPAS	Health
$I(Y; X_S\|X_{S^c}, A)$	0.0704	0.008	0.3433	0.066	0.1329	0.0714	0.1875

Table 4: Average accuracy measure error on real-world datasets.

13 DISTRIBUTION OF THE MEASURE ERRORS



(a) Error of v_1^D for DP (b) Error of v_1^D for EO (c) Error of v_2^D for DP (d) Error of v_2^D for EO

(e) Error of v_3^D for DP (f) Error of v_3^D for EO (g) Error of v_4^D for DP (h) Error of v_4^D for EO

(i) Error of v_5^D for DP (j) Error of v_5^D for EO (k) Error of v_{kh}^D for DP (l) Error of v_{kh}^D for EO

Figure 4: Distribution of the measure errors for DP and EO.

14 EXPERIMENTAL RESULTS OF SYNTHETIC DATA FOR EQUALITY OF OPPORTUNITY

This appendix extends the experimental results of the synthetic datasets (cf. Section 4) to incorporate the fairness notion of equality of opportunity (EOP) Hardt et al. (2016). EOP requires matching error rates for the "advantaged" outcome class across demographic groups. Formally, a classifier $\hat{Y} = h(X^n)$ satisfies this criterion:

- For $Y = 0$ as the advantaged outcome if $p_{\hat{Y}|A=a,Y=0} = p_{\hat{Y}|Y=0}$, for $a \in \mathcal{A}$, i.e., $\hat{Y}$ and A are independent given $Y = 0$.
- For $Y = 1$ as the advantaged outcome if $p_{\hat{Y}|A,Y=1} = p_{\hat{Y}|Y=1}$, for $a \in \mathcal{A}$, i.e., $\hat{Y}$ and A are independent given $Y = 1$.

We compute the discrimination bias of the predictor with respect to EOP as follows:

$$b^{EOP,0} \triangleq |p_{\hat{Y}|A=1,Y=0}(1) - p_{\hat{Y}|A=0,Y=0}(1)|, \quad \text{for } Y = 0 \text{ as the advantaged outcome;} \tag{14}$$

$$b^{EOP,1} \triangleq |p_{\hat{Y}|A=1,Y=1}(1) - p_{\hat{Y}|A=0,Y=1}(1)|, \quad \text{for } Y = 1 \text{ as the advantaged outcome;} \tag{15}$$

Table 5 shows the average measure errors with respect to EOP evaluated on the synthetic datasets. See Section 4 for the definition of distinct DP-EO DAGs and other DAGs.

For the DAGs DP-EO datasets, v_4^D has the lowest error when the advantaged outcome is $Y = 0$ or $Y = 1$, similar to what we have for EO in Table 1. For the other datasets, v_3^D has the lowest error for $Y = 0$, while v_5^D exhibits the lowest error for $Y = 1$. Notably, our choice of values for the threshold th_y yields positive rates $P(Y = 1)$ ranging from 20% to 50%. Hence, a larger portion of the data has $Y = 0$. Consequently, measure errors are smaller when $Y = 0$ is the advantaged outcome. This also causes v_4^D to align better with EOP with advantaged group $Y = 0$ than $Y = 1$ since v_4^D is computed as a weighted average over both target label classes.

Measure	Distinct DP-EO DAGs		Other DAGs	
	EOP ($Y = 0$)	EOP ($Y = 1$)	EOP ($Y = 0$)	EOP ($Y = 1$)
$v_1^D(X_S)$	0.329200	0.409190	0.289028	0.353112
$v_2^D(X_S)$	0.338270	0.431333	0.313405	0.397409
$v_3^D(X_S)$	0.323885	0.404792	**0.269754**	0.338302
$v_4^D(X_S)$	**0.321347**	**0.391285**	0.285868	0.336438
$v_5^D(X_S)$	0.327359	0.392154	0.278960	**0.326351**
$v_{\text{kh}}^D(X_S)$	0.328872	0.394083	0.279710	0.328981

Table 5: Average measure errors in capturing marginal discriminatory impacts for EOP.

15 A VISUALIZING THE RELATIONSHIP BETWEEN DATA-GENERATING MODEL PARAMETERS AND MEASURE ERRORS

This appendix provides visual representations illustrating the relationship between data-generating model parameters and measures' errors.

Effect of predictor accuracy on the increase/decrease of discrimination bias due to feature removal. In Figure 5, we demonstrate the relation between the predictor accuracy and the percentage of features that cause an increase in the discrimination bias when removed. DAGs are split into four equal groups based on their respective average predictor accuracy. Predictor accuracy ranges from 70% to 94.5% for 99% of the DAGs, see Appendix 12 for the exact distribution. Figure 5 shows that, for DP, predictor accuracy does not influence the number of features causing an increase in the

discrimination bias. However, for EO, higher predictor accuracy causes more features to increase discrimination bias when removed. This occurs since high-accuracy models are fairer with respect to EO. Consequently, removing features with high accuracy impact reduces predictor accuracy while increasing its discrimination with respect to EO.

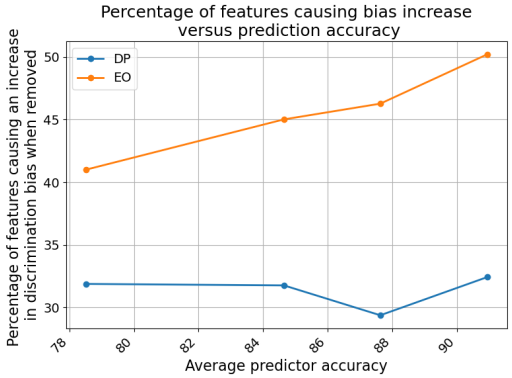


Figure 5: Average percentage of features causing increase in the discrimination bias when removed versus average accuracy of the predictor.

Effect of graph density and number of confounders on measure errors. Figure 6 shows the average measure errors for DP and EO versus the number of confounders N_c and the density of the causal links $p_d^{a,z,x}$. Increasing $p_d^{a,z,x}$ while fixing N_c increases the average measure error for both DP and EO. Lower edge density increases the likelihood of "independence or blocking" properties being satisfied (Properties 3, 4, 5, and 6), promoting better alignment with our theoretical framework. For DP, increasing data complexity by adding confounders has a similar effect to increasing edge density. For EO, adding confounders reduces measure errors since adding confounders reduces predictor accuracy as shown in Figure 7. In other words, confounders reduce the chance of features causing an increase in the discrimination bias when removed.

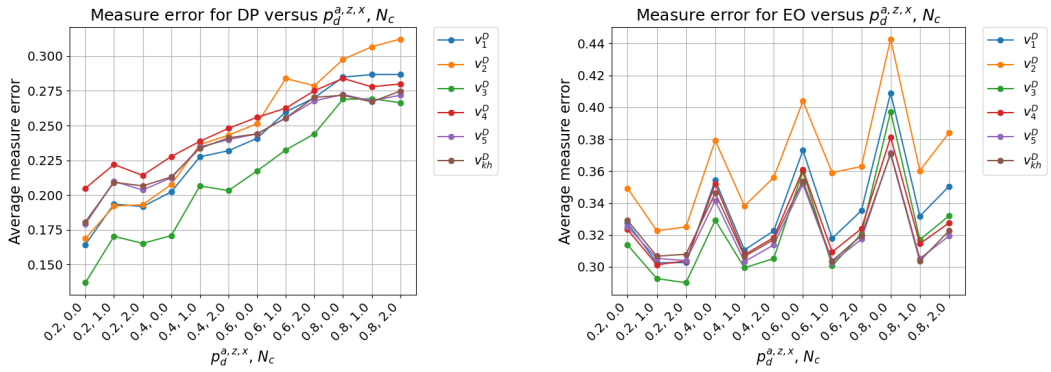


Figure 6: Average measure errors for DP and EO for different values of the parameters $p_d^{a,z,x}$ and N_c.

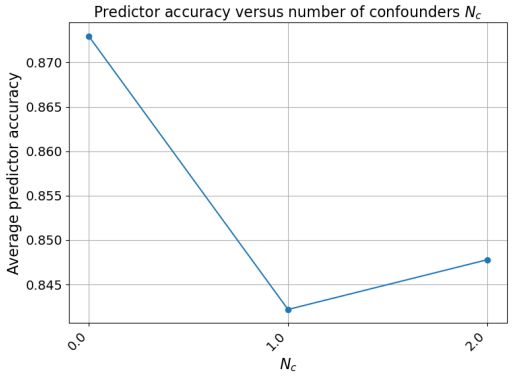


Figure 7: Average predictor accuracy for different numbers of confounders N_c.

Effect of the direct link $A \to Y$ on measure errors. Figure 8 shows the average measure errors for $\alpha_{a,y} \in \{0, 1\}$ for DP and EO. For DP, errors of the DP-aligned measures, v_1^D, v_2^D, and v_3^D, are minimally impacted by $\alpha_{a,y}$. The direct link from A to Y increases the error of the measures partially satisfying EO properties, v_4^D, v_5^D, and v_{kh}^D. For EO, the error for all the measures increases with the link $A \to Y$, with measure v_3^D showing the maximum increase.

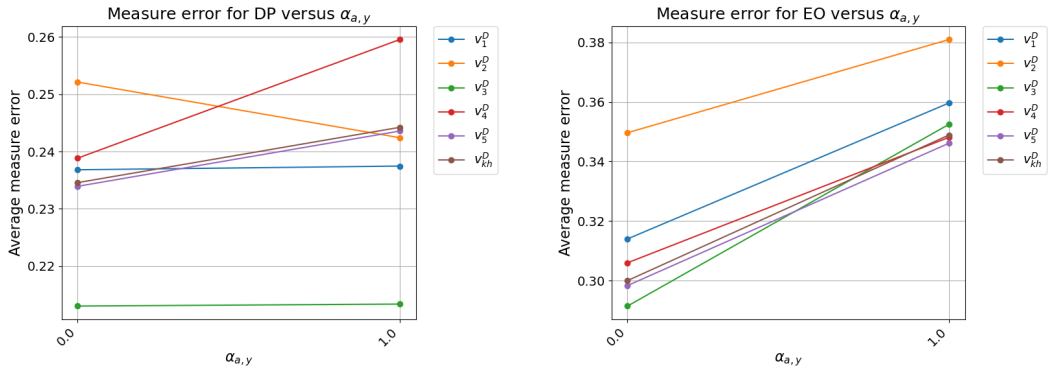


Figure 8: Average measures errors for DP and EO for $\alpha_{a,y} \in \{0, 1\}$.

Effect of the non-discriminatory components on measure errors. We compute the average ratios

$$R_A \triangleq \sum_{S \in [n]} \frac{UI(A; X_S \backslash Y)}{I(A; X_S)}, \tag{16}$$

$$R_X \triangleq \sum_{S \in [n]} \frac{UI(X_S; A \backslash Y)}{I(A; X_S)}, \tag{17}$$

and analyze their effect on the error of measures v_1^D, v_2^D, and v_3^D. We restrict this analysis to the discrimination bias with respect to DP since measures v_1^D, v_2^D, and v_3^D are aligned with DP. The correlation between R_A and the error of v_1^D and v_2^D is equal to 0.052 and 0.013, respectively. That is, including $UI(A; X_S \backslash Y)$ increases the error more than removing it does. The correlation between R_X and the error of v_1^D and v_3^D is equal to -0.017 and -0.072, respectively. This indicates that discarding the components $UI(X_S; A \backslash Y)$ slightly improves the measure correctness more than including it. This is supported by the results in Table 1.

16 GUIDELINE FOR MEASURES SELECTION

For measure selection, we provide the following guideline based on the theoretical foundations and empirical evaluations of our measures. Measures should align with the desired fairness notion: for DP, $\{v_1^D, v_2^D, v_3^D\}$ are suitable, while for EO, v_4^D is appropriate. When data is reliable and accurate predictors can be learned, removing certain features may increase discrimination bias. In such cases, v_3^D is recommended for DP as it is not monotonically non-decreasing and captures both positive and negative marginal discriminatory impacts. However, v_3^D should be avoided in two scenarios: (1) when the downstream predictor type is unknown, as v_3^D satisfies Condition 1 and fails to capture discriminatory impacts of feature subsets that are not predictive but exploitable by adversaries; and (2) when the sensitive attribute has significant unique information about features not included in the target label, as such information can be encoded—either unintentionally or intentionally—by predictors, yet it is not captured by v_3^D. For evaluating discriminatory impacts across any downstream predictor, including adversarial ones, v_1^D is recommended as it does not depend on Y and offers a more generic measure than v_3^D, though it may be less accurate when the predictor is known. Empirical evaluations show that when unique information $UI(A; X_S \setminus Y)$ is significant, v_1^D can be faulty because it primarily quantifies components exploitable by adversaries rather than predictors decoding Y (cf. Appendix 15); in such cases, v_2^D is more appropriate. When Y is highly correlated with A and the goal is to identify biased features sharing information between them, v_2^D is suitable. However, if $A \perp Y$ (independence between A and Y), v_2^D should be avoided as it fails to quantify discriminatory impacts for any feature subset. For EO, v_4^D is the aligned measure, and it can be used to identify features with positive discriminatory impacts. Yet, if conditioning on Y does not significantly affect the dependence between A and the data features, DP-aligned measures can be used to identify discriminatory impacts for EO; in our synthetic data experiment, datasets that does not show distinct DP-EO characteristics, measure v_3^D has low error for EO (cf. Section 4); hence it can be used to provide insight on features that potentially has negative or positive discriminatory impact for EO.

17 COMPREHENSIVE OVERVIEW OF BENCHMARK DATASETS

We briefly describe each of the 7 datasets as follows:

ACS Income dataset contains annual income information for over 1.66 million individuals. Target label Y indicates whether an individual earns more than \$50k. Race is selected as the sensitive attribute: Black/African-American (10% of the dataset) or White (90%). The positive rate (>\$50k) is 39% for White and 24.6% for Black individuals. We select 4 features for prediction: 'Class of worker'; 'Sex'; 'Usual weekly work hours', and 'Educational attainment'.

ACS Employment dataset contains employment status for over 3.24 million individuals. Y is the employment status. The positive rate is 57%. We select 'disability-status' as a sensitive attribute: Disabled individuals comprising 16% of the dataset (positive rate 21%) and non-disabled individuals 84% (positive rate 64%). We select 5 features for prediction: 'Race', 'Mrital status', 'Age', 'Mobility status', 'Parents' employment status'.

ACS Public Coverage dataset contains information about public-health insurance coverage for over 1.13 million low-income individuals not eligible for Medicare. Y indicates if a person has public health coverage. The positive rate is 30%. Race is selected as the sensitive attribute: White comprising 72% of the dataset (positive rate 27%) and Non-white 28% (positive rate 36%). We select 5 features for prediction: 'Age', 'Employment status of parents', 'Sex', 'Mobility status', and 'Childbirth within past 12 months'.

Adult dataset extracted from the 1994 Census database, contains information about 48,842 individuals' annual income. Y indicates whether an individual's income >\$50k. The positive rate is 25%. Gender is selected as the sensitive attribute: Females comprising 32% of the dataset (positive rate 11%) and males 68% (positive rate 31%). We selected 6 features: 'Age', 'Educational-num,' 'Capital-gain,' 'Capital-loss', 'Hours-per-week,' and 'Relationship'.

Census income (KDD) dataset derived from the 1994-1995 U.S. Census Bureau surveys; predicts annual income > \$50k; contains 399,285 records. The positive rate is 6%. Sex is selected as the sensitive attribute: Males comprising 52% of the dataset (positive rate 10%) and females 48%

(positive rate 2.5%). We select 6 features: 'Education', 'Marital status', 'Race', 'Capital gains', 'Capital loss', and 'Number of employees'.

ProPublica COMPAS dataset contains criminal history and demographic information for defendants in Broward County, Florida (2013-14); 5,334 records. Y indicates whether an individual was arrested for a crime within 2 years of release. The positive rate is 47%. Race is selected as the sensitive attribute; Caucasian individuals comprising 39% of the dataset (positive rate 39%) and Black Americans comprising 61% of the dataset (positive rate 52%). We select 5 features: 'age category', 'charge degree', 'sex', 'priors count', and 'Length of stay'.

Heritage health dataset contains insurance claims and physician records for over 60,000 patients. Y indicates whether the Charlson Comorbidity Index is zero. The positive rate is 35%. Age is selected as the sensitive attribute, with individuals older than 70 comprising 10% of the dataset (positive rate 57%) and those < 70, 90% (positive rate 53%). We select 5 features: 'Sex', 'Claim year', 'Service location', 'Payment delay duration', and 'Days since first service'.

18 Experimental results of real-world datasets for equality of opportunity

Measure errors with respect to equality of opportunity (EOP) evaluated on the real-world datasets are shown in Table 6 (See Section 4 for the definition of the measure error, and see Appendix 14 for the definition of the discrimination bias with respect to EOP).

Measure	ACS Income		ACS Employment		ACS Public Coverage		Adult		Census Income (KDD)		COMPAS		Health	
	$(Y=0)$	$(Y=1)$	$(Y=0)$	$(Y=1)$	$(Y=0)$	$(Y=1)$	$(Y=0)$	$(Y=1)$	$(Y=0)$	$(Y=1)$	$(Y=0)$	$(Y=1)$	$(Y=0)$	$(Y=1)$
v_1^D	0.346	0.469	0.098	0.044	0.414	0.454	0.048	0.143	0.302	0.299	0.100	0.112	0.343	0.461
v_2^D	0.339	0.444	0.098	0.044	0.377	0.417	0.092	0.171	0.326	0.290	0.126	0.138	0.325	0.443
v_3^D	0.180	0.428	0.081	0.029	0.343	0.321	0.062	0.128	0.348	0.353	0.040	0.025	0.214	0.309
v_4^D	0.405	0.476	0.116	0.062	0.345	0.386	0.049	0.145	0.269	0.204	0.128	0.140	0.363	0.481
v_5^D	0.393	0.479	0.102	0.048	0.380	0.420	0.050	0.145	0.274	0.245	0.127	0.140	0.411	0.530
v_{kh}^D	0.389	0.473	0.096	0.041	0.364	0.405	0.050	0.146	0.271	0.243	0.127	0.140	0.416	0.534

Table 6: Average measure errors in capturing marginal discriminatory impacts for EOP on real-world datasets.

19 Conclusion

In conclusion, we developed a model-agnostic framework to quantify individual dataset features' impact on discrimination bias of supervised ML models. Our framework proposes information-theoretic measures for feature sets' discriminatory impact, through considering inter-dependencies among data variables, and utilizes Shapley value function to deduce marginal contributions of individual features to overall discrimination. We constructed a set of discrimination measures through an axiomatic approach, while distinguishing between measures designed for demographic parity and equalized odds fairness criteria. Through a comprehensive empirical analysis on a large number of synthetic datasets, and 7 real-world benchmark datasets, we validated the efficacy of our measures in capturing discrimination bias for different fairness criteria and under distinct data conditions. Measures aligned with demographic parity accurately capture feature contributions to model discrimination under this fairness notion; similarly, for the equalized odds-aligned measure. Notably, measures satisfying both notions simultaneously were less effective in capturing feature contributions to downstream model discrimination for either fairness notions.