Demographic Parity Inspector: Fairness Audits via the Explanation Space

Carlos Mougan [1] **Laura State** [2 3] **Antonio Ferrara** [4 5] **Salvatore Ruggieri** [2] **Steffen Staab** [1 6]

Abstract

Even if deployed with the best intentions, machine learning methods can perpetuate, amplify or even create social biases. Measures of (un-)fairness have been proposed as a way to gauge the (non-)discriminatory nature of machine learning models. However, proxies of protected attributes causing discriminatory effects remain challenging to address. In this work, we propose a new algorithmic approach that measures group-wise demographic parity violations and allows us to inspect the causes of inter-group discrimination. Our method relies on the novel idea of measuring the dependence of a model on the protected attribute based on the explanation space, an informative space that allows for more sensitive audits than the primary space of input data or prediction distributions, and allowing for the assertion of theoretical demographic parity auditing guarantees. We provide a mathematical analysis, synthetic examples, and experimental evaluation of real-world data. We release an open-source Python package with methods, routines, and tutorials.

1. Introduction

There exists a range of metrics (demographic parity, equal opportunity, average absolute odds . . .) that judge the *global* degree of unfairness arising from the use of machine learning models. Together with some predictive performance measures (accuracy, precision, AUC . . .), these metrics are commonly used to deliver classifiers that achieve high degrees of predictive performance and fairness e.g. (Zafar et al., 2017; Rodolfa et al., 2021). Unfortunately, approaches that enforce low measures of unfairness at the global level may cause new kinds of unfairness *at the subgroup level*.

[1]Department of Computer Science, University of Southampton,United Kingdom [2]University of Pisa, Italy [3]Scuola Normale Superiore, Pisa, Italy [4]GESIS - Leibniz Institute for the Social Sciences Cologne, Germany [5]RWTH Aachen University Aachen, Germany [6]University of Stuttgart, Institute of Parallel and Distributed Systems, Stuttgart, Germany. Correspondence to: Carlos Mougan <c.mougan@soton.ac.uk>.

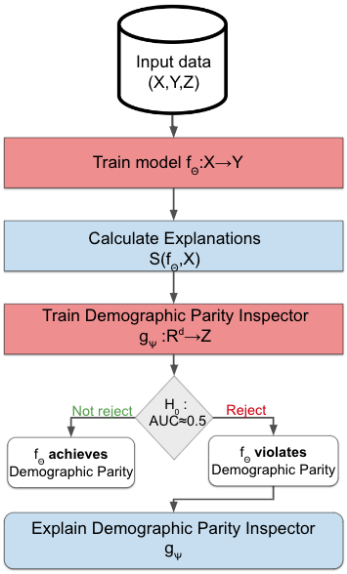

Figure 1: Demographic Parity Inspector(g_ψ) workflow. The training of models is shaded in red while the derivations of explanations are shaded in blue. The model f_θ is learned based on training data, $\{(x_i, y_i)\}$, and outputs the explanations $\mathcal{S}(f_\theta, X)$. The "DP Inspector" receives the explanations to predict the protected attribute, Z. Then if the AUC is above 0.5 then there is a demographic parity violation. We can interpret the reasons for demographic parity violations on g_ψ with explainable AI techniques

As (Ruggieri et al., 2023) shows, pushing for well-balanced fairness among groups at such a global scale may imply novel discrimination at inter-group levels leading to new kinds of unfairness introduced by an erroneous use of fair-AI methods.

The Yule's effect (Yule, 1900; Simpson, 1951; Pearson et al., 1899) occurs when positive and negative associations between the model predictions $f(X)$ and the protected attribute Z cancel out producing a vanishing correlation in the mixture of the distribution. It occurs whenever we aim at group fairness, such as independence $f(X) \perp Z$, but we wrongly disregard control for the protected attribute Z. Fair machine learning algorithms may result in disparate effects on separate distributions, with some subgroups impacted positively (higher fairness) and other subgroups impacted

negatively (lower fairness) (Ruggieri et al., 2023). For example, enforcing fairness for one discriminated ethnic group at the country level may imply novel kinds of unfairness for the subgroup of this ethnicity in a particular local state.

Previous methods of measuring demographic parity have relied on the predictions in the output space whose low dimensionality is prone to fall under Yule's effect, and on statistical measures on input data that is model-independent. Furthermore, on many occasions detecting and quantifying fairness violations is not enough, there is also the need to pinpoint *What are the specific sources of discrimination?*.

We address this problem by providing a novel kind of demographic parity measurement that analyzes how different protected groups are treated at the distributional and instance level and how each feature contributes to the demographic parity violation. This approach relies on the explanation space, a novel data space that has theoretical guarantees while performing demographic parity inspections. Our main contributions are:

- We use the explanation space to measure fairness by providing theoretical guarantees and experiments with synthetic data and real data use cases.
- We introduce a "Demographic Parity Inspector" method that allows for interpretable fairness quantification that sheds insights on the root causes of unfairness.
- We release an open-source Python package[1], which implements our "Demographic Parity Inspector" that is `scikit-learn` compatible (Pedregosa et al., 2011), together with code and tutorials for reproducibility.

2. Foundations and Related work

This section introduces the foundations of fairness and explainable AI needed to establish the context within which this contribution is made. We then compare our work to the existing one and outline the differences.

2.1. Explainable AI: Shapley values

Explainability has become an important concept in legal and ethical guidelines for data processing and machine learning applications (Selbst and Barocas, 2018). A wide variety of methods have been developed, aiming to account for the decision of algorithmic systems (Guidotti et al., 2018; Mittelstadt et al., 2019; Arrieta et al., 2020). One of the most popular approaches to explainability in machine learning (ML) has been Shapley values These values are used to attribute relevance to features according to how the model relies on them (Lundberg et al., 2019; Lundberg and Lee, 2017). Shapley values are a coalition game theory concept that aims to allocate the surplus generated by the grand coalition in a game to each of its players (Shapley, 1953).

For set of players $N = \{1, \ldots, p\}$, and a value function $\text{val} : 2^N \to \mathbb{R}$, the Shapley value $\mathcal{S}_j$ of the j'th player is defined as the averaged additional contribution of player j in all possibles coalitions of players:

$$\mathcal{S}_j = \sum_{T \subseteq N \setminus \{j\}} \frac{|T|!(p - |T| - 1)!}{p!} (\text{val}(T \cup \{j\}) - \text{val}(T))$$

In the context of machine learning models, players correspond to features $X_1, \ldots, X_p$, and the contribution of the feature X_j is with reference to the prediction of a model f for an instance $x^\star$ to be explained. Thus, we write $\mathcal{S}(f, x^\star)_j$ for the Shapley value of feature X_j in the prediction $f(x^\star)$. We denote by $\mathcal{S}(f, x^\star)$ the vector of Shapely values $(\mathcal{S}(f, x^\star)_1, \ldots, \mathcal{S}(f, x^\star)_p)$. There are two variants for the term $\text{val}(T)$ (Aas et al., 2021; Chen et al., 2020; Zern et al., 2023): the *observational* and the *interventional*. When using the observational conditional expectation, we consider the expected value of f over the joint distribution of all features conditioned to fix features in T to the values they have in $x^\star$:

$$\text{val}(T) = E[f(x^\star_T, X_{N \setminus T}) | X_T = x^\star_T] \tag{1}$$

where $f(x^\star_T, X_{N \setminus T})$ denotes that features in T are fixed to their values in $x^\star$, and features not in T are random variables over the joint distribution of features. Opposed, the interventional conditional expectation considers the expected value of f over the marginal distribution of features not in T:

$$\text{val}(T) = E[f(x^\star_T, X_{N \setminus T})] \tag{2}$$

In the interventional variant, the marginal distribution is unaffected by the knowledge that $X_T = x^\star_T$. In general, the estimation of (1) is difficult, and some implementations (e.g., SHAP) actually consider (2) as the default one. In the case of decision tree models, TreeSHAP offers both possibilities.

The Shapley value framework is the only feature attribution method that satisfies the properties of Efficiency, Symmetry, Uninformative and Additivity (Molnar, 2019; Shapley, 1953; Winter, 2002; Aumann and Dreze, 1974). We recall next the key property of efficiency and uninformative:

Efficiency Feature contributions add up to the difference of prediction for $x^\star$ and the expected value of f:

$$\sum_{j \in N} \mathcal{S}(f, x^\star) = f(x^\star)_j - E[f(X)]) \tag{3}$$

The following property only holds for the interventional variant (SHAP values), but not for the observational one:

[1] to be released upon acceptance

Uninformative A feature X_j that does not change the predicted value (i.e., for all x, x'_j: $f(x_{N\setminus\{j\}}, x_j) = f(x_{N\setminus\{j\}}, x'_j)$) have a Shapley value of zero, i.e., $\mathcal{S}(f, x^\star)_j = 0$.

In the case of a linear model $f_\beta(x) = \beta_0 + \sum_j \beta_j \cdot x_j$, the SHAP values turns out to be $\mathcal{S}(f, x^\star) = \beta_i(x_i^\star - \mu_i)$ where $\mu_i = E[X_i]$. For the observational case, this holds only if the features are independent (Aas et al., 2021).

2.2. Fairness notions

Various definitions of fairness in machine learning have been proposed (see, e.g., (Mehrabi et al., 2021; Finocchiaro et al., 2021; Barocas et al., 2019) for recent overviews).

Algorithmic fairness can be decomposed into two central notions: *between-group*, also known as group fairness, and a *within-group* component or individual fairness (Speicher et al., 2018). While the *within-group* or individual notions of fairness emphasize on *treating similar individuals similarly* (Dwork et al., 2012), *between-group* or group fairness notions aim to establish some form of parity between groups of individuals based on shared sensitive attributes like gender or race.

Selecting a measure to compare fairness between two sensitive groups has been a highly discussed topic, where results such as (Chouldechova, 2017; Hardt et al., 2016; Kleinberg et al., 2017), have highlighted the impossibility to satisfy simultaneously three type of fairness measures: demographic parity (Dwork et al., 2012), equalized odds (Hardt et al., 2016), and predictive parity (Corbett-Davies et al., 2017; Ruf and Detyniecki, 2021; Wachter et al., 2020). In this work, we focus on demographic parity as fairness metric, because it is not error dependent but relies specifically on the model predictions, allowing us to create "early warnings" in advance to avoid building/deploying machine learning models undesired behavior.

2.3. Related work

2.3.1. Fairness measuring and auditing

In this work, we rely on the following definition of audit studies: "Audits are tools for interrogating complex processes, often to determine whether they comply with company policy, industry standards or regulations" (Liu et al., 2012; Raji et al., 2020).

Algorithmic audits are closely linked to audits studies as understood in the social sciences, with a strong emphasis on social justice and participatory action (Vecchione et al., 2021). A recent survey on public algorithmic audit identified four categories of "problematic machine behaviour" that can be unveiled by audit studies: discrimination, distortion, exploitation and misjudgement (Bandy, 2021). This taxonomy is highly helpful when putting forward auditing studies, and also relevant for this work: our "Demographic Parity Inspector" can be seen as a tool to help understanding discrimination in ML models, thus, it falls into the first category.

Previous work has relied on measuring and calculating demographic parity on the model predictions (Raji et al., 2020; Kearns et al., 2018), or on the input data (Fabbrizzi et al., 2022; Yang et al., 2020; Zhao et al., 2017). In this work, we perform demographic parity measures on the explanation space, a novel model projection space that allows us to overcome sensibility issues and Yule's paradox w.r.t fairness on the prediction space and avoid false positives when there is bias in the data but there is no demographic parity violation on the model.

2.3.2. Explainability and fair supervised learning

The intersection of fairness and explainable AI has been an active topic in recent years. For example, (Stevens et al., 2020) presents an approach to explaining fairness based on adapting the Shapley value function to explain model unfairness. They also introduce a new meta-algorithm that considers the problem of learning an additive perturbation to an existing model in order to impose fairness. In our work we do not adopt the Shapley value function, instead, we use the theoretical Shapley properties to provide fairness auditing guarantees. Our *Demographic Parity Inspector* is not perturbation based but uses Shapley values to project the model to the explanation space, and then measures demographic parity violations. It also allows us to pinpoint what are the specific features driving this violation.

In (Grabowicz et al., 2022), the authors present a post-processing method based on Shapley values aiming to detect and nullify the influence of a protected attribute on the output of the system. For this, they assume there are direct causal links from the data to the protected attribute and that there are no measured confounders. Our work does not make use of causal graphs but exploits the theoretical properties of the Shapley values in order to obtain fairness model auditing guarantees.

Other works have relied on other explainability techniques such as counterfactual (Kusner et al., 2017; Manerba and Guidotti, 2021; Mutlu et al., 2022), in our work we don't focus on counterfactual explanations but on providing fairness auditing guarantees for demographic parity using feature attribution explanations.

3. Methodology

3.1. Formalization

In supervised learning, a function $f_\theta : X \to Y$, also called a model, is induced from a set of observations, called training set, $\mathcal{D}^{tr} \subseteq X \times Y$, where $X = \{X_1, \ldots, X_p\}$ are (predictive) features and Y is a target feature. f_θ belongs to a family of functions $\mathcal{F}$ parametric in θ. The domain of the target feature is $dom(Y) = \{0, 1\}$ (binary classification) or $dom(Y) = \mathbb{R}$ (regression). For binary classification, we assume a probabilistic classifier, and we actually denote by $f_\theta(x)$ the estimates of the probability $P(Y = 1|X = x)$ over the (unknown) distribution of $X \times Y$. For regression, $f_\theta(x)$ estimates $E[Y|X = x]$.

A dataset of instances $\mathcal{D} = \{x_1, \ldots, x_n\} \subseteq X$ is called an *input space*. The dataset of predictions $f_\theta(\mathcal{D}) = \{f_\theta(x) \mid x \in \mathcal{D}\}$ is called the *prediction space*. We now introduce a transformation mapping f_θ and $\mathcal{D}$ into a new space, which we call the *explanation space*.

Definition 3.1. *(Explanation Space)* Let $\mathcal{S} : \mathcal{F} \times \mathcal{X} \to \mathbb{R}^p$ be an explanation function for models in $\mathcal{F}$ and instances $x \in X$, e.g., Shapley values. For a model f_θ, we map an input space $\mathcal{D}$ into a dataset in $\mathbb{R}^p$ called the *explanation space*: $\mathcal{S}(f_\theta, \mathcal{D}) = \{\mathcal{S}(f_\theta, x)|x \in \mathcal{D}\}$.

We assume a feature modeling protected social groups, is denoted by Z, called *protected feature*, and to be binary valued in the theoretical analysis. Z can either be included in the predictive features X used by a model or not. If not, we assume that it is still available. Even in the absence of the protected feature in training data, a model can discriminate against the protected groups by using correlated features as a proxy of the protected one (Pedreschi et al., 2008).

A fairness metric quantifies the degree of discrimination or unfairness of a model (Mehrabi et al., 2021). In this work, we focus on Demographic Parity (DP), as this fairness metric does not require a ground truth target variable, allowing for our method to work in its absence (Aka et al., 2021), and under distribution shift conditions (Mougan et al., 2022) where model performance metrics are not feasible to calculate (Garg et al., 2021a;b; Mougan and Nielsen, 2023). DP requires independence of the model's output from the protected features, written $f_\theta(X) \perp Z$ viewing $f_\theta(X)$ and Z as random variables. We introduce a similar notion with reference to the explanation space by considering the multivariate random variable $\mathcal{S}(f_\theta, X)$.

Definition 3.2. *(Demographic Parity (DP))*. We have demographic parity on the prediction space if $f_\theta(X) \perp Z$, and demographic parity on the explanation space if $\mathcal{S}(f_\theta, X) \perp Z$.

We say that there is DP violation if $f_\theta(X) \not\perp Z$ and $\mathcal{S}(f_\theta, X) \not\perp Z$ respectively. DP in the explanation space will be shown not to suffer from Yule's effect. Notice that, for classifiers, DP is stronger than requiring independence of the predicted class and the protected feature (Jiang et al., 2019; Chiappa et al., 2020) .

3.2. Demographic Parity Inspector

Our approach is based on the properties of the Shapley values whose intuition is that the explanation of a certain model's output encapsulates more information than the output itself. We split the available data into three equal parts $\mathcal{D}^{tr}, \mathcal{D}^{val}, \mathcal{D}^{te} \subseteq X \times Y$. Where $\mathcal{D}^{tr}$ is the training set of f_θ. Following the intuition above, $\mathcal{D}^{val}$ is used to train another model g_ψ on the space $\mathcal{S}(f_\theta, \mathcal{D}^{val}_{X \setminus Z}) \times \mathcal{D}^{val}_Z$ where the predictive features are in the explanation space $\mathcal{S}(f_\theta, \mathcal{D}^{val}_{X \setminus Z})$ (excluding Z) and the target feature is the protected feature Z. The model g_ψ is called a "Demographic Parity Inspector", and it belongs to a family $\mathcal{G}$ possibly different from $\mathcal{F}$. In this work, we restrict to linear models for $\mathcal{G}$. The parameter ψ optimizes a loss function ℓ:

$$\psi = \arg\min_{\tilde{\psi}} \sum_{(x,z) \in \mathcal{D}^{val}} \ell(g_{\tilde{\psi}}(\mathcal{S}(f_\theta, x)), z) \qquad (4)$$

Finally, we use $\mathcal{D}^{te}$ for testing the approach and for comparison with baselines. See also workflow of Figure 1 for a visualization of the process.

Besides detecting and measuring fairness violations in machine learning models, a common desideratum is to understand what are the specific features driving the discrimination. Using the "Demographic Parity Inspector" as an auditor method that aims to depict and quantify possible fairness violations does not only report a metric but provides information on *what* features are the cause of this disparate treatment. We propose to solve this issue by applying explainable AI techniques to the "Inspector"

4. Theoretical Analysis

Our theoretical analysis consists of three parts. The first one studies the relation of the explanation of the protected attribute and DP on the prediction space. The second one focuses on the DP on explanation space, relating it to DP on the prediction space and on the input data. The third part provides a mathematical analysis on specific scenarios for the "Demographic Parity Inspector". Throughout this section, we assume an exact calculation of the Shapely values $\mathcal{S}(f_\beta, x))$ for an instance x, possibly for the interventional (2) or the observational variant (1).

4.1. Explanations of the protected feature and DP

This first question we consider is whether by looking at the Shapely values of the protected feature is a viable approach to test for DP on the prediction space. The following result

considers a linear model (with unknown coefficients) over *independent* features. In such a case, resorting to Shapley values leads to an exact test. In the following, we write $distinct(\mathcal{D}_i)$ for the number of distinct values in the i-th feature of dataset $\mathcal{D}$, and $\mathcal{S}(f_\beta, \mathcal{D})_i \equiv 0$ if the Shapley values of the i-th features are all 0's.

Lemma 1. Consider a linear model $f_\beta(x) = \beta_0 + \sum_j \beta_j \cdot x_j$. Let Z be the i-th feature, i.e. $Z = X_i$, and let $\mathcal{D}$ be a dataset such that $distinct(\mathcal{D}_i) > 1$. If the features in X are independent, then $\mathcal{S}(f_\beta, \mathcal{D})_i \equiv 0 \Leftrightarrow f_\beta(X) \perp Z$.

Proof. It turns out $\mathcal{S}(f_\beta, x)_i = \beta_i \cdot (x_i - E[X_i])$ This holds in general for the interventional variant (2), and for independent features, also for the observational variant (1) (Aas et al., 2021). Since $distinct(\mathcal{D}_i) > 1$, we have that $\mathcal{S}(f_\beta, \mathcal{D})_i \equiv 0$ iff $\beta_i = 0$. By independence of X, this is equivalent to $f(X) \perp X_i$, i.e., $f(X) \perp Z$. QED

In words, the test in the lemma consists of checking that the Shapley values of the protected feature are zero's, i.e., that the protected feature does not contribute to the output of the model. Such a result does not extend to the case of dependent features. Consider $Z = X_2 = X_1^2$, and the linear model $f(x_1, x_2) = \beta_0 + \beta_1 \cdot x_1$, with $\beta_1 \neq 0$ and $\beta_2 = 0$. In the interventional variant, since such a model does not use x_2, by the Uninformativeness property, $\mathcal{S}(f_\beta, x)_2 = 0$. However, clearly $f(X_1, X_2) = \beta_0 + \beta_1 \cdot X_1 \not\perp X_2$. For the observational variant, (Aas et al., 2021) show that:

$$val(T) = \sum_{i \in N \setminus T} \beta_i \cdot E[X_i | X_T = x_T^\star] + \sum_{i \in T} \beta_i \cdot x_i^\star \quad (5)$$

from which, we calculate:

$$\mathcal{S}(f, x^\star)_2 = \frac{\beta_1}{2} E[X_1 | X_2 = x_2^\star]$$

We have $\mathcal{S}(f, \mathcal{D})_2 \equiv 0$ iff $E[X_1 | x_2 = x_2^\star] = 0$ for all $x^\star$ in $\mathcal{D}$. For the following marginal distribution $P(X_1 = v) = 1/4$ for $v = 1, -1, 2, -2$, since $X_2 = X_1^2$, it holds that $E[X_1 | x_2 = v] = 0$. Thus $\mathcal{S}(f, \mathcal{D})_2 \equiv 0$. However, again $f(X_1, X_2) = \beta_0 + \beta_1 \cdot X_1 \not\perp X_2$.

The counterexample shows that focusing only on the Shapley values of the protected feature is not a viable way to reason about DP of the model on the prediction space.

4.2. DP on the explanation and prediction spaces

This section is divided into two parts that comprise the theoretical comparisons of DP on the explanations space against (i) the prediction space and (ii) the input data.

We start by observing that DP on the explanation space is a sufficient condition for DP in the prediction space.

Lemma 2. If $\mathcal{S}(f_\theta, X) \perp Z$ then $f_\theta(X) \perp Z$.

Proof. By the propagation of independence in probability distributions, this implies $(\sum_i \mathcal{S}_i(f_\theta, X) + c) \perp Z$ where c is any constant. By setting $c = E[f(X)]$ and by the efficiency property (3), we have the conclusion. QED

Therefore, a DP violation on the prediction space is also a DP violation in the explanation space, which accounts then for a stricter notion of fairness. The other direction does not hold. We can have dependence of Z from the explanation features, but the sum of such features cancel out resulting in perfect DP on the prediction space. This issue is also known as the Yule's effect (Ruggieri et al., 2023).

Example: *Consider the model $f(x_1, x_2) = x_1 + x_2$. Let $Z \sim Ber(0.5)$, $A \sim U(-3, -1)$, and $B \sim N(2, 1)$ be independent, and let us define:*

$$X_1 = A \cdot Z + B \cdot (1 - Z) \quad X_2 = B \cdot Z + A \cdot (1 - Z)$$

We have $f(X_1, X_2) = A + B \perp Z$ since A, B, Z are independent. Let us calculate $\mathcal{S}(f, X)$ in the two cases $Z = 0$ and $Z = 1$. If $Z = 0$, we have $f(X_1, X_2) = B + A$, and then $\mathcal{S}(f, X)_1 = B - E[B] = B - 2 \sim N(0, 1)$ and $\mathcal{S}(f, X)_2 = A - E[A] = A + 2 \sim U(-1, 1)$. Similarly, for $Z = 1$, we have $f(X_1, X_2) = A + B$, and then $\mathcal{S}(f, X)_1 = A - E[A] = A + 2 \sim U(-1, 1)$ and $\mathcal{S}(f, X)_2 = B - E[B] = B - 2 \sim N(0, 1)$. This shows:

$$P(\mathcal{S}(f, X) | Z = 0) \neq P(\mathcal{S}(f, X) | Z = 1)$$

and then $\mathcal{S}(f, X) \not\perp Z$. Notice this example holds both for the interventional and the observational cases, as we exploited Shapley values of a linear model over independent features, namely A, B, Z.

Statistical dependence between the input space, X, and the protected attribute Z, is another technique, which, however, disregards the model f_θ. For fair models, which are able not to (directly or indirectly) rely on Z, such a technique can lead to false positive detection of DP violation.

Example: *Let $X = X_1, X_2, X_3$ be independent features such that $E[X_1] = E[X_2] = E[X_3] = 0$, and $X_1, X_2 \perp Z$, and $X_3 \not\perp Z$. The target feature is defined as $Y = X_1 + X_2$, hence it is also independent from Z. Assume a linear regression model $f_\beta(x_1, x_2, x_3) = \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3$ trained from a sample data from (X, Y) with $\beta_1, \beta_2 \approx 1$ and $\beta_3 \approx 0$. Intuitively, this occurs when a number of features are collected to train a classifier without a clear understanding of which of them contributes to the prediction. It turns out that $X \not\perp Z$ but, for $\beta_3 = 0$ (which can be obtained by some fairness regularization method (Kamishima et al., 2011)), we have $f_\beta(X_1, X_2, X_3) = \beta_1 \cdot X_1 + \beta_2 \cdot X_2 \perp Z$. By reasoning as in the proof of Lemma 1, we have $\mathcal{S}(f_\beta, X) = (\beta_1 \cdot X_1, \beta_2 \cdot X_2, 0)$ and then $\mathcal{S}(f_\beta, X) \perp Z$. This holds both in the interventional and in the observational variants.*

The above represents an example where the input data depends on the protected feature, but the model and the explanations are independent. In this case, the explanation space correctly detects that there is no DP violation on the model.

4.3. DP inspecting via the explanation space

We provide here the theoretical foundations about the "Demographic Parity Inspector". First, using the Bayesian optimal classifier as inspector, we show that if the performance of the "Demographic Parity Inspector" is null, then there is DP on the prediction space. Secondly, we show how from an interpretable inspector, such as a linear model, we can derive information on the source of the DP violation.

4.3.1. The Bayesian Optimal Classifier as an auditor

Let us start with an equivalent condition of the DP in the explanation space.

Lemma 3. $\mathcal{S}(f_\theta, X) \perp Z$ iff $P(Z = 1|\mathcal{S}(f_\theta, X)) = c$ almost surely for some constant c.

Proof. By definition of independence, for any value s with non-zero probability $P(\mathcal{S}(f_\theta, X) = s) > 0$, $\mathcal{S}(f_\theta, X) \perp Z$ iff $P(Z = 1|\mathcal{S}(f_\theta, X)) = P(Z = 1)$ and $P(Z = 0|\mathcal{S}(f_\theta, X)) = P(Z = 0)$. Since Z is binary valued, the second equivalence holds iff $P(Z = 1|\mathcal{S}(f_\theta, X)) = P(Z = 1)$. This shows the conclusion, and also that c must necessarily be $c = P(Z = 1)$. QED

Another equivalent formulation can be stated in terms of AUC of the Bayes Optimal Classifier, which then provides a way of testing for the DP violation in the explanation space. Also, by Lemma 2, we have a sufficient condition for testing the DP in the prediction space. Putting together these observations we obtain the following key result.

Theorem 1. *Assume that the "Demographic Parity Inspector" g_ψ is the Bayes Optimal Classifier, i.e., such that:*

$$g_\psi(s) = P(Z = 1|\mathcal{S}(f_\theta, X) = s)$$

The AUC of g_ψ is 0.5 iff $\mathcal{S}(f_\theta, X) \perp Z$. Moreover, if the AUC of g_ψ is 0.5, then $f_\theta(X) \perp Z$.

Proof. First, we observe that the AUC of $g_\psi(s)$ is 0.5 iff $g_\psi(s)$ is constant almost surely. In fact, the AUC (Gao and Zhou, 2015) of g_ψ:

$$AUC = E[I((Z - Z')(g_\psi(\mathbf{X}) - g_\psi(\mathbf{X}') > 0) + 0.5 \cdot I(g_\psi(\mathbf{X}) = g_\psi(\mathbf{X}'))|Z \neq Z']$$

is bounded from below by 0.5, since $E[I((Z-Z')(g_\psi(\mathbf{X})-g_\psi(\mathbf{X}') > 0)|Z \neq Z'] \geq 0$ by definition of g_ψ. By Lemma 3, $AUC = 0.5$ is then equivalent to $\mathcal{S}(f_\theta, X) \perp Z$. Moreover, if $AUC = 0.5$, we have $\mathcal{S}(f_\theta, X) \perp Z$, and then, by Lemma 2, $f_\theta \perp Z$. QED

The Bayes Optimal Classifier maximizes the AUC (Gao and Zhou, 2015). Hence, there is DP violation in the explanation space iff any inspector has an AUC larger than 0.5. Moreover, if all inspectors have an AUC of 0.5 or smaller, then there is a DP violation in the prediction space. Notice, however, that all the above theoretical results refer to the population statistics, while in experiments we consider estimators of those statistics over samples of the population and approximate calculations of the explanation space.

4.3.2. Linear models and linear inspectors

This example showcases one of our main contributions: detecting the source of demographic parity violation. For this, we have used the basic case of i.i.d. data and linear models. In the following experimental sections, we will provide the experiments with real-data and non-linear models.

Example: *Let $X = (X_1, X_2, X_3)$ with X_1, X_2, X_3 independent variables with $E[X_i] = 0$, and such that $X_3 \not\perp Z$. Given a random sample of i.i.d. observations from (X, Y), a linear model $f_\beta(x_1, x_2, x_3) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3$ can be built by OLS (Ordinary Least Square) estimation. By reasoning as in the proof of Lemma 1, $\mathcal{S}(f_\theta, x)_i = \beta_i \cdot x_i$. Consider now a linear "Demographic Parity Inspector" $g_\psi(s) = \psi_0 + \psi_1 \cdot s_1 + \psi_2 \cdot s_2 + \psi_3 \cdot s_3$, which can be written in terms of the x's as: $g_\psi(x) = \psi_0 + \psi_1 \cdot \beta_1 \cdot x_1 + \psi_2 \cdot \beta_2 \cdot x_2 + \psi_3 \cdot \beta_3 \cdot x_3$. By OLS estimation properties, we have $\psi_1 \approx cov(\beta_1 \cdot X_1, Z)/var(\beta_1 \cdot X_1) = cov(X_1, Z)/(\beta_1 \cdot var(X_1)) = 0$ and analogously $\psi_2 \approx 0$. Finally, $\psi_3 \approx cov(X_3, Z)/(\beta_3 \cdot var(X_3)) \neq 0$. In summary, the coefficients of g_ψ provide information on which feature contributes (and how much it contributes) to the dependence between the prediction $f_\beta(X)$ and the protected feature Z.*

5. Experiments

We divide the experiments into two sections: a first section, using synthetic data, where we compare the "Demographic Parity Inspector" to previous methods for measuring demographic parity violations, and a second one, presenting use-cases on real data, that shows the auditing and inspection results of our approach. Thus, we show the reliability of the "Demographic Parity Inspector" in real datasets.

5.1. Experiments with synthetic data

The experiments of this section aim to show the sensibility of our method for false positives and true positives when detecting demographic parity violations.

To generate a synthetic dataset for both cases, we first draw $10,000$ samples from a normal distribu-

tion $X_1 \sim N(0,1), X_2 \sim N(0,1), (X_3, X_4) \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \gamma\\ \gamma & 1\end{bmatrix}\right)$. We then define a binary feature Z with values 1 `if` $X_4 > 0$, `else` 0. We compare the fairness auditing methods while increasing the correlation $\gamma = r(X_3, Z)$ from 0 to 1. In both cases, our model $f_\beta(X_1, X_2, X_3)$, is a function over the domain of the features. Further, we introduce two experimental scenarios:

Demographic Parity violation in the data and in the model (Indirect): In this case, there is a demographic parity violation in the input data that is learned by the model. The predictor's domain is (X_1, X_2, X_3), and the features are independent of each other. The protected attribute is Z (binary-valued) and its correlation with the predictor's domain (X_3, Z) is parameterized by $\gamma(X_3, Z)$, allowing to adjust for discrimination. To generate the synthetic demographic parity violation in the model we create the target $Y = \sigma(X_1 + X_2 + X_3)$, where σ is the logistic function

Demographic Parity violation in the data but not in the model (Uninformative): In this case, the demographic parity violation on the input data remains the same but the relationship of the target variable changes, it is now independent of the protected attribute. The target function is defined as $Y = \sigma(X_1 + X_2)$, and, the γ parameter allows adjusting for discrimination in the training data even if the model does not capture it. The target is completely independent of the protected attribute, $Y \perp X_3 \Rightarrow Y \perp Z \Rightarrow f_\beta \perp Z$.

As ablation studies, we compare the "Demographic Parity Inspector", g_ψ, that learns on the explanations space (eq. 4) against learning on the following other spaces: on the input data space, to detect if there is information about the protected attribute in the data,$\Upsilon = \arg\min_{\tilde{\Upsilon}} \sum_{(x,z)\in\mathcal{D}^{val}} \ell(g_{\tilde{\Upsilon}}(x), z)$ On the prediction space, to detect the protected attribute on the model predictions, suffering from Yule's paradox $\upsilon = \arg\min_{\tilde{\upsilon}} \sum_{(x,z)\in\mathcal{D}^{val}} \ell(g_{\tilde{\psi}}(f_\theta(x)), z)$ And in the combination of both, where it learns the protected attribute in the input data and in the model predictions $\phi = \arg\min_{\tilde{\phi}} \sum_{(x,z)\in\mathcal{D}^{val}} \ell(g_{\tilde{\phi}}(\{f_\theta(x), x\}), z)$. It overcomes the lack of dimensions of the prediction space case, but it will capture situations when there is bias in the data, that is not on the model (see Table 1).

In Figure 2 and in Table 1 we present and compare the different experiments on synthetic data. We say that a method is *Accountable* if the feature attributions identified are the ones that indeed contribute towards the synthetically generated discrimination for both methods.

Table 1: Conceptual table of "Demographic Parity Inspector" 3.2 over the different spaces for the two synthetic examples of Figure 2. The "Accountability" column refers to the ability of the algorithm to provide insights into the sources of discrimination regarding demographic parity violations of the model, and for the protected attribute.

Learning Space	Indirect	Uninform.	Accountable
Input g_Υ	✓	✗	✗
Prediction g_υ	$\sim$	✓	✗
Input + Pred. g_ϕ	✓	✗	✗
Explanations g_ψ	✓	✓	✓

5.2. Use Case: US Income data

In this section, we provide experiments on the US Income data set, [2] which is derived from the US census data (Ding et al., 2021). We divide the dataset into three equal splits $\mathcal{D}^{tr}, \mathcal{D}^{val}, \mathcal{D}^{te} \subseteq \mathcal{D}$ and select our protected attribute, Z, to be a feature that indicates the ethnicity of an individual. We train our model f_β on $\{X^{tr}, Y^{tr}\}$ and, the "Demographic Parity Inspector" g_ψ on $\{\mathcal{S}(f_\beta, X^{val}), Z^{val}\}$. Both methods are evaluated on $\{X^{te}, Z^{te}, y^{te}\}$. For the type of models, f_β, as we focus on tabular data, we choose f_θ to be a `xgboost`(Chen and Guestrin, 2016) that achieve state-of-the-art model performance (Grinsztajn et al., 2022; Elsayed et al., 2021; Borisov et al., 2021), and for the inspector g_ψ a logistic regression. The final explanations are given by the coefficients of the logistic regression.

As the baseline ("no discrimination") w.r.t. the US income data set, we shuffle the protected attribute and fit the "Demographic Parity Inspector" to predict the randomly assigned protected attribute. We then compare pair-wise between the different ethnicity combinations (see Figure3).

For the inspection part, we calculate the Wasserstein distance between the coefficients of the linear regression on the baseline and coefficients of the different pair-wise comparisons. The statistical test between the coefficients is performed within 100 bootstraps of a fraction of 0.632% (Hastie et al., 2001) of both coefficient distributions (random and ethnicity pairs). The AUC of the model f_β is 0.87 while for the DP Inspector, g_ψ it depends on which pairs of ethnicity it is trained. In Figure 3 (left), we can see the performance of the "Demographic Parity Inspector" ranging from values between 0.60 and 0.80. We observe how the "DP Inspector" identifies "Education" as a highly discriminatory proxy while the use of the feature "Worked Hours Per Week" is less discriminatory. In Figure 3 (right), the values correspond to the Wasserstein distance

[2] Please see the ACS PUMS data dictionary for the full list of variables available `https://www.census.gov/programs-surveys/acs/microdata/documentation.html`

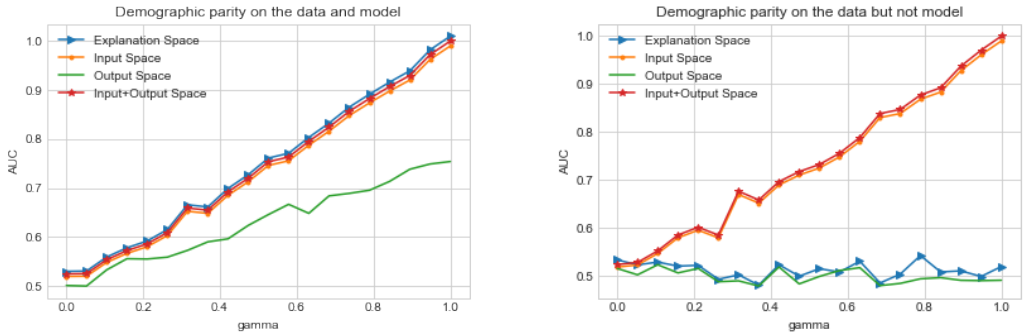

Figure 2: In the left figure, "indirect" experimental case. All spaces capture this fairness violation, only the prediction space is less sensitive due to its dimensionality. In the right figure, "uninformative" experimental case, in this case, only the explanation space, and prediction space detect that the model is non-discriminant. Measuring DP via the explanation space correctly detects false positives (DP in the data, but not on the model), and it is more sensitive to true positives.

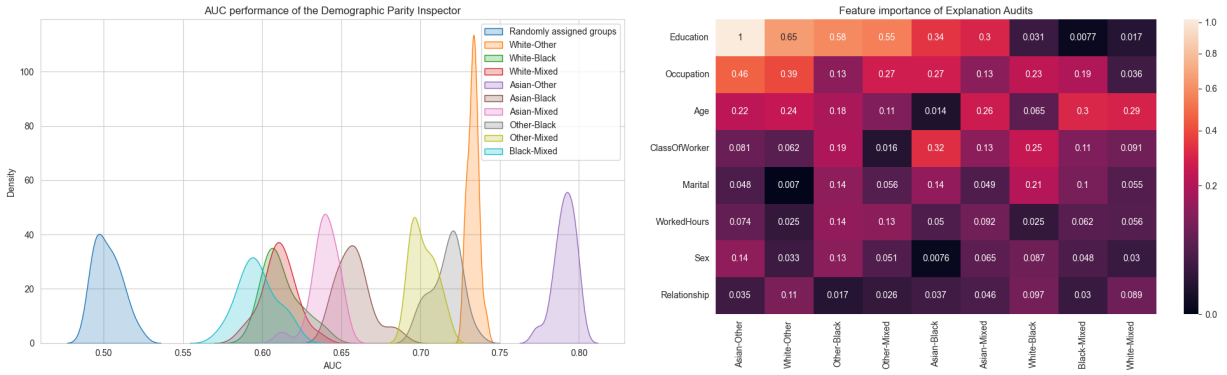

Figure 3: In the left figure the AUC of the "DP Inspector" over different pairs of protected attributes on the US Income dataset. Different pair-wise comparisons achieve different degrees of Demographic Parity Violation. On the right figure, the Wasserstein distance between the coefficients of the model "DP Inspector" between different pairs and the randomly distributed pair. Higher values imply that there is a higher probability that the feature causes demographic parity violations

between the randomly assigned coefficient distributions, and the pairwise comparisons. Higher values imply that the coefficients are more distinct, which suggests that there is a higher probability that these features are the ones causing the demographic parity violation. In the appendix, we provide analysis across different prediction tasks.

6. Conclusions

This work introduces novel measures and inspections of demographic parity fairness in machine learning models. Previous demographic parity measures relied on the model predictions, which lacked sensibility and the possibility to inspect, or, on input data, which is independent of the model.

We have provided theoretical guarantees, synthetic data experiments, and real data cases that show the benefits of our approach, which relies on the explanation space. A promising and unexplored information space that can serve to account for model behaviour.

Limitations: Our work is focused on tabular data. We have used Shapley values to derive the theoretical guarantees, whose values can be distinct from the SHAP (computational approximation). Also, we have used logistic regression as "Demographic Parity Inspector" and exploited its coefficients for accountability, but we believe that other AI explanation techniques such as other feature attribution methods, logical reasoning, argumentation, or counterfactual explanations, may be applicable and come with their own advantages. Furthermore, the usage of fair AI methods does not necessarily guarantee the fairness of AI-based socio-technical systems (Kulynych et al., 2020).

Reproducibility Statement

To ensure the reproducibility of our results, we make the data, data preparation routines, code repositories, and meth-

ods publicly available https://anonymous.4open.science/r/xAIAuditing-1841/README.md. Note that we do not perform any hyperparameter tuning throughout our work unless stated otherwise. Instead, we use default scikit-learn parameters (Pedregosa et al., 2011). We describe the system requirements and software dependencies of our experiments. Our experiments were run on a 4 vCPU server with 32 GB RAM.

Acknowledgments

This work has received funding by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Actions (grant agreement number 860630) for the project: "NoBIAS - Artificial Intelligence without Bias". Furthermore, this work reflects only the authors' view and the European Research Executive Agency (REA) is not responsible for any use that may be made of the information it contains.

References

Aas, K., Jullum, M., and Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artif. Intell.*, 298:103502.

Aka, O., Burke, K., Bäuerle, A., Greer, C., and Mitchell, M. (2021). Measuring model biases in the absence of ground truth. In Fourcade, M., Kuipers, B., Lazar, S., and Mulligan, D. K., editors, *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*, pages 327–335. ACM.

Arrieta, A. B., Rodríguez, N. D., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion*, 58:82–115.

Aumann, R. J. and Dreze, J. H. (1974). Cooperative games with coalition structures. *International Journal of game theory*, 3(4):217–237.

Bala, M., Monica S., V., Horibe, Mayumi, E., J., M., Trevor, A., David E., L., King-Wai, L. D., Shu-Fangand Kim Jerry, N., and Su-In, L. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(1):749–760.

Bandy, J. (2021). Problematic machine behavior: A systematic literature review of algorithm audits. *Proc. ACM Hum. Comput. Interact.*, 5(CSCW1):74:1–74:34.

Barocas, S., Hardt, M., and Narayanan, A. (2019). *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., and Kasneci, G. (2021). Deep neural networks and tabular data: A survey.

Chen, H., Janizek, J. D., Lundberg, S. M., and Lee, S. (2020). True to the model or true to the data? *CoRR*, abs/2006.16234.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Chiappa, S., Jiang, R., Stepleton, T., Pacchiano, A., Jiang, H., and Aslanides, J. (2020). A general approach to fairness with optimal transport. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 3633–3640. AAAI Press.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pages 797–806. ACM.

Datta, A., Sen, S., and Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *IEEE Symposium on Security and Privacy, SP 2016, San Jose, CA, USA, May 22-26, 2016*, pages 598–617. IEEE Computer Society.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. (2021). Retiring adult: New datasets for fair machine learning. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 6478–6490.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S. (2012). Fairness through awareness. In Goldwasser, S., editor, *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pages 214–226. ACM.

Elsayed, S., Thyssens, D., Rashed, A., Schmidt-Thieme, L., and Jomaa, H. S. (2021). Do we really need deep learning models for time series forecasting? *CoRR*, abs/2101.02118.

Fabbrizzi, S., Papadopoulos, S., Ntoutsi, E., and Kompatsiaris, I. (2022). A survey on bias in visual datasets. *Comput. Vis. Image Underst.*, 223:103552.

Finocchiaro, J., Maio, R., Monachou, F., Patro, G. K., Raghavan, M., Stoica, A.-A., and Tsirtsis, S. (2021). Bridging machine learning and mechanism design towards algorithmic fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 489–503.

Frye, C., Rowat, C., and Feige, I. (2020). Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Gao, W. and Zhou, Z. (2015). On the consistency of AUC pairwise optimization. In *IJCAI*, pages 939–945. AAAI Press.

Garg, S., Balakrishnan, S., Kolter, J. Z., and Lipton, Z. C. (2021a). RATT: leveraging unlabeled data to guarantee generalization. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 3598–3609. PMLR.

Garg, S., Balakrishnan, S., Lipton, Z. C., Neyshabur, B., and Sedghi, H. (2021b). Leveraging unlabeled data to predict out-of-distribution performance. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*.

Grabowicz, P. A., Perello, N., and Mishra, A. (2022). Marrying fairness and explainability in supervised learning. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, pages 1905–1916. ACM.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5).

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3315–3323.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Jiang, R., Pacchiano, A., Stepleton, T., Jiang, H., and Chiappa, S. (2019). Wasserstein fair classification. In Globerson, A. and Silva, R., editors, *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*, volume 115 of *Proceedings of Machine Learning Research*, pages 862–872. AUAI Press.

Kamishima, T., Akaho, S., and Sakuma, J. (2011). Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650.

Kearns, M. J., Neel, S., Roth, A., and Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Dy, J. G. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2569–2577. PMLR.

Kleinberg, J. M., Mullainathan, S., and Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In Papadimitriou, C. H., editor, *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*, volume 67 of *LIPIcs*, pages 43:1–43:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Kulynych, B., Overdorf, R., Troncoso, C., and Gürses, S. F. (2020). Pots: protective optimization technologies. In Hildebrandt, M., Castillo, C., Celis, L. E., Ruggieri, S., Taylor, L., and Zanfir-Fortuna, G., editors, *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 177–188. ACM.

Kusner, M. J., Loftus, J. R., Russell, C., and Silva, R. (2017). Counterfactual fairness. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4066–4076.

Liu, J. et al. (2012). The enterprise risk management and the risk oriented internal audit. *International Business Journal*, 10(4):287–292.

Lundberg, S. and Lee, S.-I. (2017). A unified approach to interpreting model predictions.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2019). Explainable ai for trees: From local explanations to global understanding.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):2522–5839.

Manerba, M. M. and Guidotti, R. (2021). Fairshades: Fairness auditing via explainability in abusive language detection systems. In *2021 IEEE Third International Conference on Cognitive Machine Intelligence (CogMI)*, pages 34–43.

Mase, M., Owen, A. B., and Seiler, B. (2019). Explaining black box decisions by shapley cohort refinement. *CoRR*, abs/1911.00467.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

Mittelstadt, B. D., Russell, C., and Wachter, S. (2019). Explaining explanations in AI. In danah boyd and Morgenstern, J. H., editors, *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 279–288. ACM.

Molnar, C. (2019). *Interpretable Machine Learning*. . `https://christophm.github.io/interpretable-ml-book/`.

Mougan, C., Broelemann, K., Kasneci, G., Tiropanis, T., and Staab, S. (2022). Explanation shift: Detecting distribution shifts on tabular data via the explanation space. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*.

Mougan, C. and Nielsen, D. S. (2023). Monitoring model deterioration with explainable uncertainty estimation via non-parametric bootstrap. In *AAAI Conference on Artificial Intelligence*.

Mutlu, E. Ç., Yousefi, N., and Garibay, Ö. Ö. (2022). Contrastive counterfactual fairness in algorithmic decision-making. In Conitzer, V., Tasioulas, J., Scheutz, M., Calo, R., Mara, M., and Zimmermann, A., editors, *AIES '22: AAAI/ACM Conference on AI, Ethics, and Society, Oxford, United Kingdom, May 19 - 21, 2021*, pages 499–507. ACM.

Pearson, K., Lee, A., and Bramley-Moore, L. (1899). Vi. mathematical contributions to the theory of evolution.—vi. genetic (reproductive) selection: Inheritance of fertility in man, and of fecundity in thoroughbred racehorses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, .(192):257–330.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Pedreschi, D., Ruggieri, S., and Turini, F. (2008). Discrimination-aware data mining. In Li, Y., Liu, B., and Sarawagi, S., editors, *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*, pages 560–568. ACM.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. (2020). Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In Hildebrandt, M., Castillo, C., Celis, L. E., Ruggieri, S., Taylor, L., and Zanfir-Fortuna, G., editors, *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 33–44. ACM.

Rodolfa, K. T., Lamba, H., and Ghani, R. (2021). Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence*, 3(10):896–904.

Ruf, B. and Detyniecki, M. (2021). Towards the right kind of fairness in AI. *CoRR*, abs/2102.08453.

Ruggieri, S., Alvarez, J. M., Pugnana, A., State, L., and Turini., F. (2023). Can we trust fair-AI? In *AAAI Conference on Artificial Intelligence*.

Selbst, A. D. and Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham L. Rev.*, 87:1085.

Shapley, L. S. (1953). *A Value for n-Person Games*, pages 307–318. Princeton University Press.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241.

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., and Zafar, M. B. (2018). A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In Guo, Y. and Farooq, F., editors, *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 2239–2248. ACM.

Stevens, A., Deruyck, P., Veldhoven, Z. V., and Vanthienen, J. (2020). Explainability and fairness in machine learning: Improve fair end-to-end lending for kiva. In *2020 IEEE Symposium Series on Computational Intelligence, SSCI 2020, Canberra, Australia, December 1-4, 2020*, pages 1241–1248. IEEE.

Strumbelj, E. and Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.*, 41(3):647–665.

Sundararajan, M. and Najmi, A. (2020). The many shapley values for model explanation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9269–9278. PMLR.

Vecchione, B., Levy, K., and Barocas, S. (2021). Algorithmic auditing and social justice: Lessons from the history of audit studies. In *EAAMO*, pages 19:1–19:9. ACM.

Wachter, S., Mittelstadt, B., and Russell, C. (2020). Bias preservation in machine learning: the legality of fairness metrics under eu non-discrimination law. *W. Va. L. Rev.*, 123:735.

Winter, E. (2002). Chapter 53 the shapley value. In ., volume 3 of *Handbook of Game Theory with Economic Applications*, pages 2025–2054. Elsevier.

Yang, K., Qinami, K., Fei-Fei, L., Deng, J., and Russakovsky, O. (2020). Towards fairer datasets: filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In Hildebrandt, M., Castillo, C., Celis, L. E., Ruggieri, S., Taylor, L., and Zanfir-Fortuna, G., editors, *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 547–558. ACM.

Yule, G. U. (1900). Vii. on the association of attributes in statistics: with illustrations from the material of the childhood society, &c. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 194(252-261):257–319.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact. *Proceedings of the 26th International Conference on World Wide Web*.

Zern, A., Broelemann, K., and Kasneci, G. (2023). Interventional shap values and interaction values for piecewise linear regression trees. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In Palmer, M., Hwa, R., and Riedel, S., editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2979–2989. Association for Computational Linguistics.

A. True to the model or true to the data?

The "Demographic Parity Inspector" proposed in this work relies on the explanation space that satisfies efficiency and uninformative theoretical properties. We have used the Shapley values as an explainable AI method that satisfies these properties. In recent times a variety of papers discuss the application of Shapley values, for feature attribution in machine learning models (Strumbelj and Kononenko, 2014; Lundberg et al., 2020; Lundberg and Lee, 2017; Bala et al., 2018). However, the correct way to connect a model to a coalitional game, which is the central concept of Shapley values, is a source of controversy, with two main approaches (i) an interventional (Aas et al., 2021; Frye et al., 2020; Zern et al., 2023) or (ii) an observational formulation of the conditional expectation (cf. Section 2)(Sundararajan and Najmi, 2020; Datta et al., 2016; Mase et al., 2019).

In the following experiment, we compare what are the differences between estimating the Shapley values using one or the other approach. We benchmark this experiment on the four prediction tasks based on the US census data (Ding et al., 2021) and using the "Demographic Parity Inspector", where both the model $f_\theta(X)$ and $g_\psi(\mathcal{S}(f_\theta, X))$ are linear models. We will calculate the Shapley values using the SHAP linear explainer. [3]

The comparison depends on a feature perturbation hyperparameter: whether the approach to compute the SHAP values is either *interventional* or *correlation dependent*. The interventional SHAP values break the dependence structure between features in the model to uncover how the model would behave if the inputs are changed (as it was an intervention). This option is said to stay "true to the model" meaning it will only give allocation credit to the features that are actually used by the model.

On the other hand, the full conditional approximation of the SHAP values respects the correlations of the input features. If the model depends on one input that is correlated with another input, then both get some credit for the model's behaviour. This option is said to say "true to the data", meaning that it only considers how the model would behave when respecting the correlations in the input data (Chen et al., 2020).

In our case, we will measure the difference between the two approaches by looking at the linear coefficients of the model g_ψ and comparing the performance of predicting protected attributes, for this case only between White-Other.

Table 2: AUC comparison of the "Explanation Shift Detector" between estimating the Shapley values between the interventional and the correlation-dependent approaches for the four prediction tasks based on the US census dataset (Ding et al., 2021). The $\%$ character represents the relative difference. The performance differences are negligible.

	Interventional	Correlation	$\%$
Income	0.736438	0.736439	1.1e-06
Employment	0.747923	0.747923	4.44e-07
Mobility	0.690734	0.690735	8.2e-07
Travel Time	0.790512	0.790512	3.0e-07

Table 3: Linear regression coefficients comparison of the "Explanation Shift Detector" between estimating the Shapley values between the interventional and the correlation-dependent approaches for one of the US census based prediction tasks(ACS Income). The $\%$ character represents the relative difference. The coefficients show negligible differences between the calculation methods

	Interventional	Correlation	$\%$
Marital	0.348170	0.348190	2.0e-05
Worked Hours	0.103258	-0.103254	3.5e-06
Class of worker	0.579126	0.579119	6.6e-06
Sex	0.003494	0.003497	3.4e-06
Occupation	0.195736	0.195744	8.2e-06
Age	-0.018958	-0.018954	4.2e-06
Education	-0.006840	-0.006840	5.9e-07
Relationship	0.034209	0.034212	2.5e-06

[3] `https://shap.readthedocs.io/en/latest/generated/shap.explainers.Linear.html`

In Table 2 and Table 3 we can see the comparison of the effects of using the aforementioned approaches to learn our proposed method, the "Explanation Shift Detector". Even though the two approaches differ theoretically, the differences become negligible when explaining the protected characteristic, i.e. when providing the linear regression coefficients.

B. Experiments on datasets derived from the US Census

In this section we apply our approach to more datasets derived from the US census database. We discuss what are the differences on three (plus the one in the main body of the paper) prediction tasks based on the US census data: ACS Income (main body), ACS Travel Time, ACS Employment and ACS Mobility.

We follow the same methodology as in the experimental section. Where we divide a dataset into three equal splits $\{\mathcal{D}^{tr}, \mathcal{D}^{val}, \mathcal{D}^{te}\} \subseteq \mathcal{D}$ and select our protected attribute, Z, to be a feature that indicates the ethnicity of an individual. We train our model f_β on $\{X^{tr}, Y^{tr}\}$ and, the "Demographic Parity Inspector" g_ψ on $\{\mathcal{S}(f_\beta, X^{val}), Z^{val}\}$. Both methods are evaluated on $\{X^{te}, Z^{te}, y^{te}\}$. For the type of models, f_β, as we are in tabular data we focus on we choose f_θ to be a `xgboost`(Chen and Guestrin, 2016) that achieve state-of-the-art model performance (Grinsztajn et al., 2022; Elsayed et al., 2021; Borisov et al., 2021),and for the inspector g_ψ a logistic regression. The final explanations are given by the coefficients of the logistic regression.

B.1. ACS Employment

The goal of this task is to predict whether an individual, is employed. In this section we are going to do a comparison within different protected attribute groups for CA14.

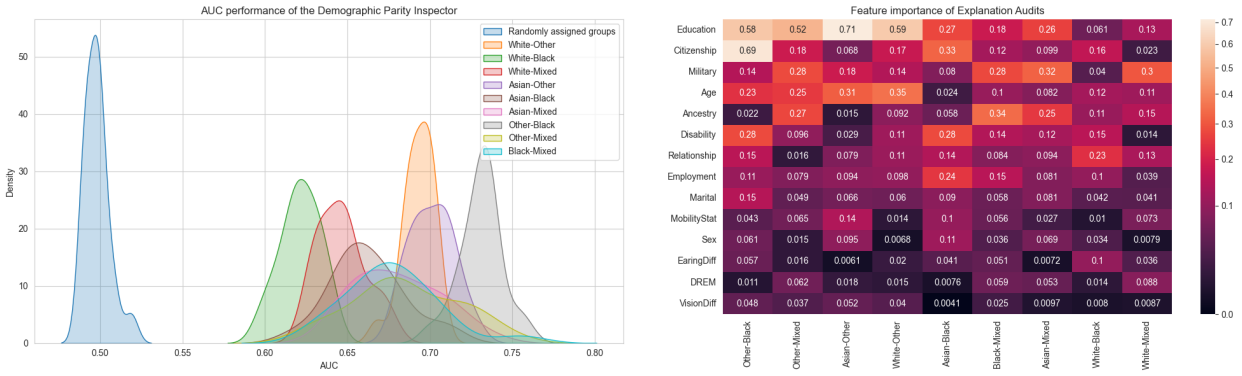

Figure 4: In the left figure, comparison of the performance of *Demographic Parity Inspector*, over the state of CA14 for the ACS Employment dataset. For this dataset most of different protected attributes comparison have different AUC distributions in terms of mean and variance. The group that receives the most comparison is Other-Black. This difference in the model behaviour is due to features such as "Education" and "Citizenship". Interestingly, features such as "difficulties on the hearing or seeing", do not play a role in the OOD model behaviour.

Compared to other prediction tasks based on the US census dataset, the measured domographic parity violation for this task is smallest. The AUC of the "Demographic Parity Inspector" is ranging from 0.55 to 0.70. For Asian-Black demographic parity violation we see that there is a lot of variation of the AUC, indicating that the method achieves different values on the bootstrapping folds. If we look for the features driving the demographic parity gap, we see particularly high values when comparing "Asian" and "Black" populations, and for features "Citizenship" and "Employment". On average, the most important features across all group comparisons are also "Education" and "Area".

B.2. ACS Travel Time

The goal of this task is to predict whether an individual has a commute to work that is longer than 20 minutes. The threshold of 20 minutes was chosen as it is the US-wide median travel time to work based on 2018 data

In general, the values of the AUC of the "Demographic Parity Inspector" are in the range of 0.50 to 0.60. By inspecting the features we can see that our proposed methods highlight different demographic parity violation drivers depending on the

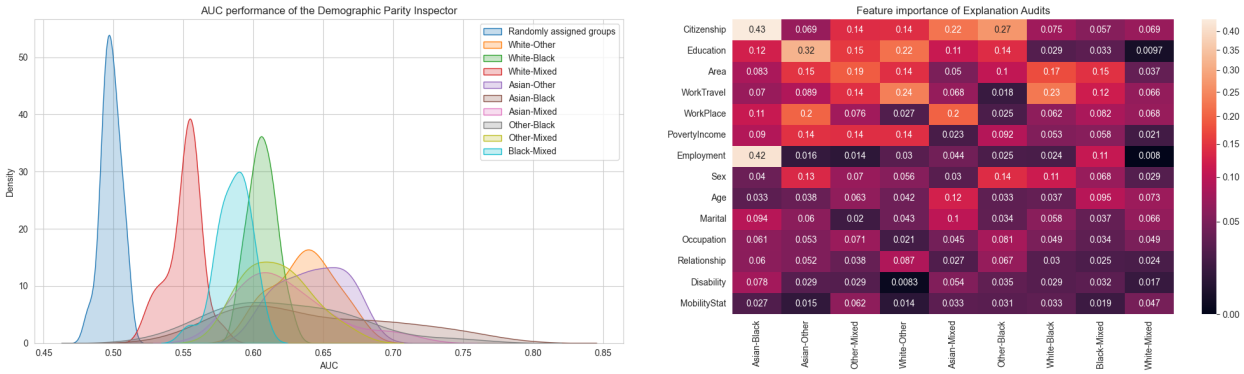

Figure 5: In the left figure, comparison of the performance of *Demographic Parity Inspector*, in different pair-wise ethnicities for the ACS Travel Time dataset. In the right image, the features that are more probable to be originating the demographic parity violation (higher implies more probable).

pair-wise comparison made. In general, the feature "Education", "Citizenship" and "Area" are the features with the highest difference. Even though for Asian-Black pairwise comparison "Employment" is also one of the most relevant features.

B.3. ACS Mobility

The goal of this task is to predict whether an individual had the same residential address one year ago, only including individuals between the ages of 18 and 35. This filtering increases the difficulty of the prediction task, as the base rate of staying at the same address is above 90% for the general population (Ding et al., 2021).

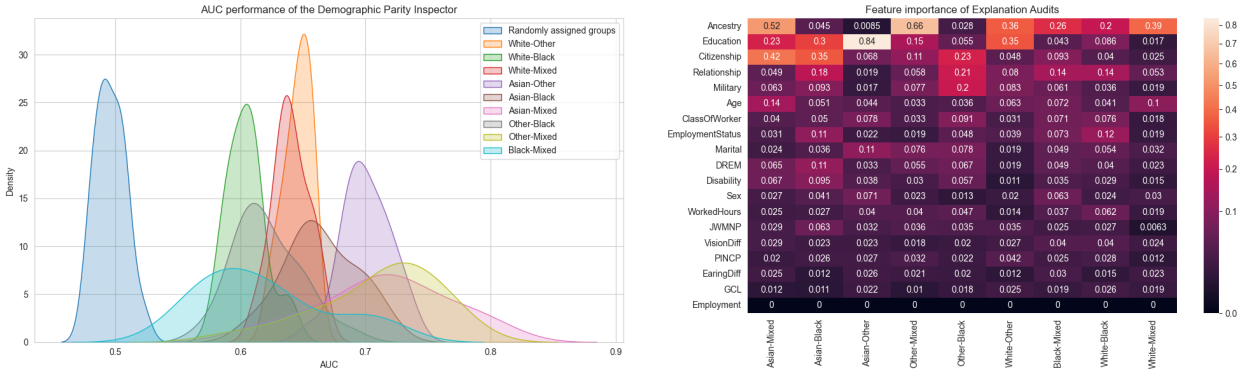

Figure 6: In the left figure, comparison of the performance of *Demographic Parity Inspector*, in different pair-wise ethnicities for the ACS Mobility dataset. In the right image, the features that are more probable to be originating the demographic parity violation (higher implies more probable).

In general, the values of the AUC of the "Demographic Parity Inspector" are in the range of 0.55 to 0.80. By inspecting the features we can see that our proposed methods highlight different demographic parity violation to be the source of the demographic parity violation depending on the ethnic pair-wise comparison selected. In general the feature "Ancestry", i.e. "ancestors' lives with details like where they lived, who they lived with, and what they did for a living", plays a high relevance when predicting demographic parity violation in the mobility prediction task. In other prediction tasks such as when predicting employment (cf Figure 4), even if the input data distribution across the protected groups is similar, the weight that the "Demographic Parity Inspector" assigns is distinct, this is due to our method measuring demographic parity violations with respect to the model not just with respect to the data.