Machine Learning (2024) 113:3111–3142
https://doi.org/10.1007/s10994-023-06319-8

PreCoF: counterfactual explanations for fairness

Sofie Goethals[1] · **David Martens**[1] · **Toon Calders**[2]

Received: 28 February 2022 / Revised: 26 January 2023 / Accepted: 10 February 2023 /
Published online: 28 March 2023

Abstract
This paper studies how counterfactual explanations can be used to assess the fairness of a model. Using machine learning for high-stakes decisions is a threat to fairness as these models can amplify bias present in the dataset, and there is no consensus on a universal metric to detect this. The appropriate metric and method to tackle the bias in a dataset will be case-dependent, and it requires insight into the nature of the bias first. We aim to provide this insight by integrating explainable AI (XAI) research with the fairness domain. More specifically, apart from being able to use (Predictive) Counterfactual Explanations to detect *explicit bias* when the model is directly using the sensitive attribute, we show that it can also be used to detect *implicit bias* when the model does not use the sensitive attribute directly but does use other correlated attributes leading to a substantial disadvantage for a protected group. We call this metric *PreCoF*, or Predictive Counterfactual Fairness. Our experimental results show that our metric succeeds in detecting occurrences of *implicit bias* in the model by assessing which attributes are more present in the explanations of the protected group compared to the unprotected group. These results could help policymakers decide on whether this discrimination is *justified* or not.

Keywords Explainable Artificial Intelligence · Counterfactual explanations · Fairness · Data science ethics

Editors: Dana Drachsler Cohen, Javier Garcia, Mohammad Ghavamzadeh, Marek Petrik, Philip S. Thomas.

✉ Sofie Goethals
sofie.goethals@uantwerpen.be

David Martens
david.martens@uantwerpen.be

Toon Calders
toon.calders@uantwerpen.be

[1] Department of Engineering Management, University of Antwerp, 2000 Antwerp, Belgium

[2] Department of Computer Science, University of Antwerp, 2000 Antwerp, Belgium

1 Introduction

More and more, Artificial Intelligence (AI) is making decisions in high stakes domains of our life, such as employment, finance, justice, and healthcare. As the influence and scope of these decisions is increasing, there are growing concerns that the models making these decisions might unintentionally encode and even amplify human bias (Corbett-Davies & Goel, 2018). This is why it is of huge importance to understand the decisions models are making and to ensure they are fair. We focus on fairness in classification, where the goal is to prevent discrimination against people based on their membership of a sensitive group, without compromising the utility of the classifier (Caton & Haas, 2020; Dwork et al., 2012).

Different automatic methods to deal with discrimination, however, make different implicit assumptions about the nature of bias in the data and the right method to apply will be case-dependent and often policy-related (Wachter et al., 2021). Arguably, the data scientist is not the right person to make this call. The necessity for the involvement of policymakers and legal scholars enlarges the need for an automated, data-driven procedure that can detect and assess the source of automated discrimination in predictive models to support decision making (Wachter et al., 2021). As other authors already argue (Rudin et al., 2018), it is misguided to focus on fairness while not obtaining transparency first as it is not fair that life-changing decisions would be made without entitlement to an explanation.

In this paper we answer the call for more transparency in the fairness domain (Rudin et al., 2018; Wachter et al., 2021) by linking Explainable AI with fairness, using Counterfactual Explanations. Counterfactual explanations form the basis of an important class of explainable AI methods (Adadi & Berrada, 2018), and a counterfactual explanation of a data instance is defined as the smallest change to the instance so that it ends up with a different classification outcome. We name our metric *PreCoF*, which stands for *Predictive Counterfactual Fairness*. *PreCoF* finds counterfactual explanations for each sensitive group by assessing for each of the attributes whether changing it to one of the default values would result in a class change. It identifies the attributes that are proportionally more present in the explanations of the protected group compared to the unprotected group. This term is not to be confused with *Counterfactual fairness* as we will explain in Sect. 2.3.1. The goal of *PreCoF* will not be to provide yet another calculation on the output of a decision making system but to shed light on underlying patterns for the discrimination in the model, so that policymakers can decide how to handle this appropriately.

A first example of something our metric is able to detect can be seen in the Adult Income dataset: the attribute *marital status* is the attribute that is proportionally the most present in the explanations of women compared to men. This offers additional insights into the model so that policymakers can decide whether this is a pattern that can be kept in the model or if the model should be modified. The results of the other datasets are also in line with patterns that we know to be present based on literature or through further analysis of the datasets.

It is important to highlight that our metric will make statements about the model but not about the underlying data. We expect them to reflect underlying patterns in the data but it is possible that two different machine learning models trained on the same data will give very different results.

2 Background

Machine learning algorithms pose a threat to fairness as they can amplify bias present in the dataset, but at the same time, they can also be leveraged to diminish this same bias. Humans are inherently biased, and there is arguably no way to verify to what extent this influences their decision-making. The bias of machines however, can be checked and this is why metrics such as the one proposed in this paper, hold great promise. In this paper, we use the term *bias* to describe the situation in which sensitive groups are substantially disadvantaged by an algorithm or model. Bias can seep into a model when it is trained on biased data, following the famous *garbage in, garbage out* principle about how flawed input data will result in flawed output (Geiger et al., 2020). There are different ways in which a dataset can contain *bias*: An example of this is *label bias*, which occurs when the ground truth and the observed outcome differ: this pattern can be seen in the criminal justice system where blacks are more likely to get arrested for minor offenses which taints the dataset (Corbett-Davies & Goel, 2018). *Subgroup validity* happens when the predictive power of features varies across groups. Furthermore, it is also possible that the training data is not representative of the whole population, which can lead to underperformance of the model in certain minority groups (Corbett-Davies & Goel, 2018). An example of this is image classification where programs have more difficulty classifying the gender of dark-skinned individuals due to the relative shortage of dark-skinned faces in facial datasets (Buolamwini & Gebru, 2018). There exist other kinds of biases that can emerge in the model but we will not enumerate them further. Programmers are not writing biases in their code on purpose; these biases emerge when the algorithms are trained on data, mimicking the biases that were already present in the data (Johnson, 2021).

Legislation is attempting to use a 'colorblind' approach that ignores socially-sensitive features, which is misguided to begin with (Johnson, 2021). The idea here is that you remove the bias from the dataset by removing the discriminatory attributes from it. However, in any sufficiently rich dataset, proxy variables will likely exist that closely correlate with the sensitive attributes (Kim, 2017) so just removing them will not work. The most famous example of this practice is 'redlining', where zip codes are used as a proxy for race in lending decisions. Removing all the attributes that are correlated with the sensitive attribute is not a good solution either (Kamiran & žliobaitÿe, 2013); in some cases, all attributes will be correlated with the sensitive attribute, or some of the correlated attributes are too informative to remove (e.g., field of study is correlated with gender but too important to remove in hiring decisions).

We make the distinction between *explicit bias*, when the model involves direct use of the sensitive attribute, and *implicit bias*, when there is a neutral attribute that substantially disadvantages the protected group. These are also called *direct* and *indirect discrimination* respectively. *Indirect discrimination* is arguably the most likely type of discrimination to arise from automated decision making due to the reliance of these system on inference and proxies of target variables and protected attributes (Wachter et al., 2020).

Many scholars see value in judging discrimination with common sense (Doyle, 2007), however, this is often ineffective in cases of *indirect discrimination*, especially when the relation between the protected attribute and the neutral attribute is not straightforward (Wachter et al., 2021). Intuition might fail us because it cannot be assumed that automated systems will discriminate in ways similar to humans or follow their patterns of discrimination: new and counterintuitive proxies for traditionally protected attributes can emerge but will not necessarily be detected (Wachter et al., 2021). If such an attribute is

found that substantially disadvantages the protected group, this is not necessarily a problem: some attributes can be justified, depending on the context of the case and the relevant legislation. *Justified indirect discrimination* occurs when the 'proportionality test' is passed, meaning that the attribute is both legally necessary and proportionate (Wachter et al., 2020). *PreCoF* is developed to fit in this mindset: can we find the attributes that explain why sensitive groups are more often predicted with a negative outcome? This can then lead to a discussion about these attributes being justified or not.

There are three main responses when such a bias is detected: First, one can do nothing and allow the bias to be amplified; second, fix the technical bias but maintain the society status quo and make sure that the machine learning does not make the society more biased which is called a *bias preserving* approach (Wachter et al., 2020). A third option is what are called *bias transforming* metrics and these aim to actively account for historical inequalities (Wachter et al., 2020). The adequate response will depend on the situation at hand, but doing nothing will in our opinion never be the right call.

2.1 Fairness metrics

There is no universal definition of fairness, which greatly complicates our research question. Some define fairness as *fairness through unawareness* (Pedreshi et al., 2008), which establishes fairness through removing the sensitive attributes from the dataset. However, this is not always possible as sometimes sensitive attributes are needed to make predictions. Even when the sensitive attribute is not directly relevant to the prediction task, correlated variables (e.g., race from zip code in the United States) make such a "blind" approach less efficient to counter discrimination (Fryer et al., 2008). Other often-used fairness metrics include *individual fairness* (Dwork et al., 2012), which states that similar individuals should be treated similarly, *demographic parity* (Calders et al., 2009) (which is also called *disparate impact* (Feldman et al., 2015) or *statistical parity* (Dwork et al., 2012)) which minimizes the absolute difference in outcome distributions of all groups, *equalized opportunities* (Hardt et al., 2016), which optimizes towards equal positive rate conditional on the target outcome and *equalized odds* (Hardt et al., 2016), which optimizes towards equal positive and negative rate conditional on the target outcome.

Demographic parity, *equalized odds* and *equal opportunity* are all group-based criteria, which are more suited to statistical analysis (Ritov et al., 2017) but can be very unfair from the point of the individual (Dwork et al., 2012): it provides protection for groups but not for specific individuals in those groups and we tend to care more about protection for individuals (Fleisher, 2021). It also does not provide protection against phenomena like cherry-picking.[1] Even more problematic, many of the group fairness metrics are mutually incompatible, which means it is impossible to satisfy all of them at the same time (Kleinberg et al., 2016; Verma & Rubin, 2018). This has as a consequence that the detection of discrimination can be 'gamed' through choosing the right fairness metric (Wachter et al., 2021). It has been shown that all these metrics suffer from deep statistical limitations and

[1] Cherry picking refers to members of sensitive groups being randomly chosen, or chosen for malicious reasons as a way to undermine members of those groups (Dwork et al., 2012; Fleisher, 2021) An example of this in college applications could be when the majority group is carefully screened, and the same number of applicants is randomly selected from the minority group. This is not fair for hard-working members of the minority group that will not get admitted, but would be compatible with a variety of group fairness criteria (Fleisher, 2021).

that they can even negatively impact the well-being of the groups they are trying to protect (Corbett-Davies & Goel, 2018). *Individual Fairness* is more strict than any group-notion fairness as it imposes a restriction on the decision for each pair of individuals. It also forbids a variety of discriminatory practices like explicit discrimination, implicit discrimination, redlining and tokenism (Fleisher, 2021). It can also detect cases of discrimination that various group fairness criteria miss like cherry-picking. However it is hard to define a metric function to measure the similarity of two inputs (Fleisher, 2021; Kim et al., 2018). A last metric is *Counterfactual Fairness* (Kusner et al., 2017), which is more related to our metric and will be discussed in Sect. 2.3.1.

All the metrics that are conditional on the target outcome such as *equalized odds* and *equal opportunity* are *bias preserving*, which means that they will preserve historical biases and just ensure that the machine learning model will not amplify these biases or insert new bias into the system (Wachter et al., 2020). They share the idea that the bias present in the target labels is meant to be there (Wachter et al., 2020). *Demographic Parity*, *Individual Fairness* and *Counterfactual Fairness* are *bias transforming metrics*. *PreCoF* is aimed to be a *bias transforming metric*, but it offers the transparency and flexibility for policy makers to decide this for each situation at hand. Choosing an appropriate metric can have political, legal and ethical implications and should be subject to more consideration and justification than is currently the case (Wachter et al., 2020). The previously discussed fairness metrics are not well suited to answer normative and legal questions on how the discrimination in the model should be handled and might ultimately prove to be irrelevant in court (Wachter et al., 2021).

2.2 Conditional fairness metrics

In practice, there often exists a certain set of attributes on which we deem it fair to discriminate (Xu et al., 2020). An example of such an attribute is the department choice in the Berkeley's graduate admission problem, where there allegedly was a bias against female applicants as they had a lower admission rate then male applicants (Xu et al., 2020; Pearl, 2009). After conditioning on department choice, this was no longer the case (Pearl, 2009). Conditional fairness is a more sound fairness metric where the outcome variables should be independent of sensitive variables conditional on these fair attributes (Xu et al., 2020). There exist various methods to implement conditional fairness such as explainable discrimination (Kamiran et al., 2013; Wachter et al., 2021) or conditional demographic disparity (Wachter et al., 2021). They have the point of view that some differences in decisions across sensitive groups can be explainable and hence tolerable (Kamiran et al., 2013). For example, in job applications the education level of a candidate can be such an explainable attribute (Kamiran et al., 2013).

The underlying fairness metrics in explainable discrimination and conditional demographic disparity are a bit different but they are based on the same principle (Kamiran et al., 2013; Wachter et al., 2021): Kamiran et al. (2013) measure the discrimination as the difference in positive rates between two sensitive groups: the discrimination that remains after subtracting the discrimination that can be explained by using the conditional attribute (*explainable discrimination*) is the *illegal discrimination*. Wachter et al. (2021) define *demographic disparity* as the difference in proportion of people from the protected group with a favorable and an unfavorable outcome. *Conditional demographic disparity (CDD)* follows the same principle but adds a conditional attribute: a decision-making system has no conditional discrimination if, after conditioning on this attribute, the decisions

are statistically independent of the sensitive attribute (Wachter et al., 2021). However, in both methods, it is not clear how the attributes on which conditional fairness is calculated are chosen: searching over all combinations of attributes would be prone to finding false positives (Wachter et al., 2021). Developers can be inclined to choose favorable conditions (Wachter et al., 2020) and it should not be up to them to choose these variables, but this should be fixed externally by law or domain experts (Kamiran et al., 2013). The selection of these conditional attributes becomes confusing and debatable as people might not agree about which combinations are reasonable (Kamiran et al., 2013). Furthermore, the conditional attributes are not necessarily the attributes that are used by the model. In large datasets, conditional variables might exist such that the data can be stratified in groups in such a way that there is no conditional demographic disparity, while that conditional variable is not even a factor used in the model. We will show this in Sect. 4.3.1.

We agree with the point of view that part of the discrimination can be explainable by other attributes, but our goal is to shed light on which attributes are making up the discrimination in the model so that policymakers can decide whether these are justified or not. Is it fair to use GPA in law admissions schools even though it is often biased against ethnic minorities? Is it desired to trade accuracy for fairness in crime recidivism prediction as this can result in a higher crime rate overall? Which biases are socially acceptable and can be maintained? Which actions are appropriate for a specific case? These are all questions that should be answered case by case in an open and transparent debate.

2.3 Related metrics

2.3.1 Counterfactual fairness

In recent years, fairness-aware machine learning has been studied from the causal perspective using causal modelling (Pearl, 2000). In line with this research, Kusner et al. (2017) define Counterfactual Fairness as a notion of fairness derived from Pearl's causal model (Pearl, 2000) where for an individual the prediction of the model is considered as fair if it is the same in the real world as it would be if the individual would belong to a different demographic group (Kusner et al., 2017; Wu et al., 2019). To measure this, they make explicit assumptions about the causal relationships in the data. One way for a predictor to be counterfactually fair is if it is a function of only non-descendants of the sensitive attribute, so this will be different depending on the chosen causal model. The biggest drawbacks of this methodology are that you need to make some untestable assumptions for such a causal model and that it is not scalable (Xu et al., 2020). It assumes that the causal relations between variables in a dataset are known, while in reality this is not the case. Furthermore, the legal frameworks that govern discrimination in multiple countries do not require a causal relationship with the protected attribute, so Counterfactual Fairness may fail to identify occurrences of legally actionable discrimination (Black et al., 2020). Several other authors also propose a causal approach to detect various forms of discrimination in a dataset (Bonchi et al., 2017; Schölkopf, 2017) but they suffer from the same drawbacks.

2.3.2 Counterfactual fairness *(bis)*

Sokol et al. (2019) already showed how counterfactual explanations can be used to check individual fairness. They consider an instance to be treated unfairly if that instance received the undesirable label and there exists a counterfactual explanation for that instance

that includes at least one protected attribute change (Sokol et al., 2019). We follow this approach when we use counterfactual explanations to identify explicit bias for an individual. On top of that, we also show that aggregating these counterfactual explanations can give more insights about the patterns of explicit bias in the algorithm.

2.3.3 CERTIFAI

CERTIFAI (Sharma et al., 2019) is a tool that can be applied to any black-box model to assess its fairness. It uses a custom genetic algorithm to generate counterfactuals and examines the explanations to assess the model's fairness, both on an individual and on a group level. The fitness of an individual is defined as the inverse distance between the input instance and its counterfactual. For an individual, if we allow the sensitive attributes to change, and the fitness goes up (distance to the counterfactual becomes smaller: desired outcome is more easily achieved), then the individual could claim the model is treating them unfairly. This tool can also be used to audit fairness on a group level: if the average fitness values of generated counterfactuals are lower for women than for men, this could be used as evidence that the model is not treating women fairly (Sharma et al., 2019). This tool is different from how we use counterfactual explanations as we will focus on the specific attributes and attribute values that occur in the explanations of both groups and not on the distance to the counterfactual instance.

2.3.4 Fairness in algorithmic recourse

The literature on algorithmic recourse has focused on finding "an actionable set of changes a person can undertake in order to improve their outcome" (Joshi et al., 2019; Karimi et al., 2021). Algorithmic recourse poses its own fairness criteria, where the effort to reach the required outcome is taken into account. If individuals from the protected group have to work harder than similar individuals from another group to achieve the desired outcome, then the concept of equal opportunity is violated (von Kügelgen et al., 2022). This notion of unfairness is not captured by predictive notions and is in line with CERTIFAI, as they both focus on the difference in effort different individuals have to make. To be able to find an 'actionable' set of changes, most authors assume, at least partial, causal knowledge. However, as in Sect. 2.3.1, the reliance on causal information creates practical issues that may limit its applicability (Black et al., 2020). As we are not necessarily interested in *actionable* counterfactuals, our method will not rely on causal assumptions about the data-generating process. We explain this further in Sect. 4.5.

2.3.5 FlipTest

FlipTest is a fairness testing approach, that also does not rely on causal information, but instead uses an optimal transport mapping to detect whether a model's behavior is sensitive to changes in the protected status (Black et al., 2020). Simply changing the protected attribute is not sufficient due to correlations in the data. Therefore, a transport map transports one probability distribution into another, for example women into men, in order to have a pair of inputs with which to query the model. An optimal transport map is used to minimize the sum of distances between a woman and the man she is mapped to (her *counterpart*), where the distance quantifies the difference between them. FlipTest analyzes the cases where the classifiers' output is different between the woman and her *counterpart*,

as these are individuals that might be harmed because of their group membership. Like FlipTest, *PreCoF* also aims to shed light on *why* the model is treating a certain subgroup differently but it uses a different method: it does not require to approximate an optimal transport mapping and does not depend on the distance function that is used to construct the mapping.

3 Counterfactual explanations as the solution

PreCoF aims to explain the discrimination in a predictive classification model, and create transparency regarding which attributes are the most discriminatory between different sensitive groups. This insight can then be used for subsequent discussions and decisions by law or domain experts on which attributes are justified and which attributes will just behave as proxies for the sensitive attribute. An example of Wachter et al. (2021) shows how some attributes can be valid in one case but not in another: when reviewing résumés for a firemen position, height can be deemed a valid discriminator but it seems highly unlikely that this will be the case when reviewing résumés for a CEO position (there it will just serve as a proxy for gender).

We agree with Wachter et al. (2021) that fairness is contextual: it is not possible to create a system that automatically detects and corrects discriminatory models as each case should be handled differently. What is needed is an 'early warning system' that provides transparency in automated discrimination (Wachter et al., 2021) which is what we aim to supply.

As Rudin et al. (2018) also state: it is arguably unfair to have life-changing decisions being made by a system without having any insights into the decisions, which brings us to the field of Explainable AI (XAI). XAI research aims at explaining how an AI system reached its decision (Gohel et al., 2021). XAI can enhance transparency as well as fairness as it provides explanations that can be understood and as such show bias that is present (Gohel et al., 2021; Sokol & Flach, 2021). There exist different sorts of explanation procedures for understanding predictive models, both on the global level as on the instance-level. Global explanations provide understanding of the complete model over the entire space of training instances and include methods like rule extraction (Craven & Shavlik, 1995; Martens et al., 2007) and global feature importance rankings (Breiman, 2001). Instance-based explanations aim to explain the model for an individual instance. Several types of instance-based explanations exist but the most popular model-agnostic ones (which means that they are applicable to any predictive model) are Counterfactual Explanations (Martens & Provost, 2014; Wachter et al., 2017), and feature importance methods on the instance-level like LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017). We want to assess fairness on the individual level so we will look at instance-based explanation methods, and we will focus on Counterfactual Explanations as they are better suited for our task than LIME or SHAP: the latter explain a prediction score rather than a decision so if we talk about unfair decisions, Counterfactual Explanations are better suited as they focus on the *treatment* an individual received (Fernandez et al., 2020). We focus on fair decision making, but in the case we want to assess fair scoring, SHAP values can be used in the same set-up. We present the results when using SHAP values instead of counterfactual explanations in Sect. 2. Our main argument that more insight is needed in the nature of the bias before deciding on a method to handle it, remains valid for both XAI techniques.

Assume we have a dataset D that consists of n instances and m attributes, where the attribute value of attribute j of an instance i is denoted by x_{ij} with $i \in \{1, 2, ..., n\}$, $j \in \{1, 2, ..., m\}$. The model M will make a decision, which is either a favorable (+, e.g. hired, credit granted) or a unfavorable (-, e.g. not hired, credit rejected) outcome.

$$M(\mathbf{x}_i) \in \{+, -\}$$

A counterfactual $\mathbf{c}$ of a factual instance $\mathbf{x}_i$ is an instance for which:

$$M(\mathbf{x}_i) \neq M(\mathbf{c})$$

and

$$d(\mathbf{x}_i, \mathbf{c}) \text{ is minimal}$$

So the counterfactual is another instance, while the counterfactual *explanation* is the difference between the two: $\mid \mathbf{c} - \mathbf{x}_i \mid$. As mentioned in Sect. 2.3, other metrics also use counterfactual explanations to assess fairness. However, our metric will be different as it does not need to assume a causal graph (Kusner et al., 2017), and does not use the distance to the counterfactual like Sharma et al. (2019), but will look at the actual explanations of decisions instead. Furthermore, we will use counterfactual explanations not only to show *explicit bias*, as done by Sokol et al. (2019), but also to get insights into the *implicit bias*, which is arguably the more challenging problem.

An advantage of also looking at *implicit bias* over *explicit bias* is that it deals with rules or patterns of behaviour, and as such can reveal underlying social inequalities and uncover structural unfairness in an algorithm (Wachter et al., 2021). Direct discrimination is simpler to detect: the action that is alleged to be discriminatory must explicitly refer to a protected characteristic while for indirect discrimination it is more difficult: a neutral attribute or criterion must be shown to substantially disadvantage the protected group, despite not explicitly addressing it (Wachter et al., 2021; Zliobaite, 2015).

4 Methodology

4.1 Materials

In this study, we focus on tabular datasets, mostly used in fairness-aware machine learning research (Le Quy et al., 2022). We use datasets from the financial (Adult Income dataset), criminological (Catalonia Juvenile dataset, Crimes and Communities dataset) and the educational (Student performance dataset, Law admission dataset) domain. All the datasets in this study are publicly available. More information about each dataset can be found in the Appendix.

4.2 Explicit bias

As already highlighted by Sokol et al. (2019), counterfactual explanations can be used to highlight explicit bias in a decision-making model, by searching for explanations that contain the sensitive attribute. We detect *explicit bias* by searching for counterfactual explanations that consist only of the sensitive attribute.

Assume we have a dataset D with sensitive attribute S, where the sensitive value is s, and the non sensitive value is ns. The group with sensitive value s is also called the *protected group* and the group with sensitive value ns is also called the *unprotected group*. The dataset consists of n instances $\mathbf{x}_i$, with m attributes, where the attribute value of attribute j for instance i is denoted by x_{ij} with $i \in \{1, 2, ..., n\}$, $j \in \{1, 2, ..., m\}$. The index of the sensitive attribute is z. The model M will make a decision, where we denote $+$ as the favorable outcome and $-$ as the unfavorable outcome.

A decision for the factual instance $\mathbf{x}_i$ that has a negative predicted outcome: $M(\mathbf{x}_i) = -$, is deemed to be unfair (*explicit bias*) if there exists a counterfactual instance $\mathbf{c}$, for the instance $\mathbf{x}_i$ that satisfies:

$$x_{iz} \neq c_z (\textit{the counterfactual instance has a different value for the sensitive attribute})$$
$$\forall j \in \{1, 2, ..., m\} \setminus z :$$
$$x_{ij} = c_j (\textit{except for the sensitive attribute, the factual and counterfactual instance are identical})$$

This means that the instance $\mathbf{c}$ that only differs from $\mathbf{x}$ with respect to the sensitive attribute receives a different classification from our prediction model M. An example of such an unfair explanation could be: "*If you would not have been a woman, you would have received the loan.*"

This analysis on the individual level could also be aggregated and as such, show patterns in the model. We aggregate the explanations by calculating how many people of each group receive such an explanation. How many negatively predicted persons from each sensitive group would have received a positive outcome, simply by changing their sensitive attribute? Which categories of the sensitive group experience *explicit bias* the most?

Machine learning models can also suffer from *fairness gerrymandering*; when there are different sensitive groups, the classifier can be fair for each individual group but can discriminate against structured subgroups (Kearns et al., 2017). Imagine we have two sensitive attributes: *race* and *gender*. When analyzing the explicit bias in the model, it is possible that no explanations are found with gender or race, but only with a combination of the two attributes (e.g., *"If you would not have been a black woman, you would have received the loan."*). Our method can take this into account by searching for all explanations that contain a combination of the sensitive attributes.

4.3 Implicit bias

We will use the same terminology as in Sect. 4.2, but now we will remove the sensitive attribute from the dataset before training the model; We will name this new dataset D'. This dataset will consist of n instances $\mathbf{x'}_i$ with $m - 1$ attributes, where the attribute value of attribute j for instance i is denoted by x'_{ij} with $i \in \{1, 2, ..., n\}$, $j \in \{1, 2, ..., m\} \setminus z$. We also have access to the original dataset D, where the sensitive attribute for each instance is still available under index z.

What our metric aims to measure, is how much more often a certain attribute is responsible (part of the counterfactual explanation) for a negative outcome decision for members of the protected group, compared to members of the unprotected group. So if changing height from short to tall is 100 times part of the counterfactual explanation for a non-hire

decision for 100 women ('if your height would have been tall instead of short, you would have been hired') (100%), and only 10 times of the counterfactual explanation for a non-hire decision for 100 men (10%), *PreCoF* will output 90% for the attribute height. We then show the features (and feature values) with the highest *PreCoF*.[2]

More formally, we test for every instance $\mathbf{x}_i$ with an unfavorable outcome for every attribute j whether changing them to one of the default values results in a counterfactual explanation E. We use a set of default values as we will not test every possible attribute value: for numerical attributes or very sparse categorical attributes, this will not be feasible. We select a set of default values, which for numerical attributes are the values of each decile. For categorical attributes, we take the most frequent (max 10) values that are at least present in 1 percent of the training set. If no attribute value is present in more than 1 percent of the training set, we will just take the 10 most occurring values.

Afterwards, we look at all negatively affected members of the protected group, and see how relatively often we can find a counterfactual explanation that consists only of attribute j. This relative number, we call $CoF(j, s)$. Similarly, we measure how often this attribute is part of the explanation for the unprotected members with negative outcome: $CoF(j, ns)$. Our final metric $PreCoF(j)$ simply calculates the difference between these two.

The mathematical definition for *PreCoF* is thus as follows (where the counterfactual explanation that leads to counterfactual instance $\mathbf{c}$ can only consist of a single attribute j [3]):

$$CoF(j,s) = \frac{\left|\left\{i | \exists \mathbf{c} : x_{iz} = s, M\left(\mathbf{x}'_i\right) = -, \mathbf{c} \text{ is a counterfactual of } \mathbf{x}'_i\right\}\right|}{\left|\left\{i | x_{iz} = s, M\left(\mathbf{x}'_i\right) = -\right\}\right|}$$

$$CoF(j,s) = \frac{\left|\left\{i | \exists \mathbf{c} : x_{iz} = ns, M\left(\mathbf{x}'_i\right) = -, \mathbf{c} \text{ is a counterfactual of } \mathbf{x}'_i\right\}\right|}{\left|\left\{i | x_{iz} = ns, M\left(\mathbf{x}'_i\right) = -\right\}\right|}$$

$$PreCoF(j) = CoF(j,s) - CoF(j,ns)$$

$$PreCoF_1 = \text{Attribute } j \text{ such that } j = \underset{\forall j \in \{1,2,\ldots,m\} \setminus z}{\operatorname{argmax}} PreCoF(j)$$

Our metric also allows us to look at the exact feature values of the factual and counterfactual instances. A difference here is that we only compare the categorical values as the numerical values are often too sparse to give us insights about the patterns in values. We define $PreCoF_f$ and $PreCoF_c$:

These are calculated in the same way as *CoF*, but CoF_f will output how often each attribute value is present as part of the factual instance and CoF_c will output how often each attribute value is present as part of the counterfactual instance. $PreCoF_f$ and $PreCoF_c$ again calculate the difference for CoF_f and CoF_c between the protected and the unprotected

[2] Like explained in Sect. 4.2, the protected group can also be a combination of multiple sensitive attributes. *PreCoF* can take this into account by comparing the explanations of this subgroup (e.g., black women) with the rest of the population.

[3] More formally, a counterfactual explanation e that only consists of attribute j means that the counterfactual explanation c satisfies:

$$M(\mathbf{c}) = +$$
$$x'_{ij} \neq c_j$$
$$\forall h \in \{1, 2, ..., m\} \setminus [j, z] : x'_{ih} = c_h$$

Table 1 A toy example

Row	Gender	School	Hobby	IQ	True Grade	Predicted Grade
1	*M*	*School1*	*Basket*	*High*	*Pass*	*Pass*
2	*M*	*School1*	*Football*	*High*	*Pass*	*Pass*
3	*M*	*School1*	*Football*	*Low*	*Fail*	*Fail*
4	*M*	*School2*	*Football*	*High*	*Fail*	*Fail*
5	*M*	*School2*	*Basket*	*Low*	*Fail*	*Fail*
6	*F*	*School2*	*Dance*	*High*	*Fail*	*Fail*
7	*F*	*School2*	*Dance*	*High*	*Fail*	*Fail*
8	*F*	*School2*	*Music*	*High*	*Pass*	*Fail*
9	*F*	*School2*	*Dance*	*High*	*Pass*	*Fail*
10	*F*	*School1*	*Music*	*High*	*Pass*	*Pass*

group, and $PreCoF_{f1}$ and $PreCoF_{c1}$ will be the attribute values for which respectively $PreCoF_f$ and $PreCoF_c$ are maximal out of all possible attribute values.

By also looking at the specific feature values in the factual and counterfactual instances, we can get more insights into the social patterns in the model. Examples of this can be seen in the results in Sects. 5.1, 5.2, and 5.4. Our metric is thus able to give us insights into the implicit bias of a prediction model, without the prediction model even having access to the sensitive attribute.

4.3.1 Toy example

We will illustrate the use of this metric with a simple toy example.

A machine learning model is trained on this toy dataset in Table 1 after removing the sensitive attribute (gender). Assume the following simple rule-based model:

If School = School2 or IQ = low, predict Fail; else predict Pass

The predicted outcome by this model can be seen in the last column of the table. This model scores an accuracy of 80 % but predicts more girls to fail than boys, even though in the dataset there are less girls that fail than boys.

We calculate the demographic disparity of our simple rule-based classifier:

$$\text{Demographic disparity} = P(\hat{y} = + \mid M) - P(\hat{y} = + \mid F) = 2/5 - 1/5 = 1/5$$

This metric just tells us that there is a difference in predicted outcome between boys and girls, but tells us nothing about *why* discrimination occurs and gives policymakers no clues on how to handle this. If the reason for this difference in predicted outcome is that the rejected girls have on average a lower IQ, and this is used by the model to predict that they will fail more often, then this could be a *justified* reason for a difference in positive rate, while for other attributes this will not be the case. This shows that group fairness metrics in general are not well suited to answer legal or normative questions as they will not provide any reasoning behind the metric.

In this small example, inspired by the Student Performance dataset, it is straightforward to see which attribute is inducing this bias. The model has learned that *School2* is associated with bad grades which disproportionally affects the female students. We will use this toy example to show that the *PreCoF* metric is able to detect this variable and as such give insights into why the discrimination occurred.

When using the *PreCoF* metric, we get the following results:

$$\begin{aligned} CoF(School, F) &= 4/4, \; CoF(IQ, F) = 0, \quad CoF(Hobby, F) = 0, \\ CoF(School, M) &= 1/3, \; CoF(IQ, M) = 1/3, \; CoF(Hobby, M) = 0 \end{aligned}$$

We calculate the attribute for which the difference between the protected (F) and the unprotected group (M) is the largest:

$$PreCoF_1 = School \qquad (CoF(School, F) - CoF(School, M) = 2/3, \text{ which is larger than 1/3 and 0})$$

We then use the *PreCoF* metric to also detect the feature values causing the differences:

$$\begin{aligned} PreCoF_{1f} &= School2 \\ PreCoF_{1c} &= School1 \end{aligned}$$

$PreCoF_1$ will be *School* as this is the attribute that is proportionally the most present in the explanations of the protected group (girls), compared to the unprotected group (boys). As will be discussed in Sect. 5.4, this will also be the case in the real dataset and could have implications in various areas such as college admissions, where girls could be incorrectly rejected because of the school they went to.

This toy example also shows that this metric will not necessarily point to the variables that are the most correlated with the sensitive attribute. Hobby is the most correlated with gender here, but it will not come out of the *PreCoF* metric as the model is not using this variable.

This toy example also allows us to highlight the difference of our metric with conditional fairness metrics; we show the difference by using the formulas of discrimination of Kamiran et al. (2013). For an explainable attribute E, which could in theory be any attribute from the dataset, Kamiran et al. (2013) consider dividing the database according to the possible values $e_1, \ldots, e_k$ of E. For each of the values e_i they compute a theoretical probability $P^*\big(\hat{y} = + \mid e_i\big)$ of being in the positive class by taking the mean $\frac{P(\hat{y}=+ \mid e_i,s)+P(\hat{y}=+ \mid e_i,ns)}{2}$, assuming that if this probability of being in the positive class differs between the protected and unprotected group, the truth must be in the middle. Based on this per-group estimate, they compute what would be the unbiased positive class probability for the protected group as follows: $\sum_{i=1}^{k} P(e_i \mid s) P^*(\hat{y} = + \mid e_i)$. The formula for the unprotected group is the same. Hence, the explainable difference between the two communities then becomes:

$$\begin{aligned} D_{explainable}(E) &= \sum_{i=1}^{k} P(e_i \mid s) P^*(\hat{y} = + \mid e_i) - \sum_{i=1}^{k} P\big(e_i \mid ns\big) P^*\big(\hat{y} = + \mid e_i\big) \\ &= \sum_{i=1}^{k} \big(P(e_i \mid s) - P(e_i \mid ns)\big) P^*\big(\hat{y} = + \mid e_i\big) \end{aligned}$$

The illegal discrimination then becomes the part of the discrimination that cannot be explained by the attribute E:

$$D_{illegal}(E) = D_{all} - D_{explainable}(E),$$

where D_{all} is equal to the demographic disparity:

$$D_{all} = P(\hat{y} = + \mid ns) - P(\hat{y} = + \mid s),$$

which is 1/5 for our toy dataset as calculated above.

With these formulas we get:

$$\begin{aligned} D_{explainable}(Hobby) =& (2/5 - 0) \times 1/2 + (3/5 - 0) \times 1/3 + (0 - 3/5) \times 0 \\ &+ (0 - 2/5) \times 1/2 \\ =& 1/5, \text{ giving } D_{illegal}(Hobby) = 0. \end{aligned}$$

Similarly we can compute $D_{illegal}(School) = -2/15$, and $D_{illegal}(IQ) = 28/75$.

This example shows that according to the explainable discrimination measure of Kamiran et al. (2013), variable (*Hobby*) could justify the discrimination, while the model is not even using this attribute. This shows the key difference with conditional fairness and our metric: we look at the factors that could change the decision of the model and where these factors differ the most between sensitive groups, while conditional fairness will search for a way to create stratified groups that satisfy a fairness metric.

4.4 Machine learning model

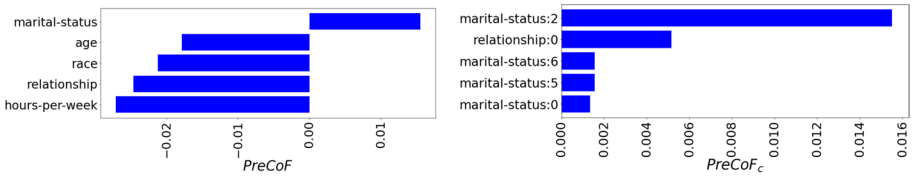
The machine learning model used for our experiments is a Random Forest model, tuned through five-fold cross-validation. The parameter grid that is used is [10, 50, 100, 500, 1000, 5000] for the number of trees and [10, 100, 500, n] for the maximum number of leaf nodes.

To measure the *explicit* and *implicit bias* we split each dataset in a training and test set, train the machine learning model on the training set, and then assess the accuracy and fairness on the test set. We generate all the counterfactuals to assess the *explicit bias* as well as the *implicit bias* on the test set. For each dataset we compare three situations: the accuracy and fairness of the model trained with the sensitive attribute (1), the accuracy and fairness of the model trained without the sensitive attribute (2) and the accuracy and fairness of the model without the sensitive attribute and $PreCoF_1$ (3). We expect the accuracy to go down and the fairness to go up going from situation 1 to situation 3 but the exact trade-off may differ per dataset. We calculate the fairness by measuring the demographic disparity, which is also equal to D_{all}.

4.5 Counterfactual methodology

As described in Sects. 4.2 and 4.3, we do not use an existing counterfactual explanation method but develop one ourselves to check for every attribute whether it results in a class change. We use this approach instead of an existing counterfactual explanation method to constrain our method to check every attribute, and hence we have a guarantee that any attribute that more often results in a class change for one group than for another is found.

There exist plenty of counterfactual explanation methods already, and they can lead to different explanations as the optimization problem is set up in a different way (Bordt et al., 2022). Even a single counterfactual explanation method could lead to a large number of explanations, where the choice of parameters (e.g., the distance metric) could determine



(a) $PreCoF$: attributes in the counterfactual explanations

(b) $PreCoF_c$: attribute values in the counterfactual explanations

Fig. 1 Difference in PreCoF for men and women in the Adult Income dataset

which explanations are returned first. This abundance of explanations is not desirable in an adversarial context, as the adversary (in this case the model developer) has considerable freedom to choose which explanation it would return and as such hide bias (Irvine et al., 2020; Bordt et al., 2022). This is why we use our own counterfactual explanation method: it will not rely on any input parameter that can be manipulated, and neither it will depend on which explanations are returned first as it will check all the attributes, even after several possible explanations are already found. This approach is needed to make tangible statements about whether there is explicit bias, or whether attributes are more often present in the explanations of one group than the other. A drawback of our method is that we limit ourselves to explanations with one feature only, as we do a complete search.

Note that in spite of this reasoning, we did also compare the results found with our counterfactual explanation method with the results when using NICE (Brughmans & Martens, 2021) as counterfactual explanation method. We see that in general the same patterns are found, i.e. the same direction of explicit bias and the same *PreCoF* attributes, but that our method is better to detect all cases of *explicit bias* and is better suited to make robust statements about the occurrence of each attribute.

Several works list *actionability* and *plausibility* (adherence to data manifold) as desirable properties of counterfactual explanations (Guidotti, 2022; Karimi et al., 2021; Verma et al., 2020, 2021). These are two distinct concepts where the former restricts actions to those that are *possible to do*, and the latter requires that the resulting counterfactual instance is *realistic* or in line with the data manifold (Karimi et al., 2021). We will not take these two properties into account, which is out of line with the *algorithmic recourse* literature: focusing on *actionability* and *plausibility* can actually decrease the ability of our metric to detect bias. After all, our goal is not to look for realistic and actionable advice but to show how the model might be discriminating. For example, the counterfactual explanations to change your race or gender are not *actionable*, however, they are valuable to show explicit bias in the model. Wachter (2022) shows that when immutable characteristics form the basis for decision-making, the decision is likely to be based on undue stereotyping and protection should be offered. That is exactly what we seek to find, while allowing both actionable and immutable features to occur in the explanations. Likewise, imagine a dataset for hiring decisions where all the men are tall and all the women are small: if we want *plausible* counterfactual explanations, women cannot receive the explanation that they should be taller because this will be out of the data manifold. However, in our case, this is, once more, exactly what we are interested in to detect implicit bias.

5 Results

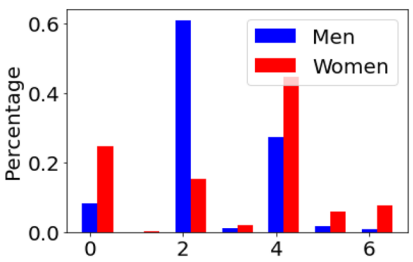
5.1 Adult Income dataset

When looking at the positive rate of both men and women in Table 2, we see that men have a higher positive rate both before and after removing the sensitive attribute. When we investigate the *explicit bias* of the model (and train the model with the sensitive attribute), we see that the explanation: *'If you would have been a man, you would have been predicted to have a high income'* is present 13 times, while the reverse explanation (*'If you would have been a woman, you would have been predicted to have a high income'*) is only present once. Afterwards, we investigate the implicit bias of the model trained without the sensitive attribute. When we compare the explanations between men and women in Fig. 1a, we see that women more often receive the explanation *marital-status*. When we look at the exact feature values of the explanations received in Fig. 1b, so the value of that feature they should change to in order to receive a favorable outcome, we see that the explanations *Marital status: Married to a civilian spouse* and *Relationship status: Husband* are much more prevalent for women than for men. The latter clearly is a proxy, as we see in Fig. 2b, that this value is only present for men. As we see in Fig. 2a, the value *Marital status: Married to a civilian spouse* is also present more often for men than for women. Whether marital status is a reasonable attribute to explain the difference in income between men and women, is not for us to decide.

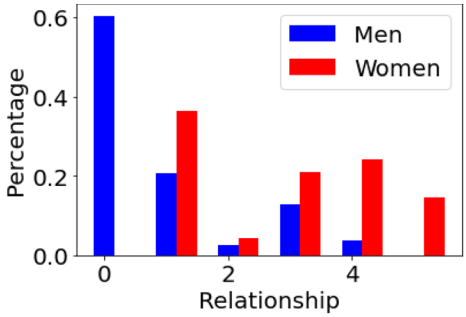
We see in Table 2 that the demographic disparity becomes even larger when we remove the sensitive attribute, which is an example of one of the cases where removing the sensitive attribute hurts the *protected group*. When we also remove $PreCoF_1$ (*marital status*) it decreases slightly but still remains very large.

5.2 Catalonia Juvenile dataset

We first use our metric to detect *explicit bias* in the model trained with the sensitive attribute. There are 7 foreigners (out of 28) that receive the explanation: *'If you would have*



(a) 0 = Divorced, 1 = Married to AF, 2= Married to Civ. Spouse, 3= Married to Abs. Spouse, 4 = Never married, 5 = Separated, 6 = Widowed



(b) 0 = Husband, 1 = Not in a family, 2 = Other relatives, 3 = Own children, 4= Unmarried, 5 = Wife

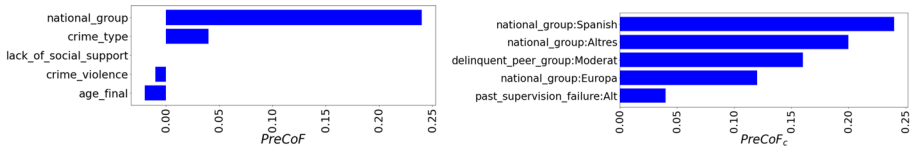
Fig. 2 Relationship between sex and the attributes marital status/relationship

Table 2 Accuracy and fairness metrics for the model trained on on the Adult Income dataset

	Situation 1 Model with sensitive attribute	Situation 2 Model without sensitive attribute	Situation 3 Model without sensitive attribute and $PreCoF_1$
Demographic disparity (Positive rate unprotected group - positive rate protected group)	**0.170 (0.242−0.073)**	**0.171 (0.242−0.071)**	**0.168 (0.236−0.068)**
Accuracy of the model	*86.28%*	*86.23%*	*86.30%*

Table 3 Accuracy and fairness metrics for the model trained on the Catalonia juvenile dataset

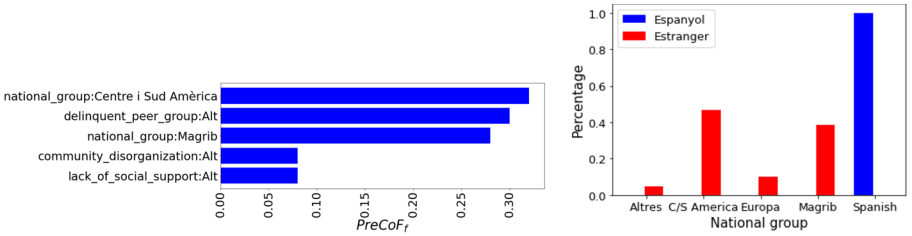
	Situation 1 Model with sensitive attribute	Situation 2 Model without sensitive attribute	Situation 3 Model without sensitive attribute and $PreCoF_1$
Demographic disparity (Positive rate unprotected group - positive rate protected group)	**0.175 (0.897−0.723)**	**0.119 (0.812−0.752)**	**0.010 (0.78−0.772)**
Accuracy of the model	*71.98%*	*72.37%*	*70.82%*



(a) $PreCoF$: attributes in the counterfactual explanations (b) $PreCoF_c$: attribute values in the counterfactual explanations

Fig. 3 Difference in *PreCoF* for foreigners and locals in the Catalonia Juvenile dataset

been a local, you would have been predicted to not reoffend' and the reverse case never happens. We also see in Table 3, that there is a large demographic disparity in Situation 1 (the model trained with the sensitive attribute). When we remove the sensitive attribute, the demographic disparity goes down but foreigners (*Estrangers*) are still disadvantaged as they are more likely to be predicted to reoffend by our model, compared to locals (*Espagnols*). When we look at the explanations in Fig. 3a, we see that *national group* is much more present in the explanations of foreigners than in the explanations of locals. As can be seen in Fig. 4b, this is a clear proxy for foreign status and should also be deleted when race attributes are not allowed. When we zoom in on the feature values in the explanations in Fig. 3b, we also see which values of national group occur most in the explanations. We see that foreigners are proportionally most likely to receive the explanation to



(a) $PreCoF_f$: attribute values of the factual instances

(b) Relationship national group - foreign status

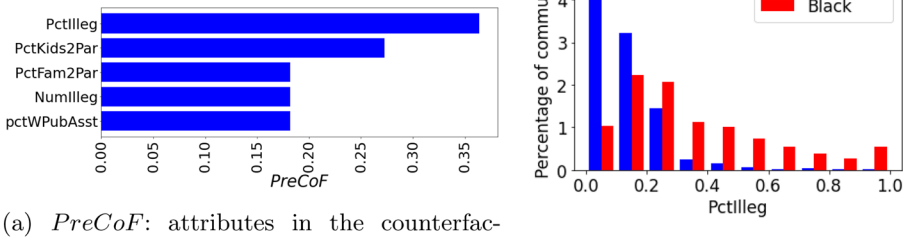
Fig. 4 Catalonia Juvenile dataset: analysis

change to *national group: Spanish* in comparison with locals, as it is a proxy for being local. Other national groups that often occur are *Altres* and *Europa*. When we look at the values occurring most often in the factual instances that receive such a class change in Fig. 4a, the national groups *Central and South America* and *Magrib* are among the most present. Hence, in this case, *PreCoF* succeeds in flagging proxy attributes which could be very helpful for deciding which attributes should be omitted from models.

We see in Table 3 that the demographic disparity goes down when removing the sensitive attribute, but nevertheless still remains quite large. When we also remove $PreCoF_1$ (*national group*), the demographic disparity almost disappears. The accuracy also goes down when removing this attribute but only slightly.

5.3 Crime and communities dataset

We find no cases of *explicit bias* in the model trained with the sensitive attributes. Next, we train a model without the sensitive attribute and assess the implicit bias. We see in Table 4 that the not-black communities in the test set are never predicted to be a violent community so their positive rate is 100 %. Black communities are predicted to be violent in around 4.5% of the cases. We hence have only explanations for the protected group, so we will just see which explanations were the most present for this group. In Fig. 5a, we



(a) $PreCoF$: attributes in the counterfactual explanations (all the attributes in this dataset are numerical so no need to investigate $PreCoF_c$)

(b) Relationship between *PctIlleg* and Black in the Crime and Communities dataset

Fig. 5 Crime and Communities dataset: analysis

Table 4 Accuracy and fairness metrics for the model trained on the Crime and Communities dataset

	Situation 1 Model with sensitive attribute	Situation 2 Model without sensitive attribute	Situation 3Model without sensitive attribute and $PreCoF_1$
Demographic disparity (Positive rate unprotected group - positive rate protected group)	**0.045 (1−0.955)**	**0.035 (1−0.965)**	**0.035 (1−0.965)**
Accuracy of the model	*84.97%*	*85.14%*	*84.81%*

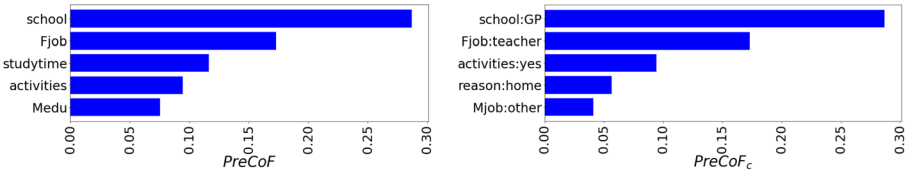
observe that the attribute *PctIlleg*, which is the percentage of kids born to people who were never married, is the most present. When we look at the distribution of this attribute for black and non-black communities in Fig. 5b, we indeed see that this percentage tends to be higher for black communities. Research on other models trained on this dataset also find this to be an important predictor of both the target value (violent community) as well as the sensitive attribute (black community) (Le Quy et al., 2022). When we assess the other top attributes in *PreCoF*, we notice that the four first are related to families with both parents being present, or being married. Earlier research already argued that marriage is linked to a reduction in crime (Sampson et al., 2006).

We also see in Table 4 that the demographic disparity goes down when we remove the sensitive attribute. It does not go down when we remove $PreCoF_1$, which makes sense as the $PreCoF_1$ attribute here (*PctIlleg*) is very correlated with other attributes of the dataset such as *NumIlleg*.

5.4 Student performance dataset

We see in Table 5 that our classifier predicts girls to be less likely to have a positive label compared to boys. Although they have on average a higher score than boys, they are more often predicted to fail in every situation. We might get some insights into this phenomenon by looking at how the explanations differ for both groups. We see in Fig. 6a that the attribute *school* is present more often in the explanations for girls and in Fig. 6b that they receive the explanation to change to *school GP* more often. Depending on what the machine learning model is used for, this kind of analysis could give very important insights. If this model would be used, for example, to determine whether the students would be successful in university and should be admitted, this analysis shows that girls could be disadvantaged compared to boys because of the school they went to. When we look at the *explicit bias* in the model trained with the sensitive attribute, boys are biased against: there are three boys that receive the explanation: *'If you would have been a girl, you would have been predicted as scoring above average instead of below'* and the reverse does not happen. This example shows that *explicit bias* and *implicit bias* can work in opposite ways.

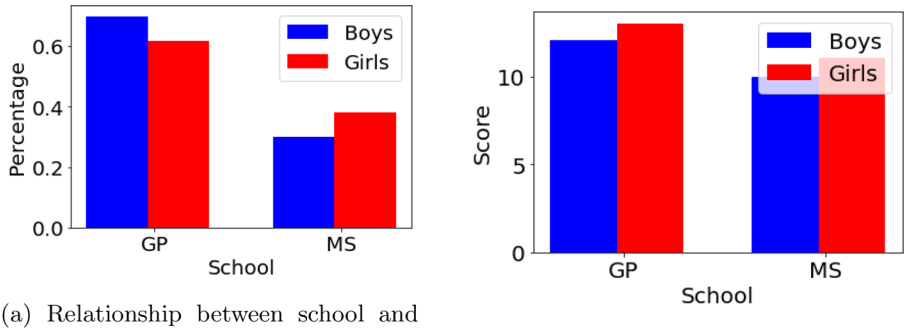
We analyse the relations of the attribute *school*. We see in Fig. 7b that for both boys and girls, their average score is higher if they went to school GP: for girls their average score on school GP is 13 and on school MS 11.03, while for boys the average score on school GP is 12.03 and on school MS 9.95. The average score of girls is also higher independent of school: on average girls have a score of 12.25 and boys of 11.41. When researching this attribute, we see in Fig. 7a that girls more often go to *school MS* which



(a) $PreCoF$: attributes in the counterfactual explanations

(b) $PreCoF_c$: attribute values in the counterfactual explanations

Fig. 6 Difference in explanations for boys and girls in the Student performance dataset



(a) Relationship between school and sex in the Student Performance dataset: percentage of this sex that goes to this school

(b) Relationship between school and sex in the Student Performance dataset: average grade

Fig. 7 Student performance dataset: analysis

has a lower average score, so they receive the explanation to change to *school GP*, which has a higher average score, more often. So due to the importance of the attribute school in the machine learning model, they are predicted to fail more often than boys while their true outcome is to fail less. The importance of the school you go to in a machine learning model to predict grades reminds of a recent case in England in 2020, where an algorithm designed to predict grades for A-level exams amidst COVID-19 increased the predicted grades at small private schools but lowered the grades at larger, state-run schools that have a larger proportion of minority students (Wachter et al., 2020). In terms of accuracy, this model performed well but as a result high performing students from 'good schools' received high marks, whereas highly performing students from 'bad schools' had their marks capped by the lower performance of classmates and got a lower mark than deserved (Wachter et al., 2020). This system was not biased on purpose: it was the ignorance of the social bias that led to the technical bias in this system (Wachter et al., 2020).

We compare the accuracy and fairness of the three situations in Table 5: We see that the accuracy of the model goes down after removing attributes, however only slightly. We see that the demographic disparity increases after removing the gender attribute, which makes sense as girls on average scored better but are disadvantaged by the school they go to: this effect will become even larger if gender information is removed. There is *explicit bias* against boys, but *implicit bias* against girls through the neutral attribute *school*. If we remove $PreCoF_1$ *School*, the demographic disparity will decrease again but

Table 5 Accuracy and fairness metrics for the model trained on the Student Performance Dataset

	Situation 1 Model with sensitive attribute	Situation 2 Model without sensitive attribute	Situation 3 Model without sensitive attribute and $PreCoF_1$
Demographic disparity (Positive rate unprotected group - positive rate protected group)	**0.043 (0.610−0.566)**	**0.115 (0.646−0.531)**	**0.066 (0.659−0.593)**
Accuracy of the model	*73.85%*	*71.28%*	*70.26%*

not until the first level. This situation shows that as mentioned in literature already (Corbett-Davies & Goel, 2018) and as seen in other datasets, removing the sensitive attribute can increase the discrimination in the dataset.

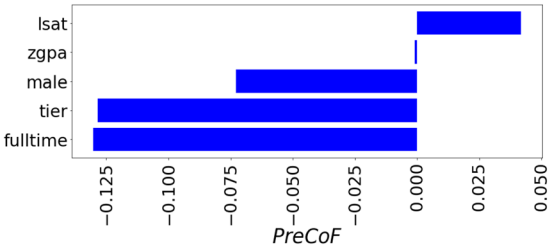
5.5 Law admission dataset

When we look at the *explicit bias*, we see that there are 45 instances in the test set that receive the explanation: *'If you would have been white, you would have been predicted as admitted to pass the bar'* and only 3 the other way around, which shows that the model that is trained with the sensitive attribute exhibits *explicit bias*. This also shows that the model is non-linear and both parties can receive such explanations. When we train the model without the sensitive attribute here, we see in Fig. 8a that the only attribute that is relatively more present in the explanations of *Non-Whites* compared to *Whites*, is the *lsat score*. The fact that almost all the attributes are relatively more present in the explanations of the unprotected group means that the rejected individuals in this group are closer to the decision boundary: Changing only one attribute more often leads to a change in outcome, while for the protected group more attribute changes are necessary. It is not surprising that *lsat scores* pop up as $PreCoF_1$ as it is often said that test scores such as GPA and LSAT are racially biased: white test-takers consistently score higher than minority test-takers (White, 2000) and there have been calls for law school admission committees to deemphasize reliance on LSAT scores and to develop new methodologies to assess the skills of each applicant (Hill, 2019). When we look at Fig. 8b, we indeed see that the average score of the LSAT is higher for *Whites* compared to *Non-Whites*.

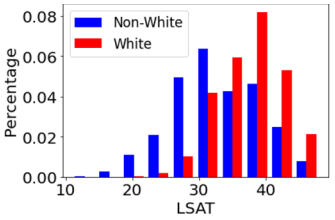
When we compare the accuracy and fairness of the three situations in Table 6, we see that the accuracy decreases when removing the sensitive attribute and $PreCoF_1$. When removing the sensitive attribute, the demographic disparity decreases slightly but after removing $PreCoF_1$, it decreases substantially. The question can be asked here whether we deem it fair that there is a difference in positive rate based on LSAT scores: are these objective scores or are they already biased in se?

6 Discussion

In this study, we use counterfactual explanations to shed light on which discrimination occurred in models trained on some well-known datasets, both in terms of *explicit* and *implicit bias*. Our experiments reveal that removing $PreCoF_1$, will decrease the



(a) *PreCoF*: attributes in the counterfactual explanations



(b) Relationship between *Race* and *LSAT* in Law Admission Dataset

Fig. 8 Law Admission dataset: analysis

Table 6 Accuracy and fairness metrics of the model trained on the Law Admission dataset

	Situation 1 Model with sensitive attribute	Situation 2 Model without sensitive attribute	Situation 3 Model without sensitive attribute and $PreCoF_1$
Demographic disparity (Positive rate unprotected group - positive rate protected group)	0.159 (0.994−0.835)	0.143 (0.990−0.847)	0.075 (0.987−0.912)
Accuracy of the model	*89.94%*	*89.82%*	*89.63%*

demographic disparity in a model, but we want to highlight that this is not the main purpose of our metric. It is possible that removing other attributes will decrease the demographic disparity even more as it is not the goal of the *PreCoF* metric to find that variable that would make the demographic disparity the smallest. Our purpose is not to give members of a protected group an advantage by giving them a better outcome (Wachter et al., 2020), but rather to shed light on which attributes resulted in a different outcome and jump-start a discussion on whether they are based on historical inequalities or are justified discriminators. The fairness results (i.e., the decrease in demographic disparity) simply show that removing the $PreCoF_1$ variable will result in a smaller difference in positive rate between the protected and the unprotected group, which can be a desirable outcome in some cases.

What does our technique add compared to other fairness metrics?

Fairness will depend on context-dependent judgements, so it is dangerous to treat the quantitative fairness metrics discussed in Sect. 2.1 as black-box fairness measures (Corbett-Davies & Goel, 2018). Using group metrics for fairness can abstract away more subtle issues that are too difficult to operationalize or to decide upon algorithmically (Yeom & Tschantz, 2021). There is not one criterion that can ensure fairness in all cases, and when a model fails on a fairness metric, this should lead to an investigation as to why this happens (Yeom & Tschantz, 2021). We also confirmed that just removing the sensitive attribute is not a viable approach as it can even amplify the discrimination of the model, and thus harm the group it was supposed to protect (Corbett-Davies & Goel, 2018). *Demographic Parity* can detect whether the model is treating the sensitive groups differently when the model does not directly use the protected attribute but correlated one(s), but it does not consider whether

there is sufficient justification for a disparity of outcomes (Yeom & Tschantz, 2021). Other tests that do take the ground truth into account such as *equalized odds* also just examine the disparities but not how they were reached (Yeom & Tschantz, 2021).

We do not state that removing $PreCoF_1$ to decrease the *demographic disparity* will be a universal solution to tackle the discrimination in a dataset. We just showcased that it is a possible approach. Our point of view is that this should be decided case by case: is this attribute a justified discriminator? Does it just behave as a proxy? Is it warranted to sacrifice accuracy for extra fairness? Is a difference in positive rate a problem when the true outcomes also differ per sensitive group or an accepted consequence? Do the observed outcomes accurately reflect the real world? This last question is related to the two worldviews that Friedler et al. (2016) suggested, namely the *'We are all equal'* worldview and the *'What you see is what you get'* worldviews. These are all questions that should be answered for each case individually, and our metric can help to decide upon them. The benefits of building more fair models could be very large, as fair machine learning models could dramatically improve the equality of consequential decisions (Corbett-Davies & Goel, 2018).

7 Future research and limitations

There are limitations to our metric, which at the same time pose opportunities for future research. The patterns detected by this metric will only be trustworthy if both groups in the test set are large enough. Therefore, we do not include the German Credit dataset into our experiments, as this is a very small and imbalanced dataset. The number of individuals with a bad outcome in each sensitive group in the test set will be so small that it is not possible to draw conclusions from them.

Furthermore, in the COMPAS Juvenile dataset we detect an interesting pattern; every attribute is relatively more present in the explanations of the not African-American group than in the African-American group. This pattern occurs because the 'rejected' individuals (individuals which are predicted the unfavorable outcome by the machine learning model) in the former group are on average closer to the decision boundary than the individuals in the latter group: for the latter, one attribute change will less often be enough to result in a class change. This is related to the fairness notion of CERTIFAI (Sharma et al., 2019) and algorithmic recourse (von Kügelgen et al., 2022), where the effort of both groups to reach the desired target outcome is taken into account. Our metric only looks at univariate changes for now but this could be expanded to changes of two or more attributes in future research.

In our experiments, we focus on the rejected individuals. Another interesting research avenue would be to focus on the *misclassified* rejected individuals and see what are the most occurring explanations for both sensitive groups. This could be a possible avenue to improve the model and reduce misclassifications.

Lastly, this study only takes tabular datasets into account but it will be valuable to analyze this on text and behavioral datasets, as they are very sparse. For some tabular datasets, we know what we can expect as proxies, however for behavioral datasets like Facebook likes, this might not be very intuitive. This will be the focus of our next research.

8 Conclusion

Fairness literature in AI has already revealed that AI creates new challenges for detecting discrimination: automated discrimination is less intuitive, subtle and intangible (Wachter et al., 2021). As the algorithmic world will make complex decisions without any reasoning behind them, it will be challenging to detect whether you are treated fairly. It is misguided to focus on fairness while not obtaining transparency first (Rudin et al., 2018). We aim to provide this transparency by providing a tool that can shed light on: how often *explicit bias* in the decision making model occurs for each subgroup, and which factors are a cause of the *implicit bias* in the decision making model in each subgroup.

Appendix A: Used datasets

UCI Adult dataset

The Adult Income dataset,[4] or 'Census Income' dataset contains information extracted from the 1994 census data with as target variable whether the income of a person exceeds $50,000 a year or not. We use it to assess whether there are gender or race inequalities present in people's annual incomes (Asuncion & Newman, 2007). The Adult dataset contains 48,842 instances with 14 features. As is common in literature, we drop the features *Fnlwgt* as it does not convey a meaning to its values, *EducationNum* as it has the same meaning as *Education* and *NativeCountry* as it has a lot of missing values. We use the features *Age*, *Workclass*, *Education*, *Marital-status*, *Occupation*, *Relationship*, *Race*, *Sex*, *CapitalGain*, *CapitalLoss* and *HoursPerWeek*. The sensitive attributes in this dataset are *Race* and *Sex*. For our experiments we use *Sex* as the protected attribute. The favorable outcome in this dataset is having an income that exceeds $50,000 a year, the unfavorable outcome is having a yearly income below $50,000.

Catalonian Juvenile Dataset

This dataset[5] consists of juvenile offenders who were incarcerated in the juvenile justice system of Catalonia and who were released in 2010 (Miron et al., 2021). Their recidivism behavior was observed between 2010 and 2015. SAVRY is a tool developed in 2003 which predicts recidivism (Miron et al., 2021). We build a model on most of the individual and criminological variables as in Miron et al. (2021),[6] but we also include the variables that are used in the SAVRY risk scores such as *History of self harm*, *Delinquent peer group*,... Our dataset contains 855 instances with 22 attributes. The target variable in this dataset is *Recid*, which is whether the offender has re-offended or not. The favorable outcome here is that there is no recidivism, the unfavorable outcome that there is. The sensitive attributes in this dataset are *Foreigner*, *Sex* and *National Group* of the offenders, but for our experiments we use *Foreigner* as protected attribute.

[4] https://github.com/EpistasisLab/pmlb/tree/master/datasets/adult
[5] https://github.com/nkundiushuti/savry/blob/master/dat/reincidenciaJusticiaMenors.csv
[6] https://github.com/nkundiushuti/savry/blob/master/Savry_Fair.ipynb

Crime and communities dataset

This dataset[7] contains 1994 samples of socio-economic data from the United States. There are 127 attributes in this dataset, but we delete all attributes related to state, race or crime, except for the target variable, so that 91 attributes remain. The target variable is whether the attribute *ViolentCrimesPerPop* is above a certain treshold, which then constitutes a violent community. In line with literature, we also add the attribute *Black* in order to divide the communities in black and non-black communities when the attribute *racepctblack* is above a certain threshold (Kamiran et al., 2013; Le Quy et al., 2022). The protected attribute here is *Black*.

Student performance dataset

This dataset[8] consists of 649 students and 30 attributes from a Portuguese high school (Cortez & Silva, 2008). The attributes of the dataset contain information about the background of the students and their social activities. As commonly done (Hamoud, 2016), we delete the results from the first and the second grade (*G1,G2*) as they are very heavily correlated with the final grade (*G3*). The target variable is scoring above average on their final exam of Portuguese, where the favorable outcome is that you score above average and the unfavorable outcome that you score below average. The protected attribute in this dataset is *Sex*.

Law admission dataset

This dataset[9] contains a Law School Admission Council (LSAC) survey conducted across 163 law schools in the United States in 1991 (Wightman, 1998). The dataset consists of 20,798 students and the following attributes: *decile1b*,*decile3b*, *lsat*, *ugpa*, *zfygpa*, *zgpa*, *fulltime*, *fam_inc*, *male*, *tier*, *race* and *pass_bar*. The target variable is whether the student will pass the bar exam or not. The protected attribute in this dataset is *Race*: 92.1% of white students pass the bar exam, while this ratio in non-white students is only 72.3%.

Appendix B: *PreSHAPF*

Alternative XAI techniques can also be employed to investigate the presence of implicit bias in a machine learning model. In this section, we use SHAP values as a means of examining disparities between two sensitive groups. The results can differ as SHAP values focuses on variations in prediction scores, rather than on decisions (which basically is the combination of a prediction score and threshold). SHAP values are a computationally efficient way to calculate Shapley values, which are defined as the average marginal contribution across all possible coalitions (Lundberg & Lee, 2017).

SHAP attributes to each feature the change in the expected model prediction when conditioning on that feature and thus reveals the extent to which each feature contributes to the

[7] https://github.com/tailequy/fairness_dataset/blob/main/experiments/data/communities_crime.csv
[8] https://archive.ics.uci.edu/ml/datasets/student+performance
[9] https://github.com/tailequy/fairness_dataset/blob/main/experiments/data/law_school_clean.csv

prediction score, either positively or negatively (Lundberg & Lee, 2017). As with *PreCoF*, we focus on the negatively affected members of both the protected and the unprotected group. We compare the mean SHAP values in both subgroups and *PreSHAPF* (*Predictive SHAP Fairness*) will reveal for which features the difference between both subgroups is the largest.[10] There are two main differences between *PreCoF* and *PreSHAPF*: First, *PreCoF* focuses on the decisions made by the model, while *PreSHAPF* focuses on the prediction scores. Second, *PreCoF* returns features (and $PreCoF_c$ the feature values for the categorical features), while *PreSHAPF* will always return feature values for the categorical features.

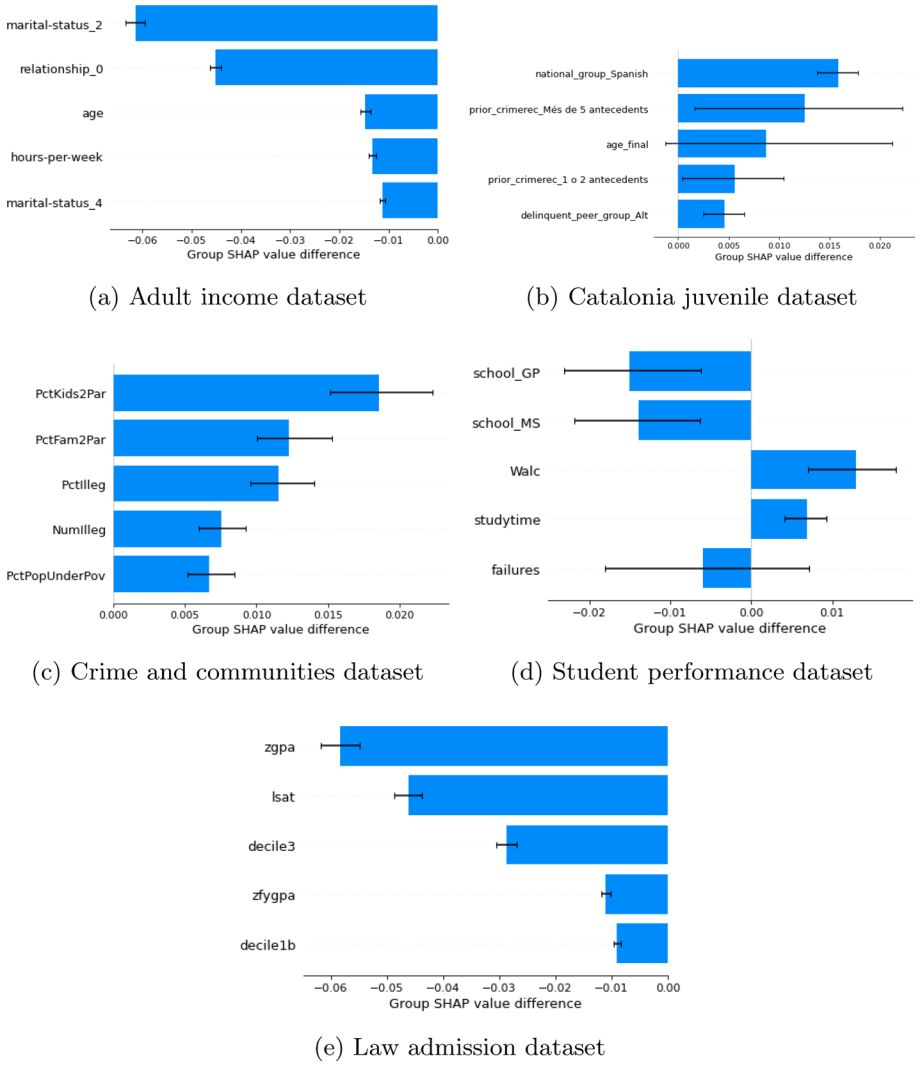
*PreCoF*vs *PreSHAPF*

The results for the datasets used in this paper can be found in Fig. 9. For each dataset, we calculate *PreSHAPF* as the discrepancy in mean SHAP values between both subgroups. As demonstrated in Fig. 9, the most salient patterns detected with *PreCoF* are also present in *PreSHAPF*, however, slight variations are observed as they measure distinct phenomena.

As depicted in Fig. 9a, the two features with the largest difference in *PreSHAPF*, namely *relationship: 0* and *marital-status: 2*, correspond to the feature values with the highest value in $PreCoF_c$, as can be seen in Fig. 1b. These features, on average, negatively impact the prediction score of women (to be predicted to have a high income) compared to men. The other top features are different between *PreCoF* and *PreSHAPF*. In Fig. 9b, we observe that the feature *national group: Spanish*, which was the $PreCoF_c$ attribute in Fig. 3b, is the feature value with the highest value in *PreSHAPF*. For foreigners, this feature, on average, has a larger positive impact on the prediction score (to be predicted to recidive) compared to locals. However, in Fig. 3b, we see that the other values for *national group*, namely *Altres* and *Europa*, are also high ranked in *PreCoF*, but they are not among the top features in *PreSHAPF*. As illustrated in Fig. 9c, the features with the highest *PreSHAPF* value are the same as the PreCoF attributes (*PctIlleg*, *PctKids2Par*, *PctFam2Par*, *NumIlleg*) in Fig. 5a in a slightly different order. Figure 9d shows that both values of *School* have the highest value in *PreSHAPF*. These attributes (on average) negatively impact the prediction score (to be predicted a good student) of girls compared to boys This is in line with the results of *PreCoF*, as *School* was the *PreCoF* attribute in Fig. 6a, but the other attributes differ. In Fig. 9e, we observe that all features, on average, have a negative impact on the prediction score (to be predicted to pass the bar) of Non-Whites compared to Whites. The two features for which this discrepancy is the largest are *zgpa* and *lsat*, which were also the two attributes with the largest value in *PreCoF* in Fig. 8a.

Overall, we notice that the global patterns seem consistent over *PreCoF* and *PreSHAPF*. However, the less important features can vary strongly, which shows that *PreCoF* and *PreSHAPF* function differently. When we change the threshold of the machine learning classifier trained on these datasets, the results of *PreCoF* will strongly change (for some thresholds, all of the top features are different), while this will have no effect on the results of *PreSHAPF*. We add a supplementary illustration which shows the effect of the decision threshold on *PreCoF* and *PreSHAPF* in Sect. 2.

[10] We use the SHAP package TreeExplainer to calculate the SHAP values (as we are explaining a random forest) (Lundberg et al., 2020) We use the group difference plots provided by SHAP to graph the difference in mean SHAP values between the two subgroups (Lundberg et al., 2020).



(a) Adult income dataset

(b) Catalonia juvenile dataset

(c) Crime and communities dataset

(d) Student performance dataset

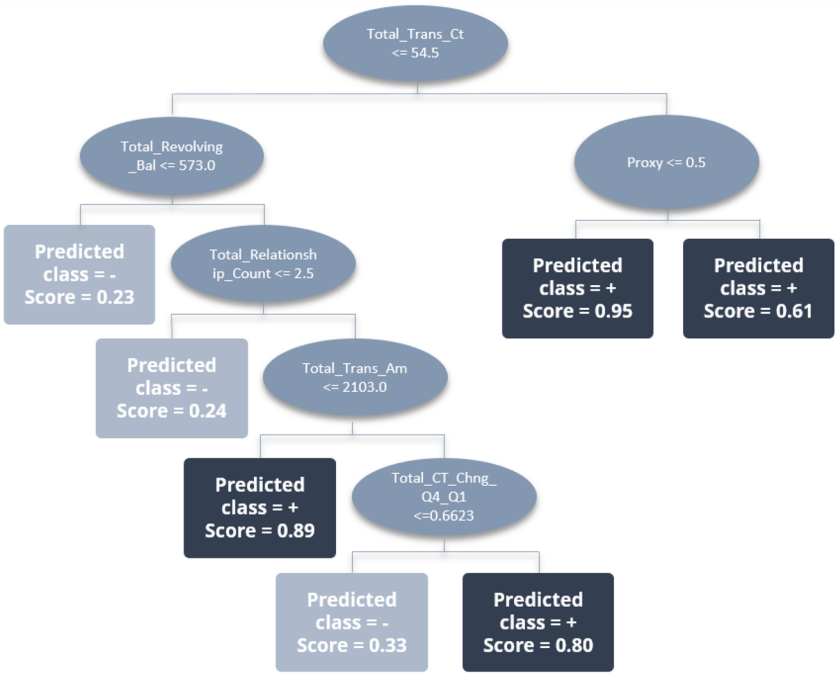
(e) Law admission dataset

Fig. 9 *PreSHAPF*

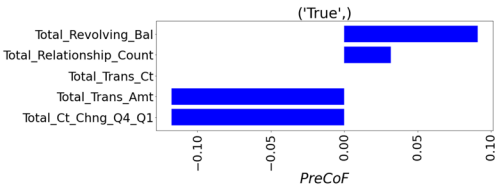
PreCoF versus *PreSHAPF* on a transparent model

To further demonstrate the functionality of *PreCoF* and its distinction with SHAP values, we present an additional illustration using an existing churn dataset set.[11] This dataset aims to predict whether a bank customer will churn or not, where the unfavorable outcome is that the customer will attrite, and the favorable outcome that the customer will remain loyal. The dataset does not contain a sensitive attribute, but we artificially introduce this

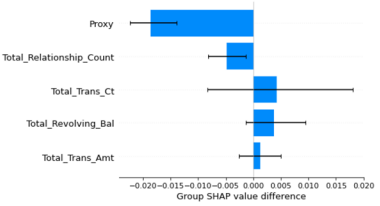
[11] https://www.kaggle.com/datasets/syviaw/bankchurners.



(a) Decision tree, where the unfavorable outcome is listed as −, and the favorable outcome as +.



(b) *PreCoF*



(c) *PreSHAPF*

Fig. 10 Additional illustration with a transparent machine learning model to show the difference between *PreCoF* and *PreSHAPF*

aspect to the dataset, randomly assigning half of the instances the gender of male and half of the instances the gender of female. In contrary to our previous experiments, we use an interpretable decision tree (with a restricted number of 7 leaf nodes) to provide insight into the model's functioning and to facilitate a comparison of how counterfactual explanations and SHAP values detect bias within the model.

To investigate the implicit bias, we add a proxy that is correlated with the target outcome and gender. This action is likely to result in the model picking up this biased pattern, even after we remove the sensitive attribute (gender) and may result in gender discrimination in the model's predictions. As in our previous experiments to detect implicit bias, we remove the sensitive attribute (*gender*) from the data, split the data into a training and

test set, and fit a machine learning on the training set. However, in this scenario, we use a simple decision tree, as opposed to a Random Forest model, to compare the results from *PreCoF* and *PreSHAPF* with the actual model, as depicted in Fig. 10a.

These results clearly illustrate how counterfactual explanations and SHAP values function differently. The *proxy* has a large impact on the prediction score, but will not have an effect on the decision for any of the instances (both leaf nodes after the biased feature split result in the same outcome as the threshold is 0.5). When using *PreSHAPF*, we see in Fig. 10c that the feature with the largest value is *proxy*. This makes sense, as we see in the decision tree, that it has a large effect on the prediction score and we know that it is correlated to gender. On the other hand, in Fig. 10b, *PreCoF* does not report this feature as it does not change the decision for any of the instances. If the threshold of the machine learning classifier changes to 0.7 or 0.8, *PreCoF* does report *proxy* as the top feature.

These results indicate that both SHAP values and counterfactual explanations are well-suited to identify patterns of indirect discrimination, but that they measure distinct phenomena. Their outcomes may vary as counterfactual explanations focus on decisions and SHAP values on prediction scores. In this paper, we use counterfactual explanations as our focus is on the actual decisions people receive, but using SHAP values is a good alternative when the focus is on fair scoring (for example with a varying or unfixed threshold). Our main argument that a deeper understanding of the nature of the bias is necessary before deciding on a method to address it, remains valid when using both XAI techniques. Finally, our experiments further confirm that both *PreCoF* and *PreSHAPF* are detecting bias in the model, and not in the underlying data. If a biased feature is added to the dataset but not picked up by the model, neither *PreCoF* or *PreSHAPF* will show this biased feature.

Author contributions Data preparation and code analysis were performed by Sofie Goethals. All authors worked on the text equally and approved the final manuscript.

Funding Funding was provided by Research Foundation-Flanders (Grant No.11N7723N) and by Flanders AI Research Program.

Availability of data and materials The data used in this paper are publicly available and appropriate references are cited for those datasets in Sect. 1.

Code availability The Python implementation of the proposed metric is available through: https://github.com/ADMAntwerp/PreCoF.

Declarations

Conflict of interest Not applicable.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160.

Asuncion, A., & Newman, D. (2007). *UCI Machine Learning Repository*.

Black, E., Yeom, S., & Fredrikson, M. (2020). Fliptest: Fairness testing via optimal transport. In: *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 111–121).
Bonchi, F., Hajian, S., Mishra, B., & Ramazzotti, D. (2017). Exposing the probabilistic causal structure of discrimination. *International Journal of Data Science and Analytics, 3*(1), 1–21.
Bordt, S., Finck, M., Raidl, E., & von Luxburg, U. (2022). *Post-hoc explanations fail to achieve their purpose in adversarial contexts*. arXiv preprint arXiv:2201.10295 .
Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
Brughmans, D., & Martens, D. (2021). In: *Nice: An algorithm for nearest instance counterfactual explanations*. arXiv preprint arXiv:2104.07411 .
Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In: *Conference on fairness, accountability and transparency* (pp. 77–91).
Calders, T., Kamiran, F., & Pechenizkiy, M. (2009). Building classifiers with independency constraints. In: *2009 IEEE international conference on data mining workshops* (pp. 13–18).
Caton, S., & Haas, C. (2020). In: *Fairness in machine learning: A survey*. arXiv preprint arXiv:2010.04053
Corbett-Davies, S., & Goel, S. (2018). In: *The measure and mismeasure of fairness: A critical review of fair machine learning*. arXiv preprint arXiv:1808.00023
Cortez, P., & Silva, A. M. G. (2008). In: *Using data mining to predict secondary school student performance* (pp. 5–12) EUROSIS-ETI.
Craven, M., & Shavlik, J. (1995). Extracting tree-structured representations of trained networks. *Advances in Neural Information Processing Systems, 8*, 24–30.
Doyle, O. (2007). Direct discrimination, indirect discrimination and autonomy. *Oxford Journal of Legal Studies, 27*(3), 537–553.
Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In: *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214–226).
Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In: *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 259–268).
Fernandez, C., Provost, F., & Han, X. (2020). Counterfactual explanations for data-driven decisions. In: *40th international conference on information systems*, ICIS 2019.
Fleisher, W. (2021). What's fair about individual fairness? In: *Proceedings of the 2021 AAAI/ACM conference on AI, ethics, and society* (pp. 480–490).
Friedler, S.A., Scheidegger, C., & Venkatasubramanian, S. (2016). In: *On the (im) possibility of fairness*. arXiv preprint arXiv:1609.07236
Fryer, R. G., Jr., Loury, G. C., & Yuret, T. (2008). An economic analysis of color-blind affirmative action. *The Journal of Law, Economics, & Organization, 24*(2), 319–355.
Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., & Huang, J. (2020). Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? In: *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 325–336).
Gohel, P., Singh, P., & Mohanty, M. (2021). *Explainable AI: Current status and future directions*. arXiv preprint arXiv:2107.07045 .
Guidotti, R. (2022). Counterfactual explanations and how to find them: Literature review and benchmarking. *Data Mining and Knowledge Discovery*, 1–55.
Hamoud, A. (2016). Selection of best decision tree algorithm for prediction and classification of students' action. *American International Journal of Research in Science, Technology, Engineering & Mathematics, 16*(1), 26–32.
Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems, 29*, 3315–3323.
Hill, L. (2019). Less talk, more action: How law schools can counteract racial bias of LSAT scores in the admissions process. *University of Maryland Law Journal of Race, Religion, Gender & Class, 19*, 313.
Irvine, CA, USA. Barocas, S., Selbst, A. D., & Raghavan, M. (2020). The hidden assumptions behind counterfactual explanations and principal reasons. In: *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 80–89).
Johnson, G. M. (2021). Algorithmic bias: On the implicit biases of social technology. *Synthese, 198*(10), 9941–9961.
Joshi, S., Koyejo, O., Vijitbenjaronk, W., Kim, B., & Ghosh, J. (2019). *Towards realistic individual recourse and actionable explanations in black-box decision making systems*. arXiv preprint arXiv:1907.09615 .
Kamiran, F., & žliobaitÿe, I. (2013). Explainable and non-explainable discrimination in classification. Discrimination and privacy in the information society (pp. 155–170) Springer.

Kamiran, F., žliobaitÿe, I., & Calders, T. (2013). Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems, 35*(3), 613–644.

Karimi, A. -H., Barthe, G., Schölkopf, B., & Valera, I. (2021). A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. In: *ACM Computing Surveys (CSUR)*.

Kearns, M., Neel, S., Roth, A., & Wu, Z.S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In: Dy, J. & Krause, A. (Eds.) *Proceedings of the 35th international conference on machine learning* (Vol. 80, pp. 2564–2572). PMLR. Retrieved from https://proceedings.mlr.press/v80/kearns18a.html

Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. *Advances in Neural Information Processing Systems*, *30*.

Kim, M., Reingold, O., & Rothblum, G. (2018). Fairness through computationally-bounded awareness. *Advances in Neural Information Processing Systems*, *31*.

Kim, P. T. (2017). Auditing algorithms for discrimination. *University of Pennsylvania Law Review Online, 166*, 189.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). In: *Inherent trade-offs in the fair determination of risk scores*. arXiv preprint arXiv:1609.05807 .

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in Neural Information Processing Systems*, *30*.

Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., & Ntoutsi, E. (2022). *A survey on datasets for fairness-aware machine learning*, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery (p. e1452).

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

Lundberg, S. M., & Lee, S. -I. (2017). A unified approach to interpreting model predictions. In: *Proceedings of the 31st international conference on neural information processing systems* (pp. 4768–4777).

Martens, D., Baesens, B., Van Gestel, T., & Vanthienen, J. (2007). Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research, 183*(3), 1466–1476.

Martens, D., & Provost, F. (2014). Explaining data-driven document classifications. *MIS Quarterly, 38*(1), 73–100.

Miron, M., Tolan, S., Gómez, E., & Castillo, C. (2021). Evaluating causes of algorithmic bias in juvenile criminal recidivism. *Artificial Intelligence and Law, 29*(2), 111–147.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Pearl, J., et al. (2000). *Models, reasoning and inference* (p. 19). Cambridge, UK: Cambridge University Press.

Pedreshi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. In: *Proceedings of the 14th ACM SIGKDD International conference on knowledge discovery and data mining*, pp. 560–568.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).

Ritov, Y., Sun, Y., & Zhao, R. (2017). In: *On conditional parity as a notion of nondiscrimination in machine learning*. arXiv preprint arXiv:1706.08519 .

Rudin, C., Wang, C., & Coker, B. (2018). In: *The age of secrecy and unfairness in recidivism prediction*. arXiv preprint arXiv:1811.00731 .

Sampson, R. J., Laub, J. H., & Wimer, C. (2006). Does marriage reduce crime? A counterfactual approach to within-individual causal effects. *Criminology, 44*(3), 465–508.

Sharma, S., Henderson, J., & Ghosh, J. (2019). In: *CERTIFAI: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models*. arXiv preprint arXiv:1905.07857 .

Sokol, K., & Flach, P. (2021). In: *Explainability is in the mind of the beholder: Establishing the foundations of explainable artificial intelligence*. arXiv preprint arXiv:2112.14466 .

Sokol, K., Santos-Rodriguez, R., & Flach, P. (2019). In: *FAT Forensics: A Python toolbox for algorithmic fairness, accountability and transparency*. arXiv preprint arXiv:1909.05167 .

Verma, S., Dickerson, J., & Hines, K. (2020). In: *Counterfactual explanations for machine learning: A review*. arXiv preprint arXiv:2010.10596 .

Verma, S., Dickerson, J., & Hines, K. (2021). In: *Counterfactual explanations for machine learning: Challenges revisited*. arXiv preprint arXiv:2106.07756

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In: *2018 IEEE/ACM international workshop on software fairness (fairware)* (pp. 1–7).
von Kügelgen, J., Karimi, A. -H., Bhatt, U., Valera, I., Weller, A., & Schölkopf, B. (2022). On the fairness of causal algorithmic recourse. In: *Proceedings of the AAAI conference on artificial intelligence*, (Vol. 36, pp. 9584–9594).
Wachter, S. (2022). In: *The theory of artificial immutability: Protecting algorithmic groups under anti-discrimination law*. arXiv preprint arXiv:2205.01166
Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *The Harvard Journal of Law & Technology, 31*, 841.
Wachter, S., Mittelstadt, B., & Russell, C. (2020). Bias preservation in machine learning: The legality of fairness metrics under EU non-discrimination law. *West Virginia Law Review, 123*, 735.
Wachter, S., Mittelstadt, B., & Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review, 41*, 105567.
White, D. M. (2000). The requirement of race-conscious evaluation of LSAT scores for equitable law school admissions. *Berkeley La Raza Law Journal, 12*, 399.
Wightman, L. F. (1998). In: *LSAC National longitudinal bar passage study*. LSAC research report series.
Wu, Y., Zhang, L., & Wu, X. (2019). Counterfactual fairness: Unidentification, bound and algorithm. In: *Proceedings of the twenty-eighth international joint conference on Artificial Intelligence*.
Xu, R., Cui, P., Kuang, K., Li, B., Zhou, L., Shen, Z., & Cui, W. (2020). Algorithmic decision making with conditional fairness. In: *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2125–2135).
Yeom, S., & Tschantz, M. C. (2021). Avoiding disparity amplification under different worldviews. In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 273–283).
Zliobaite, I. (2015). In: *A survey on measuring indirect discrimination in machine learning*. arXiv preprint arXiv:1511.00148