ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230306

Counterfactual Reasoning for Bias Evaluation and Detection in a Fairness Under Unawareness Setting

Giandomenico Cornacchia[a;*], Vito Walter Anelli[a], Fedelucio Narducci[a], Azzurra Ragone[b] and Eugenio Di Sciascio[a]

[a]Polytechnic University of Bari
[b]Università degli Studi di Bari Aldo Moro
ORCiD ID: Giandomenico Cornacchia https://orcid.org/0000-0001-5448-9970, Vito Walter Anelli https://orcid.org/0000-0002-5567-4307, Fedelucio Narducci https://orcid.org/0000-0002-9255-3256, Azzurra Ragone https://orcid.org/0000-0002-3537-7663, Eugenio Di Sciascio https://orcid.org/0000-0002-5484-9945

Abstract. Current AI regulations require discarding sensitive features (e.g., gender, race, religion) in the algorithm's decision-making process to prevent unfair outcomes. However, even without sensitive features in the training set, algorithms can persist in discrimination. Indeed, when sensitive features are omitted (*fairness under unawareness*), they could be inferred through non-linear relations with the so-called proxy features. In this work, we propose a way to reveal the potential hidden bias of a machine learning model that can persist even when sensitive features are discarded. This study shows that it is possible to unveil whether the black-box predictor is still biased by exploiting counterfactual reasoning. In detail, when the predictor provides a negative classification outcome, our approach first builds counterfactual examples for a discriminated user category to obtain a positive outcome. Then, the same counterfactual samples feed an external classifier (that targets a sensitive feature) that reveals if the modifications to the user characteristics needed for a positive outcome moved the individual to the non-discriminated group. When this occurs, it could be a warning sign for discriminatory behavior in the decision process. Furthermore, we leverage the deviation of counterfactuals from the original sample to determine which features are proxies of specific sensitive information. Our experiments show that, even if the model is trained without sensitive features, it often suffers discriminatory biases.

1 Introduction

The use of AI systems that deal with life-changing tasks is increasing daily, raising social concerns. One of the most worrisome is the *discrimination* of minority groups or individuals. As an example, in the financial domain, the decision to approve or deny credit has been regulated with precise and detailed regulatory compliance requirements (i.e., Equal Credit Opportunity Act , Federal Fair Lending Act , and Consumer Credit Directive for EU Community).These rules aim to prevent discrimination in human decision-making processes. However, they do not fit scenarios involving Machine Learning (ML) or, more broadly, Artificial Intelligence (AI) systems. In the wake of the GDPR (i.e., a regulatory oversight to protect personal information), the EU Commission has issued "Ethics Guidelines for Trustworthy AI" and, more recently, "The Proposal for Harmonized Rule on AI" to prevent irresponsible uses of such technology. Several studies have shown that deploying AI systems without considering ethical concerns might encourage discrimination, although system designers train their models without discriminatory intent [9, 17, 4, 10, 11]. Moreover, omitting sensitive features to train AI models does not guarantee the absence of biases in the outcomes [1, 37, 19]. Indeed, there may exist *proxy features*, which behave as implicit sensitive feature (e.g., I can infer the gender from the individual's habits).

This study proposes an approach for detecting biased decisions in a realistic scenario where sensitive features are omitted and an unknown number of proxy features may be present. Our designed pipeline is composed of three main modules. The **Decision Maker** is the first module, which wraps the critical classifier to analyze, and is trained without using the sensitive features (e.g., it decides to grant or not a loan). The second module trains a different classifier, named **Sensitive-Feature Classifier**, on the same features to predict the sensitive characteristics (e.g., it classifies if the individual involved in the previous decision is a member of the protected or non protected group). The third module, the **Counterfactual generator**, calculates the minimal counterfactual samples by modifying the values of non-sensitive features to obtain the desired outcome on the Decision maker (e.g., loan approved). Eventually, the sensitive feature predictor classifies the generated counterfactual samples to check whether the samples do still belong to the original sensitive class (e.g., gender female). If this does not occur (e.g., the new counterfactual sample obtaining the loan is classified now as male, opposite to the original class), the outcome predictor is biased, and its unfairness can be quantified.

Let us provide a real-life example to explain our model:

Example 1. Anna is a young researcher who wants to buy a house and decides to apply for a loan. She got a permanent contract two months ago, has a good income, and likes cinema, open-air sports, and visiting museums. The bank AI system analyzed Anna's financial profile, non-sensitive information, bank account transactions, and income data and decided to deny the loan. At this point, our system comes into play to discover whether the decision is affected

* Corresponding Author. Email: giandomenico.cornacchia@poliba.it

by bias. It acquires all of Anna's non-sensitive data exploited by the AI bank system and starts to generate counterfactual samples. Consequently, Anna's profile is slightly changed regarding employment-contract duration, monthly income, and recurrent transactions until the new counterfactual profile gets a favorable decision by the AI system and the loan is approved. Anna's new profile is now used to feed the sensitive-feature classifier. The sensitive classifier decides that Anna's original profile belongs to the female class, but the new counterfactual profile belongs to the male class. This demonstrates decision bias since, even though the system does not exploit sensitive features and does not know Anna's gender, it classifies Anna's counterfactual profile (who gets the loan) as belonging to the (privileged) male class.

To summarize, we propose an approach for detecting bias in machine learning models making use of counterfactual reasoning, even when those models are trained without sensitive features, i.e., the setting known as *Fairness Under Unawareness* [8, 29]. In detail, our work aims to answer the following research questions:

- **RQ1:** Is there a method for determining whether a dataset contains proxy features or not?
- **RQ2:** Does the Fairness Under Unawareness setting ensure that decision biases are avoided?
- **RQ3:** Is counterfactual reasoning effective for discovering decision biases?
- **RQ4:** Is it possible to define a strategy for identifying the proxy features?

The remainder of the paper is organized as follows: Section 2 provides an overview of the most relevant research in the fields of fairness and counterfactual reasoning, Section 3 introduces the preliminaries, Section 4 depicts our methodology, Section 5 describes the experimental evaluation, Section 6 discusses the results, and the conclusion and future work are drawn in Section 7. **Code is available at:** https://github.com/giandos200/ECAI23.

2 Related Work

The section illustrates the most significant studies on *fairness under unawareness* and *counterfactual reasoning* research.

Fairness, Fairness Under Unawareness, and Proxy Features. In machine learning research, fairness is a well-studied topic with a considerable body of knowledge to draw from [2, 35, 14]. In fact, due to the critical impact of decision-making on people's lives (e.g., Financial Services, Education, and Health), the use of sensitive characteristics is strictly prohibited. While designing the decision-making algorithm not to leverage sensitive information is simple, assuring the same accuracy as before and demonstrating that the predictor is unbiased, despite the absence of such sensitive characteristics, is another matter [7, 12, 13].

The system may infer sensitive features from variables, i.e. proxy variables, that nonetheless represent protected characteristics, even if the user does not explicitly give sensitive information. The models that infer sensitive features from proxy variables are known as "probabilistic proxy models" [8, 6]. The majority of the approaches for identifying proxy features proposed in the literature rely on strategies for determining multicollinearity between variables [1, 39]. The relationships, however, may not be linear. Cosine similarity and mutual information are the most commonly employed methods in this scenario, according to Agarwal and Mishra [1]. Chen et al. [8] studied the relationship between proxy features, sensitive variables (i.e., geolocation and race), and classifiers threshold. Fabris et al. [20] use a quantification approach to measure group fairness in an unawareness setting.

Counterfactual Reasoning. Counterfactual reasoning is a flourishing field in artificial intelligence research [22, 30, 13]. This research was initially born to investigate causal links [34], and today it can count on several contributions [21]. Most of them define and employ counterfactuals as helpful tools to explain the decisions taken by modern decision support systems. The study that takes the first steps from the same motivations was conducted by Kusner et al. [28] which proposed a Counterfactual metric. The metric exploits causal inference to assess fairness at an individual level by requiring that a sensitive attribute not be the cause of a change in a prediction. In the same direction, Wu et al. [38] propose a bounded linear constrained optimization of countefactual faireness addressing the limitation of Kusner et al. [28] work. For what concerns the risk assessment domain, Mishler et al. [31] put forward a similar working hypothesis. In detail, they propose a counterfactual equalized odds ratio criterion to train predictors operating in the post-processing phase. They extend and adapt previous post-processing approaches [23] to the counterfactual setting.

Our Contribution Differently from the majority of the recent studies, our investigation aims to leverage a counterfactual generation tool to reveal the presence of implicit biases in a decision support system seen as a black box. Interestingly, this motivation is similar to Bottou et al. [5]. Indeed, we both aim to answer the question: "How would the system have decided if we had replaced some user characteristics?". The paper focuses on exploiting possible system behavior changes in real-life applications of minimal intervention. In contrast, our contribution focuses on the same work philosophy but on exploiting possible unethical consequences of machine learning systems. Beyond this commonality, the two studies differ significantly as Bottou et al. [5] focus on measuring the fidelity level of the system.

3 Preliminaries

This section introduces the notation adopted hereinafter.
Data points: We assume the dataset $\mathcal{D}$ is an m-dimensional space containing n non-sensitive features, l sensitive features, and a target attribute. In other words, we have $\mathcal{D} \subseteq \mathbb{R}^m$, with $m = n + l + 1$.[1] A data point $d \in \mathcal{D}$ is then represented as $d = \langle \mathbf{x}, \mathbf{s}, y \rangle$, with $\mathbf{x} = \langle x_1, x_2, ..., x_n \rangle$ representing the sub-vector of non-sensitive features, $\mathbf{s} = \langle s_1, s_2, ..., s_l \rangle$ the sub-vector of sensitive features and y being a binary target feature. Given a vector of sensitive festures, $\forall s_i \in \mathbf{s}$, s_i^- refers to the *unprivileged* group and s_i^+ to the *privileged* group of the i-th sensitive feature.
Proxy Features: $\mathbf{p}_i \subseteq \mathbf{x}$ is a subset of features for a sensitive feature s_i such that there exists a function $h(\cdot)$ which can estimate s_i (i.e., $h(\mathbf{p}_i) = s_i$).
Target Labels: Given a target feature $y \in \{0, 1\}$, $y = 1$ is the positive outcome and $y = 0$ is the negative one.
Outcome Prediction: $\hat{y} \in \{0, 1\}$ represents the prediction for $\mathbf{x} \subset d$ estimated by $f(\cdot)$, a function such that $f(\mathbf{x}) = \hat{y}$.

[1] Without loss of generality, we assume that categorical features can always be transformed into features in $\mathbb{R}$ via one-hot-encoding.

Sensitive Feature Prediction: $\hat{s}_i \in \{-,+\}$ represents the prediction of the i-th sensitive feature for a given data point estimated by $f_{s_i}(\cdot)$, a function s.t. $f_{s_i}(\mathbf{x}) = \hat{s}_i$.
Counterfactual samples: Given a vector $\mathbf{x}$ and a perturbation $\epsilon = \langle \epsilon_1, \epsilon_2, ..., \epsilon_n \rangle$, we say that a vector $\mathbf{c_x} = \langle c_{x_1}, c_{x_2}, ..., c_{x_n} \rangle = \mathbf{x}+\epsilon$ is a counterfactual (CF) of $\mathbf{x}$ if $f(\mathbf{c_x}) = 1-f(\mathbf{x}) = 1-\hat{y}$. We use the set $\mathcal{C}_\mathbf{x}$, with $|\mathcal{C}_\mathbf{x}| = k$, to denote the set of possible **counterfactual samples** for $\mathbf{x}$. A function $g(\mathbf{x})$ is used to compute k counterfactuals for $\mathbf{x}$.
For simplicity, we denote $f(\cdot)$, $f_{s_i}(\cdot)$, and $g(\mathbf{x})$ as the **Decision Maker**, the **Sensitive-Feature Classifier**, and the **Counterfactual Generator** respectively.

4 Methodology

The *fairness under unawareness* setting (see Section 2) poses several challenges to the identification of discriminatory behaviors performed by intelligent systems. Proxy traits can be non-linearly associated with sensitive ones, making typical statistical procedures ineffective. Figure 1 depicts the principal components of our model, namely the *Decision-Maker*, the *Counterfactual Generator*, and the *Sensitive-Feature Classifier*, as well as the flow of our pipeline. As a relevant case study, we refer to the financial domain, considering the tasks of predicting loan-repayment default. However, focusing on this specific domain does not compromise the model generality.

Example 2. Let us imagine that some users with certain characteristics apply for a loan (see Figure 1). The *Decision Maker* analyzes their requests and computes a positive or negative decision. If a request is denied (e.g., for users 2 and 3), the *counterfactual generator* starts to produce a series of counterfactuals until it gets a positive decision ($\hat{y} = 1$, loan accepted). In the example, only two counterfactuals for users 2 and 3 are generated. Once the decision has been changed, the *Sensitive-Feature Classifier* analyzes the original characteristics of users' 2 and 3 profiles and the newly generated counterfactuals to assess how many counterfactuals changed the sensitive-feature *gender*. User 2 was originally classified as female and, then, as male for both counterfactuals (profile with counterfactual changes for getting loan approval). For user 3, this does not happen, the gender classification is the same before and after the counterfactual changes (loan approved).

To define a discrimination score of a given decision model, we propose a metric that we call *Counterfactual Flips* providing a snapshot of the discriminatory behavior the model might put in place.

Definition 1 (*Counterfactual Flips*)**.** *Given a sample* $\mathbf{x}$ *belonging to a demographic group* s *whose model output is denoted as* $f(\mathbf{x})$*, generated a set* $\mathcal{C}_\mathbf{x}$ *of* k *counterfactuals with a desired* y^* *outcome* $f(\mathbf{c}^i_\mathbf{x}) = y^* \quad \forall \mathbf{c}^i_\mathbf{x} \in \mathcal{C}_\mathbf{x}$*, the Counterfactual Flip indicates the percentage of counterfactual samples belonging to another demographic group (i.e.,* $f_s(\mathbf{c}^i_\mathbf{x}) \neq f_s(\mathbf{x})$*, with* $f_s(\mathbf{x}) = s$*).*

$$\text{CFlips}(\mathbf{x}, \mathcal{C}_\mathbf{x}, f_s(\cdot)) = \frac{\sum_{i=1}^{k} (\mathbb{1}(\mathbf{c}^i_\mathbf{x}))}{k} \tag{1}$$

where the function $\mathbb{1}(\mathbf{c}^i_\mathbf{x})$ correspond to:

$$\mathbb{1}(\mathbf{c}^i_\mathbf{x}) = \begin{cases} 1 & \text{if } f_s(\mathbf{c}^i_\mathbf{x}) \neq f_s(\mathbf{x}) \neq s \\ 0 & \text{if } f_s(\mathbf{c}^i_\mathbf{x}) = f_s(\mathbf{x}) = s \end{cases} \tag{2}$$

and it measures if a counterfactual sample is linked with a *Flip*, i.e. a change, for the sensitive characteristic.

The bigger the CFlips value is, the stronger the biases and the discrimination the model suffers from. In Example 2, $CFlips = 1$ for user 2, and $CFlips = 0$ for user 3, thus the sensitive classification changed 2 out of 2 CF samples for user 2 and 0 out of 2 CF samples for user 3, unveiling implicit bias in the *Decision Maker* in favor of male characteristics. Therefore, we can introduce the proposed *Counterfactual Fair Flips* desiderata.

Definition 2 (Counterfactual Fair Flips)**.** -A binary classifier shows Counterfactual Fair Flips if the probability of generating Counterfactual samples belonging to a different demographic group (privileged vs unprivileged) is the same:

$$\mathbb{P}(f_s(\mathcal{C}_{\mathcal{D}|_{s^-}}) \neq s^- \mid f(\mathcal{C}_{\mathcal{D}|_{s^-}})) = \mathbb{P}(f_s(\mathcal{C}_{\mathcal{D}|_{s^+}}) \neq s^+ \mid f(\mathcal{C}_{\mathcal{D}|_{s^+}})) \tag{3}$$

which implies the complement:

$$\mathbb{P}(f_s(\mathcal{C}_{\mathcal{D}|_{s^-}}) = s^- \mid f(\mathcal{C}_{\mathcal{D}|_{s^-}})) = \mathbb{P}(f_s(\mathcal{C}_{\mathcal{D}|_{s^+}}) = s^+ \mid f(\mathcal{C}_{\mathcal{D}|_{s^+}})) \tag{4}$$

In our work, we only take into account samples negatively predicted by the *Decision Maker* (i.e., $f(\mathbf{x}) = 0$), that we denote as $\mathcal{X}^- \subseteq \mathcal{D}$, as we are interested in quantifying the discrimination of the minority group (e.g. women) in the process to achieving a positive counterfactual result (i.e., $f(\mathbf{c_x}) = 1 \wedge f_s(\mathbf{c_x}) \neq s$). For the set of samples $\mathcal{X}^-$, the metric in Equation 1 can be generalized to the *privileged* and *unprivileged* group (Equation 5 is restricted to the *privileged* samples negatively predicted, while Equation 6 is restricted to the *unprivileged* samples negatively predicted).

$$\textit{Privileged} \quad \begin{cases} \text{CFlips} = \frac{\sum_{i=1}^{n} \text{CFlips}(\mathbf{x_i}, \mathcal{C}_{\mathbf{x_i}}, f_s(\cdot))}{|\mathcal{X}|^-_{s^+}|} & \text{with } \mathbf{x} \in \mathcal{X}|^-_{s^+} \end{cases} \tag{5}$$

$$\textit{Unprivileged} \quad \begin{cases} \text{CFlips} = \frac{\sum_{i=1}^{n} \text{CFlips}(\mathbf{x_i}, \mathcal{C}_{\mathbf{x_i}}, f_s(\cdot))}{|\mathcal{X}|^-_{s^-}|} & \text{with } \mathbf{x} \in \mathcal{X}|^-_{s^-} \end{cases} \tag{6}$$

Definition 2 requires that the quantities of counterfactuals flipped in equations 5 and 6 must be equal. Thus, we are interested in the difference between the result of the two equations, i.e. ΔCFlips, being close to zero.

5 Experimental Design

This section provides the description of our experimental settings, designed to answer the four RQs outlined in Section 1.

Table 1: Adult, Adult-debiased, Crime, and German datasets' characteristics. $|n|^*$ correspond to the number of non-sensitive features, and $|n|^\dagger$ to the number of feature w/o the sensitive one (i.e., *gender* and *crime*).

Dataset	Split	$\lvert\mathcal{D}\rvert$	$\lvert n\rvert^\dagger$	Target (Y)	$Y=1$
Adult	**Train**	40699	13	income	$\geq \$50,000$
	Test	4523	13	income	$\geq \$50,000$
Adult-debiased	**Train**	40699	6	income	$\geq \$50,000$
	Test	4523	6	income	$\geq \$50,000$
Crime	**Train**	1794	98	Violent State	<*median*
	Test	200	98	Violent State	<*median*
German	**Train**	900	17	credit score	Good
	Test	100	17	credit score	Good

Datasets. Experiments have been conducted on three state-of-the-art (SOTA) datasets, used as benchmarks in several works [3, 16, 35,

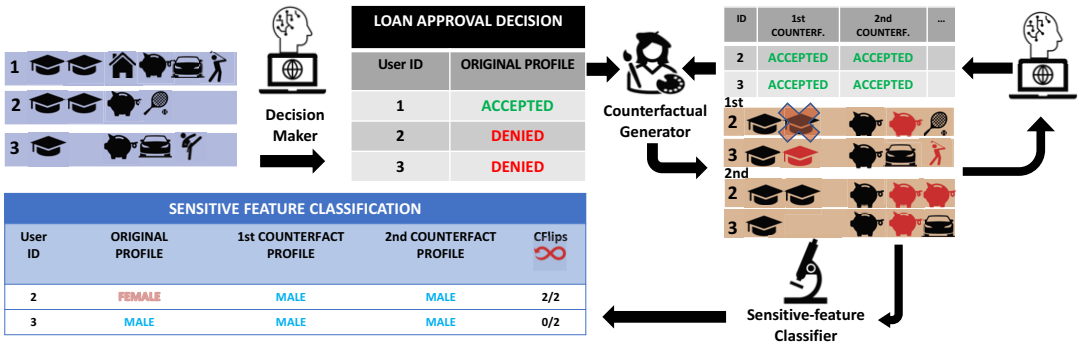


Figure 1: An example of a loan-approval decision process analyzed through our model. From left to right, we have the decision made on the original user profiles, the counterfactual generation for users with loan denied, the sensitive-feature classification of the original profile, and counterfactual profiles with the decision changed. For user 2, both counterfactual profiles change the sensitive-feature category.

15]. These are: Adult [27], a real-world dataset used for income prediction,[2] German [24], a real-world dataset for default prediction,[3] and Crime [36], a real-world Census dataset for violent state prediction[4] (i.e., a state is violent if the number of crime in a state is higher with respect to the median($|\mathcal{C}_{\mathbf{x}}|$) of all the states). We point to the reader attention that the privileged and unprivileged have been chosen based on the ex-ante Statistical Parity in Table 2 (i.e., if positive, the first group is the *privileged*, otherwise the second). For Adult and German, the sensitive attribute we considered is *gender*, with *male* and *female* corresponding to the *privileged* and *unprivileged* group respectively. For the Crime dataset, the sensitive attribute we considered is the *race* that indicates the race with the largest number of crimes committed in a specific state. As a second sensitive feature, for Adult, we chose *maritalStatus*, with *married* and *not married* as *privileged* and *unprivileged*, and for German, we chose age as > 25 *years* and $<= 25$ *years* as *privileged* and *unprivileged* [25]. In this dataset, each sample consists of the name of the *state* and the number of crimes associated with each race, i.e., *white*, *black*, *asian*, and *hispanic*. For our task, we split races into two groups *White* and *Others* where *Others* groups the crimes of *Black*, *Asian*, and *Hispanic* races [3]. The *privileged* group is the *White* one, and the *unprivileged* is *Others* (i.e., Blacks, Asians, and Hispanics).

Regarding the Adult dataset, we decided to create two settings: (a) - *Adult*: the original dataset where we only discarded the sensitive features *gender* and *marital-status*; (b) - *Adult-debiased*: where we remove all the sensitive features (i.e., gender, age, marital status, and race), and all the features highly correlated with at least one of the sensitive features.[5] As regards the non-sensitive features used for training the models, we used: *education num*, *occupation*, *work class*, *capital gain*, *capital loss*, *hours per week*. Furthermore, the feature *work class* has been condensed into three classes: *Private*, *Public*, and *Unemployed*. We replaced the categories in *work class Private*, *SelfEmpNotInc*, *SelfEmpInc*, with *Private*, the categories *FederalGov*, *LocalGov*, *StateGov*, with *Private*, and the category *WithoutPay* with *Unemployed*. This diversification ensures coherence with the *fairness under unawareness* setting and makes possible comparisons with completely biased approaches. As for the Adult dataset, German contains other sensitive characteristics (i.e., age and race) that we do not include for learning the model to guarantee the *fairness under awareness* setting. Additional information on the datasets, target distribution, sensitive-feature distribution, and ex-ante Statistical Parity are available in Table 1 and Table 2.

[2] Adult: https://archive.ics.uci.edu/ml/datasets/adult
[3] German: https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
[4] Crime: https://archive.ics.uci.edu/ml/datasets/US+Census+Data+(1990)
[5] Both Spearman and Pearson's correlation coefficient greater than 0.35.

Table 2: Overview of relevant dataset information, including sensitive feature distribution, target distribution, name of privileged group, and ex-ante Statistical parity respectively for the Adult, Adult-debiased, German, and Crime datasets.

Dataset	s	privileged (s^+)	$\Phi(s)^\dagger$	$\Phi(Y)^{\dagger\dagger}$	ex-ante SP*
Adult	*gender*	*male*	0.68/0.32	0.25/0.75	0.199
	maritalStatus	*married*	0.48/0.52	0.25/0.75	0.378
Adult-deb.	*gender*	*male*	0.68/0.32	0.25/0.75	0.199
	maritalStatus	*married*	0.48/0.52	0.25/0.75	0.378
Crime	*race*	*white*	0.58/0.42	0.50/0.50	0.554
German	*gender*	*male*	0.69/0.31	0.70/0.30	0.075
	age	> 25 *year*	0.81/0.19	0.70/0.30	0.149

† Probability distribution of the *privileged* and *unprivileged* group: $\mathbb{P}(s^+)/\mathbb{P}(s^-)$
†† Probability distribution of the target variable: $\mathbb{P}(y=1)/\mathbb{P}(y=0)$
* A priori Statistical Parity, based on Independence criteria: $\mathbb{P}(y=1 \mid s^+) - \mathbb{P}(y=1 \mid s^-)$

Decision Maker. To keep the approach as general as possible, we have chosen seven largely adopted learning models to handle the classification task and implement the *Decision Maker*. In detail, we used: Logistic Regression (LR), Decision Tree (DT), Support-Vector Machines (SVM), LightGBM (LGBM), XGBoost (XGB), Random Forest (RF), and Multi-Layer Perceptron (MLP). We completed our analysis with three *in-processing* debiasing algorithms, Linear Fair Empirical Risk Minimization (LFERM) [16], Adversarial Debiasing (Adv) [41], and Fair Classification (FairC) [40]. Those debiasing models have not been casually chosen, but they are consistent with the *fairness under unawareness* setting. Indeed, those models can not use sensitive features in the inference phase and can be trained only on non-sensitive features (i.e., $\mathbf{x}$), using the sensitive features (i.e., s_i) as debiasing constraints.

Counterfactual Generator. For the sake of reproducibility and reliability, the counterfactuals are generated by a third-party counterfactual framework. We used DiCE [32, 33], an open-source framework developed by Microsoft. DiCE not only offers several strategies for generating counterfactual samples but also is a model-agnostic approach. Furthermore, it is built upon *Proximity*, *Sparsity*, *Diversity*, and *Feasibility* constraints. We used the Genetic and KDtree strategies. The number of counterfactuals generated for each negatively predicted sample is equal to 100 (i.e., $|\mathcal{C}_{\mathbf{x}}| = 100$). We tried to use another agnostic generator, MACE [26], but the only compatible

models are LR, DT, RF, and MLP (only for a 10-neuron single layer). Moreover, MACE never managed to generate the required number of counterfactuals making the analysis impractical.

Sensitive-Feature Classifier. This component plays a crucial role in our methodology since it allows the system to discover discriminatory models. For each sensitive feature, a classifier is thus learned. We exploited RF, MLP, and XGB for implementing this component.

Metrics. We evaluate the classifiers' performance with confusion matrix-based Accuracy (ACC) and F1 metrics, the Area Under the Receiver Operative Curve (AUC), and separation fairness metric Difference in Equal Opportunity[6] (DEO).

Split and Hyperparameter Tuning. The datasets have been divided with the 90/10 train-test split hold-out method. We restricted the test set to 10% due to computational cost of generating counterfactual samples for each data point. The *Decision Maker* and the *Sensitive-Feature Classifier* models have been tuned on the training set with a Grid Search 5-fold cross-validation methodology, the first optimizing AUC metric and the latter optimizing F1 score to prevent unbalanced predictions on the sensitive feature.

6 Experiments and discussion

The Section introduces the experiments and answers the RQs. The analysis of the results will be limited to XGB as $f_s(\cdot)$ due to space constraints.

Table 3: AUC, Accuracy, and F1 score on the Adult, Adult-debiased, Crime, and German test set of the Sensitive Feature Classifiers. We mark the best-performing method for each metric in bold font.

			$f_s(\cdot)$		
Dataset	s	metric	**RF**	**MLP**	**XGB**
Adult	*gender*	AUC	0.9402	0.9363	**0.9413**
		ACC	0.8539	**0.8559**	0.8463
		F1	0.8900	**0.8914**	0.8769
Adult-debiased	*gender*	AUC	**0.8028**	0.8010	0.7896
		ACC	**0.7482**	0.7480	0.7444
		F1	**0.8274**	0.8227	0.8106
Crime	*race*	AUC	0.9885	0.9893	**0.9910**
		ACC	**0.9500**	0.9450	0.9450
		F1	**0.9576**	0.9532	0.9532
German	*gender*	AUC	0.7106	0.5091	**0.7139**
		ACC	**0.7300**	0.6900	0.6900
		F1	**0.8344**	0.8166	0.7704

Is there a method for determining whether a dataset contains proxy features or not? (RQ1) This experiment aims to assess how well the sensitive-feature classifier can identify if a subject belongs to the *privileged* or *unprivileged* group, without exploiting sensitive features in the training phase. We trained a sensitive-feature classifier for each dataset. The investigated sensitive features are *gender* or *race* and results are shown in Table 3.

- The first observation is that every *Sensitive-Feature Classifier* shows to be accurate for all the datasets. Among them, XGB exhibits the best performance in terms of AUC while RF shows the most promising ACC and F1 performance. For the Adult, Adult-debiased, and Crime datasets, MLP performance is comparable to XGB and RF, while shows a low AUC on the German dataset. We recall we seek models with high F1 and AUC since it indicates that the classifiers provide accurate and balanced predictions.
- The careful reader may have noticed that, on the Adult-debiased, the sensitive-feature classifiers exhibit the worst performance (in comparison with the original Adult dataset). This behavior is due to the debiasing process, as the Adult-debiased dataset has been deprived of features highly correlated with the sensitive ones. Noteworthy, the prediction capability remains high despite the absence of sensitive and sensitive-correlated features.

Final comments. *Results show that, due to proxy features, it is possible to train a classifier able to predict sensitive characteristics. Moreover, it is still possible to predict sensitive information even when only low correlated features with the sensitive information are available (i.e., Adult-debiased).*

Does the Fairness Under Unawareness setting ensure that decision biases are avoided? (RQ2) The *Fairness Under Unawareness* setting aims to ensure fair treatment by removing sensitive features from training data. However, as demonstrated previously, it is possible to predict sensitive information due to proxy features. Before evaluating fairness metrics, we evaluated the accuracy performance of the classifiers exploited to implement the *Decision Maker*.

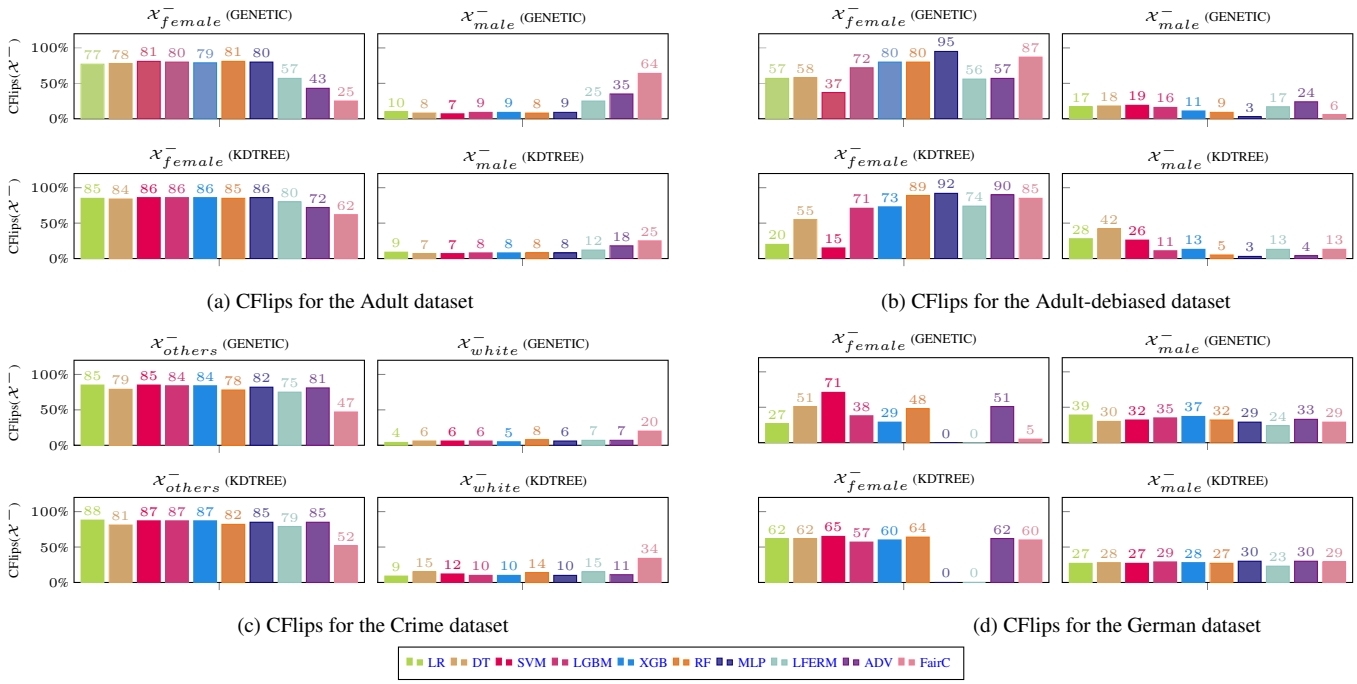
- Table 4 indicates that all classifiers work well in terms of accuracy metrics. However, as expected, adopting *Fairness Under Unawareness* – i.e., removing all the sensitive information and, for Adult-debiased, also removing highly correlated features – has caused a worsening of the performance for all the classifiers (see the comparison between Adult and Adult-debiased results). This observation suggests that sensitive and sensitive-correlated information may be "necessary" to predict the target label correctly. Table 4 reports the fairness evaluation computing the Difference in Equal Opportunity (DEO). It is worth noticing that removing the considered sensitive information (i.e., gender and race) has not improved model equity. This result shows that not removing proxy features makes the *Fairness Under Unawareness* setting useless since the model can implicitly learn them. A clear example is the Adult-debiased dataset, where DEO values are generally better than on the Adult dataset. However, some degree of discrimination is still present due to non-linear proxy features. Furthermore, despite adopting the *in-processing* debiasing constrained optimization, the debiased *Decision Maker* does not seem to improve fairness performance consistently.

Final comments. *The classifiers seem to be affected by discrimination even when the sensitive information is omitted. Accordingly, imposing Fairness Under Unawareness setting is not sufficient to avoid decision biases and discrimination.*

Is counterfactual reasoning effective to discover decision biases? (RQ3) This experiment aims to unveil potential decision biases by counterfactual reasoning. Figure 2 reports the CFlips values (see Definition 1) for each classifier and category – i.e., *privileged*, see Eq. 5, and *unprivileged*, see Eq. 6, with XGB as $f_s(\cdot)$. The metric tells us how frequently a change in the decision (from negative to positive) for a sample is followed by a change in the sensitive-feature classification (e.g., from *female* to *male* and vice-versa).

- The proposed metric seems to operate as expected since some hidden discriminatory behaviors emerge. For instance, the counterfactuals belonging to the *unprivileged* category, i.e., *female* or *others*, have a much higher CFlips than counterfactuals of *privileged* samples, i.e., *male* or *white*. This high percentage of flips for the unprivileged category means that the counterfactuals for

[6] $\text{DEO} = |\mathbb{P}(\hat{Y} = 1 \mid S = 1, Y = 1) - \mathbb{P}(\hat{Y} = 1 \mid S = 0, Y = 1)|$



(a) CFlips for the Adult dataset

(b) CFlips for the Adult-debiased dataset

(c) CFlips for the Crime dataset

(d) CFlips for the German dataset

Figure 2: CFlips (i.e., Counterfactual Flips, see Definition 1) for samples of the *unprivileged* and *privileged* sensitive group. The groups are *others* (s^-) and *white* (s^+) for the Crime dataset, and *female* (s^-) and *male* (s^+) for the Adult, Adult-deb, and German dataset.

Table 4: Accuracy, DEO, and ΔCFlips (%) metrics with XGB as $f_s(\cdot)$ on the Adult, Adult-debiased, Crime, and German test set. We mark the best-performing method for each metric in bold font.

	Adult (*gender*)				Adult-debiased (*gender*)				Crime (*race*)				German (*gender*)			
			ΔCFlips ↓				ΔCFlips ↓				ΔCFlips ↓				ΔCFlips ↓	
$f(\cdot)$	ACC↑	DEO↓	Genetic	KDtree	ACC↑	DEO↓	Genetic	KDtree	ACC↑	DEO↓	Genetic	KDtree	ACC↑	DEO↓	Genetic	KDtree
LR	0.8099	0.0546	67.05	76.03	0.7367	0.0695	40.02	**8.37**	**0.8700**	0.3294	81.70	78.76	0.7600	0.1400	11.89	35.65
DT	0.8161	0.0760	70.23	77.14	0.8017	0.0492	39.65	12.74	0.8200	0.4039	72.80	66.25	0.7600	0.0500	21.61	**22.43**
SVM	0.8541	0.0644	73.99	79.35	0.8061	0.0353	**18.27**	11.66	**0.8700**	0.3843	79.17	75.15	0.7600	0.0300	38.64	37.29
LGBM	0.8658	0.0379	70.87	78.40	0.8371	0.0470	56.16	59.11	0.8400	0.2824	78.22	76.52	0.7500	0.1900	**3.54**	28.54
XGB	**0.8698**	0.0635	70.40	78.42	**0.8375**	0.0400	69.02	59.22	0.8500	0.2941	78.89	76.79	**0.7900**	0.0400	7.85	32.37
RF	0.8534	0.0216	73.09	76.68	0.8267	0.0703	71.12	83.87	0.8400	0.2824	70.23	68.19	0.7600	0.0800	15.44	37.17
MLP	0.8494	0.0529	71.32	77.95	0.8156	**0.0173**	92.38	89.37	0.8650	0.3294	76.41	75.22	0.7600	0.0300	29.07	29.00
LFERM	0.8428	**0.0194**	32.40	68.25	0.7953	0.0179	39.67	60.55	0.8400	0.2941	68.68	63.69	0.7200	**0.0200**	24.24	23.23
ADV	0.8512	0.1399	**7.96**	53.90	0.8196	0.0326	33.17	85.81	0.8100	0.1882	74.09	74.12	0.7300	0.2200	18.46	31.51
FairC	0.8395	0.2451	38.72	**36.27**	0.8054	0.0529	80.12	72.24	0.7500	**0.1373**	**27.06**	**18.40**	0.7400	0.0500	23.94	30.81

the female (and "others" race) group must show male (and white) characteristics to get a positive decision. In this respect, Adult and Crime are characterized by the highest CFlips values and the largest difference between the *privileged* and *unprivileged* groups (see Table 4). We underline that CFlips and ΔCFlips results complement, they explain (e.g., highlighting if privileged group characteristics lead to a positive outcome) but also overturn the DEO metric in Table 4, shedding light on how the discriminatory classifiers work.

- The debiasing models perform particularly well (see the DEO metric in Table 4) on datasets where sensitive features can be easily identified, i.e., the datasets characterized by accurate sensitive-feature classifiers. Notwithstanding, the ΔCFlips in Table 4 highlights that a certain degree of discrimination persists for the datasets i) with features with a low correlation with the sensitive features or ii) composed by just a few samples. For instance, for the Crime dataset, even though all the sensitive-feature classifiers exhibit high accuracy, only FairC succeeds in decreasing discrimination at the expense of a significant loss in accuracy.
- The Adult-debiased dataset shows a smaller number of CFlips than Adult, especially for LR and SVM. However, even the most accurate XGB and LGBM show an evident discrepancy between the *privileged* and *unprivileged* groups. This might indicate that both classifiers learned correlations between the proxy features and the target. The German dataset has a similar trend with MLP and SVM as the most affected by unfairness. Finally, German's small test set size and the low accuracy of XGB as $f_s(\cdot)$ drive MLP, and LFERM to have 0 flips in the *female* category.
- Figure 3 analyzes the impact of the number of generated counterfactuals and the validity of the metric ΔCFlips when different sensitive features are present in the same dataset. The figure shows that the values of ΔCFlips are stable and reliable if the metric is

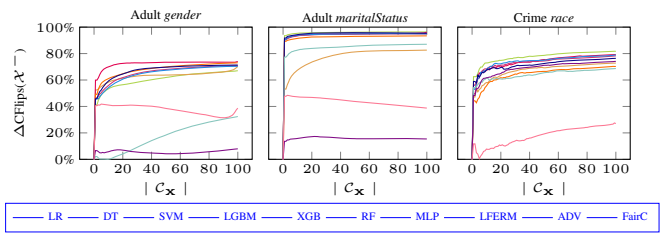


Figure 3: Ablation study at different number of generated CF (i.e. $|\mathcal{C}_\mathbf{x}|$) for each sample with Genetic strategy. For Adult, we also investigate a second sensitive feature (i.e., *marital-status*).

computed on at least 20 counterfactuals for each sample. A different behavior can be observed for LFERM in Adult-gender and FairC in Crime-race that start with an optimal performance, and then their ΔCFlips increases linearly. This trend needs further investigation, but it could be related to the higher number of counterfactuals. Indeed, with several counterfactuals, some could be farther from the original sample, and this distance probably entails a higher ΔCFlips. However, in fairness research, measuring the distance of a sample from the decision boundary of the classifier is a timely challenge [18]. It is worth mentioning that the analysis can be conducted for different features on the same dataset, even when it contains more than one sensitive feature (see *gender* and *maritalStatus* plots for Adult dataset).

Table 5: Leveraging different *sensitive-feature classifiers* (i.e., RF, MLP, and XGB) and testing CFlips metrics. As expected, for Adult and Crime datasets whose classifiers perform accurately, the differences in the predictions of CFlips are negligible.

			ΔCFlips					
			RF		**MLP**		**XGB**	
Dataset	s	$f(\cdot)$	Genetic	KDtree	Genetic	KDtree	Genetic	KDtree
Adult	*gender*	LR	67.35	76.09	66.75	75.28	67.05	76.03
		DT	70.10	77.16	70.07	77.14	70.23	77.14
		SVM	74.03	79.33	73.99	79.82	73.99	79.35
		LGBM	70.72	78.40	70.26	77.57	70.87	78.40
		XGB	70.34	78.59	69.54	77.42	70.40	78.42
		RF	73.24	77.01	72.25	75.68	73.09	76.68
		MLP	71.17	77.86	71.24	77.80	71.32	77.95
		LFERM	32.15	68.13	32.33	67.86	32.40	68.25
		ADV	**4.36**	53.70	**4.25**	53.80	**7.96**	53.90
		FairC	43.09	**35.86**	43.51	**35.78**	38.72	**36.27**
Crime	*race*	LR	80.56	78.33	82.57	86.31	81.03	78.76
		DT	72.48	66.04	74.13	71.76	72.59	66.25
		SVM	79.02	75.31	83.60	83.36	78.29	75.15
		LGB	78.23	76.65	82.07	85.11	77.82	76.52
		XGB	78.93	76.84	81.46	85.35	78.52	76.79
		RF	70.27	68.12	72.87	73.84	69.32	68.19
		MLP	76.90	74.66	82.32	82.31	76.04	75.22
		LFERM	68.42	63.85	71.16	71.44	68.32	63.69
		ADV	74.51	74.10	74.42	81.80	73.62	74.12
		FairC	**25.50**	**17.60**	**31.77**	**25.13**	**25.46**	**18.40**

- Table 5 extends the analysis to different $f_s(\cdot)$ (for the sake of space, the previous experiments adopted XGB as $f_s(\cdot)$). The results show that, also for Adult and Crime, the proposed metric is stable and reliable.
- The counterfactual generation strategies reveal some important findings: there exist similar real samples (i.e., similar people) for which the switch to the privileged group led to a positive outcome. Indeed, KDtree searches among dataset samples – i.e., $\mathbf{c_x} \in \mathcal{D}$ – thus CFlips ≥ 0 is critical. Genetic strategy, instead, analyzes the unexplored space and confirms the discriminatory behavior.

Final comments. *In the various plots emerges that the unprivileged samples, to achieve favorable decisions, must take on the characteristics of privileged samples. The results demonstrate that counterfactual reasoning effectively discovers decision biases and complements SOTA fairness metrics.*

Is it possible to define a strategy for identifying the proxy features? (RQ4) In RQ1, we highlighted how it is possible to determine if a dataset contains proxy features. Here, we define a strategy to identify them in the dataset. A counterfactual $\mathbf{c_x}$ can be seen as a perturbation from a starting sample $\mathbf{x}$ of a quantity ϵ (i.e., $\mathbf{c_x} = \mathbf{x} + \epsilon$). For a numerical or ordinal feature i, ϵ_i can be expressed as the difference between the counterfactual and the feature of the sample $c_{x_i} - x_i$. For a categorical feature j, ϵ_j can be expressed in a *one-hot encoding* form as -1 to the category that is removed and 1 to the category that is engaged. Let be δ the difference between the posterior conditional probability of predicting a counterfactual sample and the original sample as belonging to the privileged group (i.e., $\delta = \mathbb{P}(f_s(\mathbf{c_x}) = 1|\mathbf{c_x}) - \mathbb{P}(f_s(\mathbf{x}) = 1|\mathbf{x})$). We can identify the most influential features for $f_s(\cdot)$ evaluating the Pearson correlation between ϵ and δ: $\rho(\epsilon, \delta)$ (a demonstrative example in Table 6). In Figure 4 we can find the top-6 most correlated features with a Flip in $f_s(\cdot)$ with MLP as $f(\cdot)$ decision boundary for the generation of $\mathbf{c_x}$ and also as $f_s(\cdot)$ for the Adult-debiased dataset. A negatively correlated feature (e.g., *Adm-Clerical*) is a feature that has an opposite direction respect $\mathbb{P}(\hat{s} = 1 \mid \mathbf{x})$ while a positively correlated one (e.g., *hours per weeek*) has the same direction.

Table 6: Demonstrative example of ρ computation based on ϵ and δ for a numeric and categorical feature of Adult-debiased dataset.

	numeric	ordinal	Category (*workclass*)			*gender*
	capital gain	*education-n.*	*Private*	*Public*	*Unemployed*	$\mathbb{P}(\hat{s} = 1\|x)$
$\mathbf{c_{x_1}}$	5000	6	1	0	$0\|f_s(\mathbf{c_{x_1}})$	0.7
$\mathbf{x_1}$	2000	2	0	0	$1\|f_s(\mathbf{x_1})$	0.1
$\epsilon_{\mathbf{x_1}}$	3000	4	1	0	$-1\|\delta_{\mathbf{x_1}}$	0.6
$\mathbf{c_{x_2}}$	2800	5	0	1	$0\|f_s(\mathbf{c_{x_2}})$	0.3
$\mathbf{x_2}$	600	5	1	0	$0\|f_s(\mathbf{x_2})$	0.7
$\epsilon_{\mathbf{x_2}}$	2200	0	-1	1	$0\|\delta_{\mathbf{x_2}}$	-0.4
			...			
$\epsilon_{\mathbf{x_3}}$	1200	-1	0	1	$-1\|\delta_{\mathbf{x_3}}$	-0.6
$\rho(\epsilon, \delta)$	0.91	0.93	0.78	-0.99	-0.36	

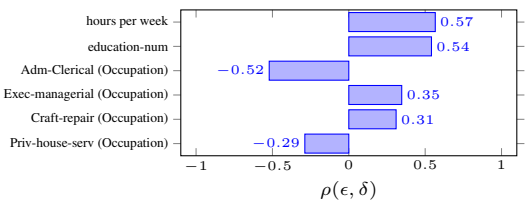


Figure 4: Top-6 most correlated features with a *gender* Flip (i.e., $\rho(\epsilon, \delta)$) on the Adult-debiased dataset with Genetic strategy as $g(\mathbf{x})$ and MLP as both $f(\cdot)$ and $f_s(\cdot)$.

7 Conclusion and Future Work

In the context of *fairness under unawareness*, where the sensitive information is omitted, *Decision Makers* can often be biased as they leverage proxy features of sensitive ones. In this work, we present a novel methodology to detect bias in *Decision-Making* models by analyzing the sensitive behavior of (counterfactual) samples belonging to the positive side of the decision boundary. In detail, we exploit three different counterfactual generation strategies to do so. Specifically, the newly generated counterfactuals could belong to another sensitive group, thus suggesting potential discrimination in the decision process. To comprehensively assess the proposed methodology, we conducted experiments with ten decision makers (including state-of-the-art debiasing models) and three sensitive-feature classifiers. To measure the extent of the discriminatory behavior, we introduce a new metric to find how many counterfactuals belong to another sensitive group. The contribution of this work is manifold: (i) we demonstrate that *fairness under unawareness* assumption is not sufficient to mitigate bias, (ii) we propose a methodology for the bias auditing task, (iii) we show that counterfactual reasoning is an effective methodology to unveil the bias, and (iv) we define a procedure to identify proxy features leveraging counterfactual reasoning. In the future, we plan to define a strategy to generate fair and actionable counterfactual samples with the aim of developing a model that could be effectively fair in the context of *fairness under unawareness*.

Acknowledgements

This work was partially supported by the following projects: LUTECH DIGITALE 4.0, OVS Fashion retail Reloaded, SECURE SAFE APULIA, CT_FINCONS III, VAI2C - Virtual Artificial Intelligence Contact Center, Progetto IDENTITA - rete Integrata meDiterranea per l'osservazione ed Elaborazione di percorsi di Nutrizione personalizzaTa contro la malnuTrizione.

References

[1] Sray Agarwal and Shashin Mishra, *Responsible AI*, Springer, 2021.
[2] Ashwathy Ashokan and Christian Haas, 'Fairness metrics and bias mitigation strategies for rating predictions', *Inf. Process. Manag.*, **58**(5), 102646, (2021).
[3] Mislav Balunovic, Anian Ruoss, and Martin T. Vechev, 'Fair normalizing flows', in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, (2022).
[4] P. J. Bickel, E. A. Hammel, and J. W. O'Connell, 'Sex bias in graduate admissions: Data from berkeley', *Science*, **187**(4175), 398–404, (1975).
[5] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson, 'Counterfactual reasoning and learning systems: The example of computational advertising.', *Journal of Machine Learning Research*, **14**(11), (2013).
[6] Consumer Financial Protection Bureau, 'Using publicly available information to proxy for unidentified race and ethnicity: a methodology and assessment', (2014).
[7] Jiahao Chen, 'Fair lending needs explainable models for responsible recommendation', in *FATREC'18 Proceedings of the Second Workshop on Responsible Recommendation*, Vancouver, British Columbia, Canada, (October 8 2018).
[8] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell, 'Fairness under unawareness: Assessing disparity when protected class is unobserved', in *FAT*, pp. 339–348. ACM, (2019).
[9] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq, 'Algorithmic decision making and the cost of fairness', in *KDD*, pp. 797–806. ACM, (2017).
[10] Giandomenico Cornacchia, Vito Walter Anelli, Giovanni Maria Biancofiore, Fedelucio Narducci, Claudio Pomo, Azzurra Ragone, and Eugenio Di Sciascio, 'Auditing fairness under unawareness through counterfactual reasoning', *Information Processing & Management*, **60**(2), 103224, (2023).
[11] Giandomenico Cornacchia, Vito Walter Anelli, Fedelucio Narducci, Azzurra Ragone, and Eugenio Di Sciascio, 'Counterfactual reasoning for decision model fairness assessment', in *Companion Proceedings of the ACM Web Conference 2023*, pp. 229–233, (2023).
[12] Giandomenico Cornacchia, Vito Walter Anelli, Fedelucio Narducci, Azzurra Ragone, and Eugenio Di Sciascio, 'A general architecture for a trustworthy creditworthiness-assessment platform in the financial domain', *Annals of Emerging Technologies in Computing (AETiC)*, **7**(2), (2023).
[13] Giandomenico Cornacchia, Fedelucio Narducci, and Azzurra Ragone, 'A general model for fair and explainable recommendation in the loan domain (short paper)', in *KaRS/ComplexRec@RecSys*, volume 2960 of *CEUR Workshop Proceedings*. CEUR-WS.org, (2021).
[14] Kimberlé Williams Crenshaw, 'Mapping the margins: Intersectionality, identity politics, and violence against women of color', in *The public nature of private violence*, 93–118, Routledge, (2013).
[15] Sanjiv Das, Michele Donini, Jason Gelman, Kevin Haas, Mila Hardt, Jared Katzman, Krishnaram Kenthapadi, Pedro Larroy, Pinar Yilmaz, and Muhammad Bilal Zafar, 'Fairness measures for machine learning in finance', *The Journal of Financial Data Science*, **3**(4), 33–64, (2021).
[16] Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil, 'Empirical risk minimization under fairness constraints', in *NeurIPS*, pp. 2796–2806, (2018).
[17] Julia Dressel and Hany Farid, 'The accuracy, fairness, and limits of predicting recidivism', *Science Advances*, **4**(1), eaao5580, (2018).
[18] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel, 'Fairness through awareness', in *ITCS*, pp. 214–226. ACM, (2012).
[19] Marc Elliott, Allen Fremont, Peter Morrison, Philip Pantoja, and Nicole Lurie, 'A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity', *Health services research*, **43**, 1722–36, (05 2008).
[20] Alessandro Fabris, Andrea Esuli, Alejandro Moreo, and Fabrizio Sebastiani, 'Measuring fairness under unawareness of sensitive attributes: A quantification-based approach', *J. Artif. Intell. Res.*, **76**, 1117–1180, (2023).
[21] Roberta Ferrario, 'Counterfactual reasoning', in *CONTEXT*, volume 2116 of *Lecture Notes in Computer Science*, pp. 170–183. Springer, (2001).
[22] Matthew L. Ginsberg, 'Counterfactuals', *Artif. Intell.*, **30**(1), 35–79, (1986).
[23] Moritz Hardt, Eric Price, and Nati Srebro, 'Equality of opportunity in supervised learning', in *NIPS*, pp. 3315–3323, (2016).
[24] Hans Hofmann, *UCI Statlog (German Credit Data) Data Set.*, UCI Machine Learning Repository, 2000.
[25] Max Hort and Federica Sarro, 'Privileged and unprivileged groups: An empirical study on the impact of the age attribute on fairness', in *Proceedings of the 2nd International Workshop on Equitable Data and Technology*, FairWare '22, p. 17–24, New York, NY, USA, (2022). Association for Computing Machinery.
[26] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera, 'Model-agnostic counterfactual explanations for consequential decisions', in *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pp. 895–905. PMLR, (2020).
[27] Ronny Kohavi and Barry Becker, *UCI Adult Data Set.*, UCI Machine Learning Repository, May 1996.
[28] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva, 'Counterfactual fairness', in *NIPS*, pp. 4066–4076, (2017).
[29] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan, 'A survey on bias and fairness in machine learning', *ACM Computing Surveys (CSUR)*, **54**(6), 1–35, (2021).
[30] Tim Miller, 'Explanation in artificial intelligence: Insights from the social sciences', *Artif. Intell.*, **267**, 1–38, (2019).
[31] Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova, 'Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds', in *FAccT*, pp. 386–400. ACM, (2021).
[32] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan, 'Explaining machine learning classifiers through diverse counterfactual explanations', in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 607–617, (2020).
[33] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan, 'Explaining machine learning classifiers through diverse counterfactual explanations', in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, p. 607–617, New York, NY, USA, (2020). Association for Computing Machinery.
[34] Judea Pearl, 'Causation, action and counterfactuals', in *ECAI*, pp. 826–828. John Wiley and Sons, Chichester, (1994).
[35] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini, 'Discrimination-aware data mining', in *KDD*, pp. 560–568. ACM, (2008).
[36] Michael Redmond, *UCI Communities and Crime Data Set*, UCI Machine Learning Repository, 1995.
[37] Boris Ruf and Marcin Detyniecki, 'Active fairness instead of unawareness', *CoRR*, **abs/2009.06251**, (2020).
[38] Yongkai Wu, Lu Zhang, and Xintao Wu, 'Counterfactual fairness: Unidentification, bound and algorithm', in *IJCAI*, pp. 1438–1444. ijcai.org, (2019).
[39] Samuel Yeom, Anupam Datta, and Matt Fredrikson, 'Hunting for discriminatory proxies in linear regression models', *Advances in Neural Information Processing Systems*, **31**, (2018).
[40] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi, 'Fairness Constraints: Mechanisms for Fair Classification', in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, eds., Aarti Singh and Jerry Zhu, volume 54 of *Proceedings of Machine Learning Research*, pp. 962–970. PMLR, (20–22 Apr 2017).
[41] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell, 'Mitigating unwanted biases with adversarial learning', in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, p. 335–340, New York, NY, USA, (2018). Association for Computing Machinery.