Secondary Publication

Heidrich, Louisa; Slany, Emanuel; Scheele, Stephan; Schmid, Ute

FairCaipi : A Combination of Explanatory Interactive and Fair Machine Learning for Human and Machine Bias Reduction

Date of secondary publication: 02.06.2025

Version of Record (Published Version), Article

Persistent identifier: urn:nbn:de:bvb:473-irb-1084942

Primary publication

Heidrich, Louisa; Slany, Emanuel; Scheele, Stephan; Schmid, Ute (2023): FairCaipi : A Combination of Explanatory Interactive and Fair Machine Learning for Human and Machine Bias Reduction, in: Machine learning and knowledge extraction, Basel: MDPI, Vol. 5, Nr. 4, pp. 1519–1538, doi: 10.3390/make5040076.

Legal Notice

This work is protected by copyright and/or the indication of a licence. You are free to use this work in any way permitted by the copyright and/or the licence that applies to your usage. For other uses, you must obtain permission from the rights-holders.

This document is made available under a Creative Commons license.

The license information is available online:
https://creativecommons.org/licenses/by/4.0/legalcode

machine learning & knowledge extraction

MDPI

Article

FAIRCAIPI: A Combination of Explanatory Interactive and Fair Machine Learning for Human and Machine Bias Reduction

Louisa Heidrich [1], Emanuel Slany [1,2,*], Stephan Scheele [1,2] and Ute Schmid [1,2]

[1] Cognitive Systems, University of Bamberg, An der Weberei 5, 96047 Bamberg, Germany; louisa-marie.heidrich@stud.uni-bamberg.de (L.H.); stephan.scheele@iis.fraunhofer.de (S.S.); ute.schmid@uni-bamberg.de (U.S.)

[2] Fraunhofer Institute for Integrated Circuits IIS, Sensory Perception & Analytics, Comprehensible AI, Am Wolfsmantel 33, 91058 Erlangen, Germany

* Correspondence: emanuel.slany@iis.fraunhofer.de

Abstract: The rise of machine-learning applications in domains with critical end-user impact has led to a growing concern about the fairness of learned models, with the goal of avoiding biases that negatively impact specific demographic groups. Most existing bias-mitigation strategies adapt the importance of data instances during pre-processing. Since fairness is a contextual concept, we advocate for an interactive machine-learning approach that enables users to provide iterative feedback for model adaptation. Specifically, we propose to adapt the explanatory interactive machine-learning approach CAIPI for fair machine learning. FAIRCAIPI incorporates human feedback in the loop on predictions and explanations to improve the fairness of the model. Experimental results demonstrate that FAIRCAIPI outperforms a state-of-the-art pre-processing bias mitigation strategy in terms of the fairness and the predictive performance of the resulting machine-learning model. We show that FAIRCAIPI can both uncover and reduce bias in machine-learning models and allows us to detect human bias.

Keywords: fair machine learning; explanatory and interactive machine learning

check for updates

Citation: Heidrich, L.; Slany, E.; Scheele, S.; Schmid, U. FAIRCAIPI: A Combination of Explanatory Interactive and Fair Machine Learning for Human and Machine Bias Reduction. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1519–1538. https://doi.org/10.3390/make5040076

Academic Editors: Federico Cabitza, Fang Chen, Jianlong Zhou and Andreas Holzinger

Received: 11 August 2023
Revised: 22 September 2023
Accepted: 14 October 2023
Published: 18 October 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The discovery of discriminatory machine-learning applications has refuted the popular belief that machine-learning (ML) algorithms make objective decisions. For instance, the *COMPAS* tool (short for Correctional Offender Management Profiling for Alternative Sanctions) erroneously assigns Black defendants a higher risk of recidivism than white defendants, indicating a clear racial bias [1]. Another example is Microsoft's *TayAI* chatbot, which generated racist and anti-Semitic tweets [2]. Despite this, machine-learning algorithms are increasingly used in sensitive domains such as employment hiring assistance [3] and credit request evaluation [4]. We argue that ML engineers have an ethical obligation to ensure that ML algorithms do not reproduce data-inherent biases and thus systematically disadvantage certain groups, despite legal requirements that may exist.

Numerous approaches incorporate fairness—or bias mitigation—as an additional objective, e.g., by satisfying fairness metrics during model optimization [5–7] or entirely new classification systems [8,9]. Celis et al. (2019) derive a generalized classification algorithm that arranges multiple arbitrary fairness metrics in a linear group performance function as an optimization objective. Satisfying fairness constraints *during* the model optimization is obviously an entirely different approach to bias mitigation than improving fairness *before* training. For instance, Reweighing modifies the proportion of potentially deprived groups before fitting a classification model such that it satisfies a single specific fairness metric [10]. Methods like Reweighing are likely to outperform more generic approaches regarding specific metrics, whereas the class of the former methods offers advantages when aggregating multiple fairness constraints.

Common to state-of-the-art bias-mitigation approaches is that they tend to treat fairness as a context-free, stationary concept. In general, what is considered to be fair is determined by the specific cultural background, be it a nation or a company, which highlights the highly subjective perception of fairness [11]. Berman et al. (1985) investigate the following example where colleagues of an Indian and a US company are likewise asked to distribute corporate benefits among each other. Although for the Indian employees, it was fair to distribute the corporate benefits according to the economic need of each individual, the US employees agreed to a distribution proportional to individual merit. Given the observation that fairness is highly context-sensitive, we argue that the current approaches of stationary fairness metrics need to be extended to interactive ML methods to take user-specific cultural presuppositions of fairness into account. To reflect individually perceived fairness, users must be able to guide the optimization process of ML models. In active learning [12], users iteratively label instances that maximize the information gain of the classification model. Coactive learning is an active learning modification where users interactively correct predictions of the classifier [13].

Active and coactive learning presumes that users capture the decision-making mechanism of problems modeled by ML, which we argue to be an overstating assumption. Given context-sensitive fairness interpretations, users have individual perceptions of fair decision-making. Explanatory and interactive ML (XIML) enriches the interactive component of active learning with explanatory ML techniques and thus equips the user with the awareness of *how* a ML model obtains its decision [14]. Hence, XIML discloses the model's decision-making mechanism by explanations and allows the user to correct both the prediction and the explanation. Angerschmid et al. (2022) [15] show that including an explanatory component in the decision-making process improves the user's perception of fairness. Indeed, the objective of CAIPI [14], a state-of-the-art XIML algorithm, is to leverage user feedback that drives the model toward a presumably correct decision-making mechanism from the perspective of a domain expert. For this purpose, CAIPI has been enriched with a user interface in the context of medical image classification [16]. Others propose algorithmic adaptations. For instance, Schramowski et al. (2020) [17] specifically tailor CAIPI for deep learning.

Let us introduce a fictional example within the domain of credit risk management to motivate our interactive learning approach. Suppose a student is applying for credit to build or buy real estate, as shown in Figure 1. The primary objectives of credit risk management include evaluating creditworthiness of the applicant, therefore minimizing risks, and maintaining a robust loan portfolio. Initially, suppose that the applicant is a philosophy student in his first semesters, aspiring to pursue research in academia in the future, and whose current income is limited to part-time work. This application is automatically rejected due to the applicant's current age, income situation, and employment status. In this case, the machine-learning system's assessment was accurate and fair with respect to the contextual situation of the student and the perspective of the lending organization, i.e., the explanation is not biased, and no additional interaction is needed. Second, envision a scenario where the applicant has a time contract in academia in the domain of computer science. However, the model, solely considering the weak employment status of a temporary contract, might initially reject the credit application. Here, the domain expert intervenes because he can reasonably anticipate a strong job market in the computing domain and, therefore, proceeds to rectify the label accordingly. Subsequently, he augments the dataset with this corrected information and retrains the model for improved accuracy. Third, suppose that the applicant is a female MSc student in biology. The system rejects the application due to gender female. The domain expert accepts the system recommendation, but he questions the explanation that the decision was based on gender. His reasoning leads to the observation that the explanation is biased; specifically, female applicants are considered less creditworthy. Despite this, he concludes that the decisive features of the credit rejection should be based on the fact that the applicant is not employed and does not provide enough equity capital to fulfill the banks' lending policy. Consequently, the

expert intervenes, rectifying this explanation bias by incorporating counterexamples into the dataset for model refinement.

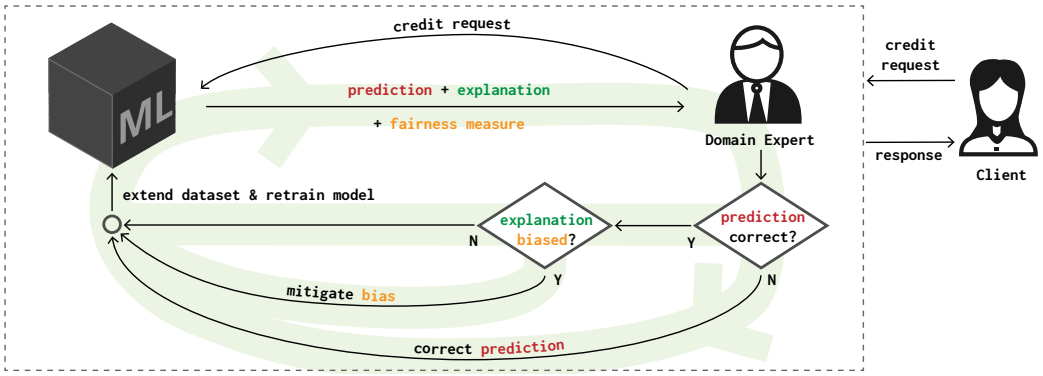

Figure 1. Interactive learning in the domain of lending.

We present FAIRCAIPI, a XIML approach based on the CAIPI algorithm. Its **novelty** is that FAIRCAIPI lets the user interact with the model's decision-making mechanism from a fairness perspective. Instead of embedding presumably correct mechanisms, users are asked to provide iterative feedback to improve the fairness of the model. Biased decision-making mechanisms are mitigated by user feedback. By FAIRCAIPI, we **contribute** a tool to (i) uncover and (ii) reduce machine bias as well as (iii) detect human bias *during* model optimization. This work presents the theoretical basis and a formal derivation of the FAIRCAIPI method. We assess our approach through a simulation study and additionally compare FAIRCAIPI to Reweighing, a state-of-the-art bias-mitigation pre-processing strategy [10]. The research questions addressed by our experiments are as follows:

(R1) Does the correction of explanations for fairness lead to fairer models?
(R2) Does correcting explanations for fairness lead to fairer explanations?
(R3) Does correcting for fair explanations have a negative impact on the predictive performance of the model?
(R4) Which is superior, FAIRCAIPI or the state-of-the-art Reweighing strategy?

2. Approaches to Fairness in Machine Learning

Bias-mitigation strategies can be classified into three stages: pre-processing, in-processing, and post-processing [18]. For an intuitive overview, we summarize and cluster related approaches to fairness in ML in Table 1. We briefly address them in this section to locate FAIRCAIPI in the research field fair ML:

Pre-processing is responsible for satisfying fairness quality criteria during the data collection process [19]. Prominent examples exist for facial recognition [20]. More general pre-processing approaches are based on sampling. Their objective is to modify the weight of certain instances to increase fairness. For example, Reweighing optimizes the proportion of deprived and favored groups in the dataset [10]. Other examples include Disparate Impact Removal [21] or Optimized Pre-Processing [22].

Post-processing, on the contrary, changes the output of the ML model by verifying fairness constraints. Some methods tune the probability threshold to minimize outcome differences between deprived and favored groups, e.g., using simpler linear models [23] or classification models [24]. Another example is Reject Option Classification, which alternates labels based on critical regions of the decision boundaries [25].

In-processing methods are commonly associated with FAIRCAIPI because fairness objectives are incorporated during model training by satisfying fairness metrics [5–8]. More sophisticated approaches include specific cost functions for different instances in their classification objective [26]. The latter approach has also been extended towards regressions [27]. In-processing bias-mitigation strategies also exist for more specific ML niches like discrimination-free word embedding for Natural Language Processing [28,29], fair

generative models, such as Generative Adversarial Networks [30], or Variational Autoencoders [31], discrimination-free image recognition with deep learning [32,33], or fair causal models and graphs such as Bayesian networks [34–38]. Some in-processing methods exploit explanatory ML methods, such as counterfactuals [39], causal explanations [40], or Shapley values [41] to reveal biased decision-making. In particular, the latter is closely associated with our method, as FAIRCAIPI accesses the model's mechanism by local explanations with Shapley values. Interaction with fairness in ML models is currently mostly reserved for visualization techniques that reveal biases within models in user interfaces [42,43]. A closely related approach to FAIRCAIPI is the XIML method by Nakao [44], which integrates human feedback on explanations into model optimization to increase fairness. By borrowing a mechanism for user explanation interaction [45], it allows users to change the disadvantageous behavior of a classifier for several protected attributes at once. By contrast, FAIRCAIPI aligns the user explanation interaction in an optimization cycle.

Table 1. High-level clustering of existing bias-mitigation strategies into pre-, in-, and post-processing [18]. FAIRCAIPI can be subsumed into the *XIML for bias mitigation* category.

Pre-Processing	In-Processing	Post-Processing
Fair data collection [19,20]	Fairness constraints [5–8]	Threshold tuning [23,24]
Sampling-based bias mitigation [10,21,22]	Fair cost functions [26,27]	Prediction alternation [25]
	Fair Natural Language Processing [28,29]	
	Fair generative models [30,31]	
	Discrimination-free image classification [32,33]	
	Fair causal models [34–38]	
	Explanatory bias mitigation [39–41]	
	Interactive bias mitigation [42,43]	
	XIML for bias mitigation [44]	

3. Materials and Methods

FAIRCAIPI combines the technical concepts of bias mitigation and XIML. Consequently, this section covers both. It begins with a brief description of the German Credit Risk dataset, which serves as a running example for the remainder of this paper and is the subject of the simulation study that underpins FAIRCAIPI. German Credit Risk suffers from a gender bias. Since FAIRCAIPI is an extension of the CAIPI algorithm, the derivation of XIML will mostly be centered around CAIPI. We will adapt CAIPI to satisfy fairness objectives and present a human ML architecture that can (i) detect and (ii) reduce machine bias, and (iii) detects human bias. We will specify the experimental setup at the end of this section. First, let us specify the basic notation used in this article:

Notation. *A binary classification model $f : \mathcal{X}^n \to \mathcal{Y}$ is a function that maps a feature space $\mathcal{X}^n$ of n features to a target space $\mathcal{Y} = \{0, 1\}$. For brevity, we omit the superscript n in the following and just write $\mathcal{X}$. We denote an inference by $y = f(x)$. An instance $x \in \mathcal{X}$ can be represented as feature value vector $x = (x_1, x_2, \ldots, x_n)^T \in \mathcal{X}$, where x_i denotes a single feature in x at index i. Furthermore, let $l : \mathcal{X} \to \mathcal{Y}$ be a labeling function from instances to class labels. Moreover, let $\mathcal{L} \subseteq \mathcal{X} \times \mathcal{Y}$ and $\mathcal{U} \subseteq \mathcal{X}$ denote subsets of labeled and unlabeled instances, where we write $\mathcal{X}_\mathcal{L}$ and $\mathcal{Y}_\mathcal{L}$ for instance data and labels of $\mathcal{L}$, respectively. Furthermore, we will write $x^{(n)}$ ($y^{(n)}$) for the n-th feature (label) instance in $\mathcal{X}$ ($\mathcal{Y}$) when the associated set is clear from the context, or add a subscript like $x^{(n)}_\mathcal{L}$ ($y^{(n)}_\mathcal{L}$) and $x^{(n)}_\mathcal{U}$ to indicate its associated set explicitly. We assume a procedure* FIT *to train and update a classification model.*

3.1. Bias in the German Credit Risk Dataset

The German Credit Risk data set (https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data, accessed on 17 July 2023) suffers from a gender bias, as shown in Figure 2. In the context of fairness, bias systematically favors a privileged group while penalizing an unprivileged group [18]. Privileged and unprivileged groups are determined by the distribution of a *favorable label* [18] regarding a *protected attribute* [46,47]. Discrimination occurs when an unprivileged group is systematically disadvantaged because of a protected characteristic. Notice that discrimination can occur directly or indirectly. *Direct discrimination* is based on a protected attribute, while *indirect discrimination* is caused by apparently neutral features that are highly correlated with the protected attribute [37]. To quantify bias, we split the data into four groups [10]:

DP **D**eprived (unprivileged) group with **P**ositive (favorable) label
DN **D**eprived (unprivileged) group with **N**egative (unfavorable) label
FP **F**avored (privileged) group with **P**ositive (favorable) label
FN **F**avored (privileged) group with **N**egative (unfavorable) label

Consider Figure 2 that displays the distribution of the favorable label *good credit risk* conditioned on the protected attribute *gender*: This reveals a gender bias, as the favored group *male* receives the positive label more likely than the deprived group *female*.

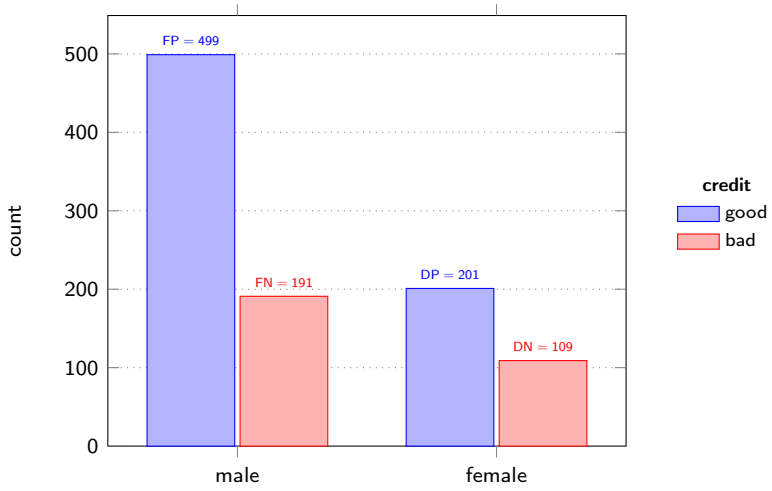

Figure 2. Visualization **F**avored group with **P**ositive label (**FP**), **F**avored group with **N**egative label (**FN**), **D**eprived group with **P**ositive label (**DP**), and **D**eprived group with **N**egative label (**DN**). This figure investigates the favorable label *good credit risk* regarding the protected feature *gender* and reveals a gender bias, as it is more likely to have a good credit risk for males than for females.

By now, we have presented bias mostly narratively, raising the question: How can it be quantified? Metrics measuring bias are called *bias-detection metrics*. Most bias-detection metrics measure conditional classification model performance differences. For instance, a fair classifier requires that the recall of the model— in our case, the correctly detected good credit risks— is stable with respect to gender as a protected attribute.

Definition 1 (Protected Attribute)**.** *Let S be the protected attribute and let $S = s$ and $S = \bar{s}$ mark the privileged and unprivileged groups, respectively.*

Definition 2 (Favorable Label)**.** *Let $\hat{y} = f(x)$ be a prediction, where $\hat{y} = d$ denotes the favorable label and $\hat{y} = \bar{d}$ the unfavorable outcome label, respectively.*

In the following, we present five bias-detection metrics: Statistical Parity [48], Equalized Odds [23], Equalized Opportunity [23], False Positive Error Rate Balance [49], and Predictive Parity [49]. According to the definition of Statistical Parity (SP) [48], a classifier

is fair if the probability of receiving the unfavorable label is distributed equally across privileged and unprivileged groups, i.e.,

$$P(\hat{y} = \bar{d}|S = s) \stackrel{!}{=} P(\hat{y} = \bar{d}|S = \bar{s}), \tag{1}$$

where P is the probability of the predictive outcome $\hat{y}$ conditioned on the protected attribute S. When we now condition the probability of receiving the unfavorable label, i.e., $P(\hat{y} = \bar{d})$, on the privileged and unprivileged groups $S = s$ and $S = \bar{s}$, we obtain $P(\hat{y} = \bar{d}|S = s)$ and $P(\hat{y} = \bar{d}|S = \bar{s})$, respectively. From the definition in Equation (1) directly follows the assumption that Statistical Parity exists if the conditional difference of receiving the unfavorable label is zero between the privileged and unprivileged groups. The differential form of (1) is given in Equation (2), which states that a fair classifier should yield a Statistical Parity estimate $SP = 0$, i.e.,

$$SP = P(\hat{y} = \bar{d}|S = s) - P(\hat{y} = \bar{d}|S = \bar{s}). \tag{2}$$

The main difference between Statistical Parity compared to the remaining bias-detection metrics is that it does not require access to the ground truth, as it solely relies on the prediction conditioned on a known protected attribute. In contrast, Equalized Odds (EqOdds) [23] includes the ground truth label y in the condition of each side:

$$P(\hat{y} = \bar{d}|y = \bar{d}, S = s) \stackrel{!}{=} P(\hat{y} = \bar{d}|y = \bar{d}, S = \bar{s}). \tag{3}$$

Equation (4) formulates EqOdds as an average of two differences. Each part of the sum is a performance difference between the privileged and the unprivileged group. The conditional probability is now replaced by performance metrics of a binary classification. Equalized Odds likewise considers the false positive (fpr) and the true positive rate (tpr):

$$EqOdds = \frac{1}{2} \times [(fpr_{S=\bar{s}} - fpr_{S=s}) + (tpr_{S=\bar{s}} - tpr_{S=s})]. \tag{4}$$

In fact, the idea of (3) forms the basis of Equal Opportunity (EqOpp) [23], False Positive Error Rate Balance (FPERB) [49], and Predictive Parity (PP) [49], where EqOpp uses the tpr in (5) and FPERB the fpr in (6). Predictive Parity is slightly different as it exploits the false discovery rate (fdr) in (7), calculated by $fp \times (fp + tp)^{-1}$.

$$EqOpp = tpr_{S=\bar{s}} - tpr_{S=s} \tag{5}$$

$$FPERB = fpr_{S=\bar{s}} - fpr_{S=s} \tag{6}$$

$$PP = fdr_{S=\bar{s}} - fdr_{S=s} \tag{7}$$

Our explanatory and interactive FairCaipi cycle will present the bias-detection metrics, as defined above, to the user. Its purpose is to notify and educate users about the impact of their changes have on the fairness of the classifier. Furthermore, Statistical Parity (2) will play an important role in the benchmark test in the simulation study as Reweighing is optimized for Statistical Parity.

3.2. Explanatory Interactive Machine Learning

The state-of-the-art XIML algorithm Caipi [14] involves users by iteratively including prediction and explanation corrections. Caipi has three prediction outcome states: **R**ight for the **R**ight **R**easons (**RRR**), **R**ight for the **W**rong **R**easons (**RWR**), or **W**rong for the **W**rong **R**easons (**W**). The two erroneous cases require human intervention. Although users correct the label in the **W** case, they give feedback for wrong explanations in the **RWR** case, where so-called counterexamples are generated. Counterexamples are additional novel instances, containing solely the decisive features. They are supposed to shape the model's mechanism into a presumably correct direction from a user's perspective. Practically, using suitable

data augmentation procedures, a single explanation correction yields multiple different counterexamples. For the remainder of this work, let us assume that a procedure GEN generates counterexamples using a user-induced explanation correction.

Definition 3 (Counterexample Generation)**.** *Consider the prediction $\hat{y} = f(x)$ and suppose that, according to a user, the vector x^* is decisive for $\hat{y}$, in the sense that the attribution effect of each non-zero feature in x^* exceeds a threshold value, where vector x^* is derived from x such that the set of non-zero components of x^* is a subset of the set of non-zero components in x. Let us assume a procedure* GEN *that takes x, $\hat{y}$, x^*, c as inputs and returns counterexample feature and target data sets $\mathcal{X}'$ and $\mathcal{Y}'$. For $\mathcal{Y}'$, we repeat $\hat{y}$ for c times. And for $\mathcal{X}'$, x is repeated c times. Whenever x_i is not set in x^*, x_i is disturbed, e.g., by randomization.*

We use *SHapley Additive exPlanations (SHAP)* [50] to obtain local explanations of a classification model, whereas traditional CAIPI uses LIME. SHAP tends to be a fruitful option at this point, as LIME's performance is sensitive to segmentation algorithms [51]. SHAP, in contrast, performs reliably on tabular data [50]. The SHAP explanation model approximates a model f for a specific input x by an explanation g that uses a simplified input x' that maps to the original input x via a mapping $h_x(x') = x$. It ensures that $g(z') \approx f(h_x(z'))$ whenever $z' \approx x'$ and where $z' \in \{0,1\}^M$ and M is the number of simplified input features. The SHAP method relies on Shapley values, which measure the importance of a feature x_i for a prediction by calculating the impact of knowledge about this feature for the prediction. The contribution of a feature value x_i to a prediction outcome is known as the SHAP value $\phi_i \in \mathbb{R}$, such that the sum of all feature attributions approximates the output $f(x)$ of the original model.

The SHAP method is built upon three desirable properties: *local accuracy*, *missingness*, and *consistency* [50]. SHAP's simplified representation of a classification model $f(x)$ including a simplified feature space representation x', i.e.,

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i, \text{ where } x = h_x(x') \tag{8}$$

directly satisfies the *local accuracy property*, whenever $x = h_x(x')$ holds [50]. SHAP approximates a model f for an input x in the sense that attributions are *added* such that the explanation model $g(x')$ for simplified input x' should at least match the output of the original model $f(x)$: In the vanilla case, attributions ϕ are added linearly, where each ϕ_i represents the importance of a feature (or a combination of a feature subset) with size M. The baseline attribution ϕ_0 is calculated by $\phi_0 = f(h_x(\mathbf{0}))$.

Apart from local accuracy, SHAP satisfies the missingness and consistency properties [50]. *Missingness* states that zero (in this case missing) feature values have an attribution value of zero. However, *consistency* guarantees that a stronger feature importance for f is also represented by the attribution value ϕ. Ensuring all three properties, SHAP attributions are given as:

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} \left[f_x(z') - f_x(z' \setminus i) \right], \tag{9}$$

where $z' \subseteq x'$ represents all z' vectors where the non-zero entries are a subset of the non-zero entries in x', $|z'|$ is the number of non-zero entries in z', $f_x(z') = f(h(_x(z')))$ and $z' \setminus i$ denotes setting $z'_i = 0$. The right-hand side of Equation (9) reflects the original idea behind Shapley values, as it is the difference for f_x between including versus excluding the i-th feature of z'. Equation (9) is the single possible solution that follows from the properties: local accuracy, missingness, and consistency [50]. Young (1985) [52] originally proves that Shapley values are the only possible representations that satisfy local accuracy and missingness. Young (1985) [52] utilizes an additional axiom, which Lundberg and Lee (2017) prove to be non-essential. According to Lundberg and Lee (2017), the missingness property

is non-essential for Shapley values themselves. However, it is essential for additive feature attribution methods such as SHAP.

Consider Figure 3 as an exemplary SHAP explanation for an arbitrary instance and a classification model with a prediction score of 0.5, where values lower than 0.5 indicate a good credit risk and values higher or equal to 0.5, a bad credit risk. SHAP's base value lies approximately at 0.27. From this point, some attributes have a positive impact, i.e., a positive attribution value—they drag the decision toward bad credit risk. Here, having a bank account's `status` that is not None (it exists), but the average incoming monthly payment within the last year is smaller than 200 German Mark, being female ($\texttt{sex} = 0$) and not been employed for over 4 years ($\texttt{employment} = 4 + \texttt{years} = 0$) are the major reasons for bad credit risk. On the contrary, the investigated instance requested a `credit_amount` of 2278 German Mark, which is the only feature that receives a negative attribution value and thus contributes to a good credit risk.

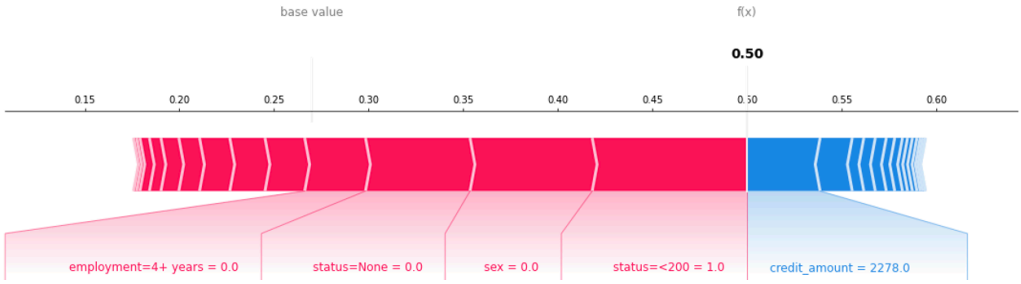

Figure 3. SHAP explanation for a test instance of a classification model trained on the German Credit Risk dataset. The classification score matches the decision boundary of 0.5. SHAP estimates attribution values for each feature. It mimics the classification model's outcome and builds a sum of each feature value, which is scaled by its attribution. The algebraic sign of the attribution value determines whether the feature contributes to a positive or negative classification result. SHAP starts from a base value—SHAP's attribution value for **0**. The base value is approximately 0.27. From this point, features in red contribute to a bad credit risk, whereas features in blue drag the classification score toward a good credit risk. (Figure generated with https://shap-lrjball.readthedocs.io/en/latest/generated/shap.force_plot.html, accessed on 19 July 2023).

CAIPI has a local explanation procedure that takes a feature instance and a classification model as input and reveals the decisive features to the user. Mathematically, our procedure Exp is built upon SHAP.

Definition 4 (Local Explanation)**.** *Let ϕ be attribution values assigned to x, highlighting the importance of each feature x_i in x for $f(x)$ (9). Furthermore, let α be an importance threshold. Let $e \subseteq x$ denote the set of features such that $|\phi_i| > \alpha$ holds. We assume an explanation procedure* Exp *that takes x, f, and α as input and returns e.*

CAIPI [14] leverages user feedback regarding the model's prediction and explanation depending on the prediction outcome state. In each iteration, it selects the most informative instance from an unlabeled dataset—that is, the instance with prediction score closets to the decision boundary (in our case 0.5) regarding the classifier trained on a smaller pool of labeled data. We argue that this instance is *most informative*, as we associate predictions close to the decision boundary with high uncertainty. Knowledge about its label maximizes the information gain for the classifier in the next iteration. The procedure MII retrieves the index of the most informative instance. At this point, we assume access to both the prediction scores and the decision boundary.

Definition 5 (Most Informative Instance)**.** *Let the procedure* MII *take a set of predictions $\hat{y}$ and a decision boundary β as input. It returns the index m of the* most informative *instance, i.e., the prediction with the score closest to β.*

CAIPI requires human feedback at two points: to evaluate the correctness of the prediction and the explanation. In the second case, by correcting the local explanation, users can induce a desirable decision mechanism. We denote the interaction points by the procedure INTERACT and summarize CAIPI in Algorithm 1. In each iteration, CAIPI trains a model on the labeled data (line 2) and draws predictions on the unlabeled feature instances (line 3) to obtain the most informative instance (line 4). The user examines the prediction and provides the correct label if the prediction is incorrect (line 7). Otherwise, if the prediction is correct, CAIPI presents the corresponding local explanation, which can be corrected by the user if necessary (line 9). If the explanation is correct, the instance is added to the labeled dataset (line 12), otherwise, counterexamples are generated (line 14). The current most informative instance is removed from the set of unlabeled data to prepare the next iteration (line 15). In contrast to the original CAIPI algorithm [14], we formalize each component explicitly. We will utilize our explicit formalization to adapt CAIPI to fair ML in the next section.

Algorithm 1 CAIPI $[\mathcal{L}, \mathcal{U}, c, n]$ [14]

Input: Labeled dataset $\mathcal{L}$, unlabeled dataset $\mathcal{U}$, number of counterexamples c, number of iterations n
Output: Classification model f

1: **for** $i \leftarrow 1 : n$ **do**
2: $\quad f \leftarrow \text{FIT}(\mathcal{L})$
3: $\quad \hat{\mathcal{Y}} \leftarrow \{f(u) \mid u \in \mathcal{U}\}$
4: $\quad m \leftarrow \text{MII}(\hat{\mathcal{Y}}, 0.5)$
5: $\quad y^{(m)} \leftarrow \text{INTERACT}(x_{\mathcal{U}}^{(m)})$ ▷ *Label retrieved from human annotator*
6: $\quad$ **if** $\hat{y}^{(m)} \neq y^{(m)}$ **then**
7: $\quad\quad \mathcal{L} \leftarrow \mathcal{L} \cup \{(x_{\mathcal{U}}^{(m)}, y^{(m)})\}$ ▷ *Case **W**: label correction*
8: $\quad$ **else**
9: $\quad\quad e \leftarrow \text{EXP}(x_{\mathcal{U}}^{(m)}, f, 0.005)$
10: $\quad\quad x^* \leftarrow \text{INTERACT}(e)$ ▷ *Decisive features retrieved from human annotator*
11: $\quad\quad$ **if** $e = x^*$ **then**
12: $\quad\quad\quad \mathcal{L} \leftarrow \mathcal{L} \cup \{(x_{\mathcal{U}}^{(m)}, \hat{y}^{(m)})\}$ ▷ *Case **RRR**: no further interaction needed*
13: $\quad\quad$ **else**
14: $\quad\quad\quad \mathcal{L} \leftarrow \mathcal{L} \cup \text{GEN}(x_{\mathcal{U}}^{(m)}, \hat{y}^{(m)}, x^*, c) \cup \{(x_{\mathcal{U}}^{(m)}, \hat{y}^{(m)})\}$
▷ *Case **RWR**: Explanation correction and generation of counterexamples*
15: $\quad \mathcal{U} \leftarrow \mathcal{U} \setminus x_{\mathcal{U}}^{(m)}$
16: **return** f

3.3. Fair Explanatory and Interactive Machine Learning

FAIRCAIPI adapts the original CAIPI framework with a fairness objective in two ways: (i) it evaluates the local explanation to detect biased decision-making, and (ii) it thus accounts for protected attributes during the counterexample generation. Regarding adaptation (i), we recapture the groups DP, DN, FP, and FN from Section 3.1, where D indicates the deprived and F the favored group, each with either the desirable positive or undesirable negative outcome, P or N, respectively. We argue that the over-proportional presence of DN and FP manifests a bias, as the first assigns the undesirable outcome to the deprived and the second the desirable outcome to the favored group. Consequently, we define an explanation—a decision-making mechanism—as unfair if the fact of belonging to the deprived group is a reason for receiving the undesirable label. Conversely, belonging to the favored group is a reason to receive the desirable label. Regarding adaptation (ii), our goal is to remove protected attributes from the decision-making mechanism. This is achieved if the protected attributes are randomized, and all remaining features are held constant during counterexample generation. Randomization, in our case, means that if the fact of being male is a reason for a good credit risk, our counterexample is the identical instance, but the gender is female. Let us formalize the notion of *biased decision making*:

Definition 6 (Biased Decision Making). *Consider features e_i from explanation e, written $e_i \in e$, for a model's outcome $\hat{y} = f(x) \in \{d, \bar{d}\}$, and a protected attribute S with deprived group $\bar{s}$ and favored group s. We define the* decision-making *mechanism of f to be* biased *if it holds that*

$$\hat{y} = \bar{d} \text{ and } \exists e_i \in e, e_{i=S} = \bar{s}, \quad \text{or} \quad \hat{y} = d \text{ and } \exists e_i \in e, e_{i=S} = s.$$

FairCaipi's bias-mitigation strategy takes place in the counterexample generation procedure Gen′ (Algorithm 2), where we identify the parameterization of the protected attribute that would reproduce a bias regarding the prediction (line 2). For example, if the prediction would be a bad credit risk, then the bias-reproducing parameterization of the protected attribute gender would be female. Next, we build a set of all possible values of the protected attribute without the bias-reproducing value (line 3). To generate a counterexample instance, we repeat each feature but randomly replace the protected attribute (line 5), and add the input label (line 6). The resulting counterexample dataset contains all initial correlations except for the correlation between the protected attribute and the target.

Algorithm 2 Gen′ $[x, y, S, c]$ (Counterexample Generation)

Input: feature instance x, label y, protected attribute S, number of counterexamples c
Output: Data sets of labeled counterexamples $\mathcal{X}', \mathcal{Y}'$

1: $\mathcal{X}' \leftarrow \emptyset; \mathcal{Y}' \leftarrow \emptyset$
2: $s^* \leftarrow \bar{s}$ **if** $y = \bar{d}$ **else** $s^* \leftarrow s$
3: $s' \leftarrow \{x_{i=S} | x_i \in \mathcal{X}\} \setminus s^*$
4: **for** $n \leftarrow 1 : c$ **do**
5: $\quad x' \leftarrow \text{sample}(s')$ **if** $x_{i=S} = s^*$ **else** x_i **for** $x_i \in x$
6: $\quad \mathcal{X}' \leftarrow \mathcal{X}' \cup \{x'\}; \mathcal{Y}' \leftarrow \mathcal{Y}' \cup \{y\}$
7: **return** $\mathcal{X}', \mathcal{Y}'$

In contrast to original Caipi, FairCaipi (Algorithm 3) takes the protected attribute as an input. The users still provide feedback on the model's explanation. However, their task is no longer to induce correct mechanisms but to mitigate bias. Thus, according to Definition 6, the interaction in line 10 returns `True` if the explanation is biased and `False` otherwise, where an identified bias yields bias mitigating counterexamples (line 14).

FairCaipi is capable of (i) detecting and (ii) reducing machine bias and (iii) uncovering human bias. Note, however, that technically, so far FairCaipi does not require human feedback as long as we have access to all elements of Algorithm 3, which often is a realistic assumption when experiments are only pseudo-unlabeled. Although this qualifies FairCaipi to meet the first two capabilities, the latter is still not met. Practically, we explicitly want to include a human user in the optimization cycle. The user can override FairCaipi's biased decision (Definition 6). This might be due to good human intentions—passive bias, when the users fail to correct a biased mechanism—or active bias, when users intentionally miss corrections or correct the mechanism such that the bias-detection metrics suffer. We enrich FairCaipi to present bias detection metrics (Section 3.1) to the user at the beginning and the end of each iteration, i.e., before and after refitting the model (line 2). Furthermore, we include an interaction step where the user has the opportunity to provide feedback, and FairCaipi notifies the user in the case conducted or missed corrections negatively affect the bias-detection metrics. Placing FairCaipi in the priory-described design (Figure 4) has an essential benefit: It educates users, as it relates human feedback directly to bias-detection metrics.

Algorithm 3 FAIRCAIPI $[\mathcal{L}, \mathcal{U}, S, c, n]$

Input: Labeled dataset $\mathcal{L}$, unlabeled dataset $\mathcal{U}$, protected attribute S, number of counterexamples c, number of iterations n
Output: Classification model f

for $i \leftarrow 1 : n$ **do**
  $f \leftarrow \text{FIT}(\mathcal{L})$ **and** *ComputeFairnessMetrics*$(f, S, \mathcal{L})$
    ▷ *Compare fairness before and after refitting the model and present to the user*
  $\hat{\mathcal{Y}} \leftarrow \{f(u) \mid u \in \mathcal{U}\}$
  $m \leftarrow \text{MII}(\hat{\mathcal{Y}}, 0.5)$
  $y^{(m)} \leftarrow \text{INTERACT}(x_{\mathcal{U}}^{(m)})$ ▷ *Label retrieved from human annotator*
  if $\hat{y}^{(m)} \neq y^{(m)}$ **then**
    $\mathcal{L} \leftarrow \mathcal{L} \cup \{(x_{\mathcal{U}}^{(m)}, y^{(m)})\}$ ▷ *Case* ***W***
  else
    $e \leftarrow \text{EXP}(x_{\mathcal{U}}^{(m)}, f, 0.005)$
    $b \leftarrow \text{INTERACT}(e)$ ▷ *Bias information retrieved from human annotator*
    if not b **then**
      $\mathcal{L} \leftarrow \mathcal{L} \cup \{(x_{\mathcal{U}}^{(m)}, \hat{y}^{(m)})\}$ ▷ *Case* ***RRR***
    else
      $\mathcal{L} \leftarrow \mathcal{L} \cup \text{GEN}'(x_{\mathcal{U}}^{(m)}, \hat{y}^{(m)}, S, c) \cup \{(x_{\mathcal{U}}^{(m)}, \hat{y}^{(m)})\}$ ▷ *Case* ***RWR***
  $\mathcal{U} \leftarrow \mathcal{U} \setminus x_{\mathcal{U}}^{(m)}$
return f

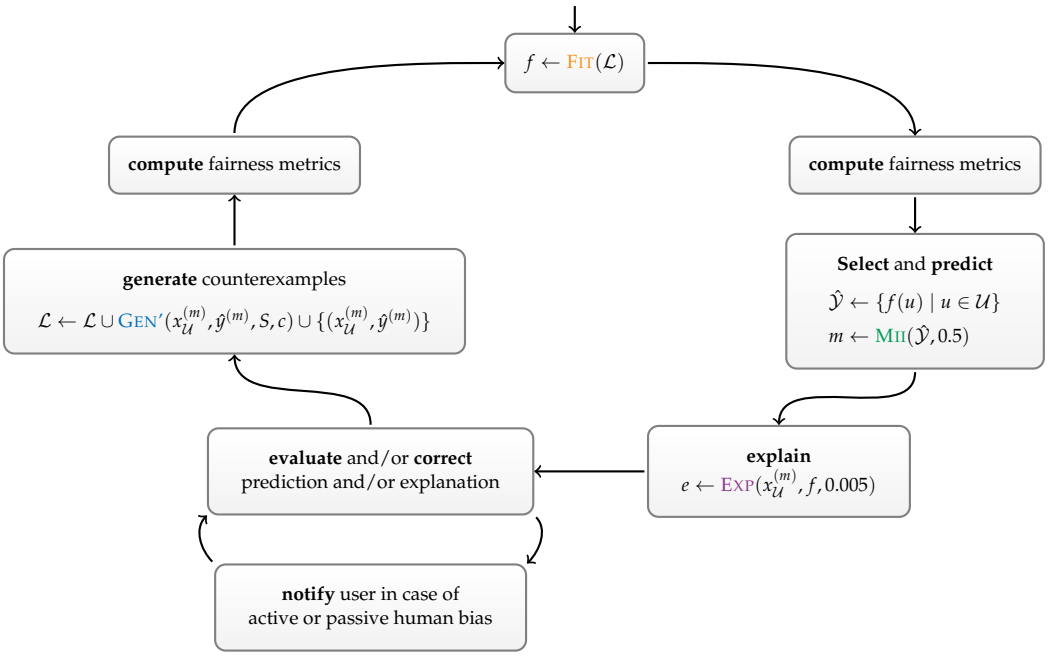

Figure 4. Human FAIRCAIPI interaction. FAIRCAIPI seeks user feedback regarding the prediction and explanation of a classification model. Placing bias-detection metrics at the beginning and the end of each iteration lets the users relate their correction to the model's bias. Furthermore, CAIPI notifies users if conducted or missed corrections negatively affect bias-detection metrics. FAIRCAIPI, operated by users that act out of good intentions, will make classification models fairer.

Our approach aims at improving the model quality while minimizing the number of queries, interactions, and overall cost. According to [14], active learning benefits from a significantly improved sample complexity compared to passive learning due to the specific selection of instances for labeling. Regarding its computational complexity, our approach is influenced by its model-agnostic nature, inheriting the complexities associated with fitting

a model and making predictions for specific instances based on the underlying machine-learning algorithm. The core components of our approach, including FAIRCAIPI and its associated computations, such as computing fairness metrics and SHAP values, maintain reasonable complexity: Notably, SHAP computation being part of the EXP procedure, although generally intractable, can be efficiently approximated in polynomial time [50,53]. Procedures MII and GEN′ for counterexample generation, as well as used set operations are at most of polynomial complexity. Hence, we argue that FAIRCAIPI is a computationally viable XIML approach.

3.4. Simulation Study

We demonstrate FAIRCAIPI in a simulation study, whose objective is to reduce gender bias in the German Credit risk dataset (code that reproduces the simulation study and code to execute FAIRCAIPI interactively according to Figure 4 is available under: https://github.com/emanuelsla/faircaipi, accessed on 11 August 2023). This makes the gender variable the only protected attribute S. Moreover, we have access to the labels of the actually unlabeled instances by $\mathcal{Y}_{\mathcal{U}} = l(\mathcal{U})$ and evaluate our explanations according to Definition 6 with an attribution threshold of 0.005. We benchmark FAIRCAIPI with Reweighing [10], which is a pre-processing technique and assigns weights to the training instances to achieve Statistical Parity (1). From the German Credit Risk dataset, we assume 550 instances to be labeled and 150 to be unlabeled. The remaining 300 instances serve as test data. We train a Random Forest classifier with balanced class weights. A priori, we achieve an accuracy of 75%. A comparatively high starting performance is important for FAIRCAIPI, as its objective is to make decision-making mechanisms fairer. Prior research shows that fewer a priori labeled instances suffice for a high predictive performance after CAIPI optimization [16]. We run FAIRCAIPI for 100 iterations. In RWR iterations, we add a single counterexample, the identical instance with the opposite gender—we neutralize the instance's gender effect.

4. Results

During 100 FAIRCAIPI iterations shown in Figure 5, we observe a slight trend toward more label corrections compared to explanation corrections. This implies more iterations of type W than of type RWR. Within FAIRCAIPI, approximately 35 out of 100 iterations correct only the label, i.e., solely the predictive accuracy. Yet, within 30 iterations, the SHAP explanation is corrected. These are the only iterations where the simulated user tries to adapt the model's mechanism to mitigate the classifier's inherent bias. However, in about 35 FAIRCAIPI iterations, the prediction is correct, including an unbiased decision-making mechanism representing RRR cases.

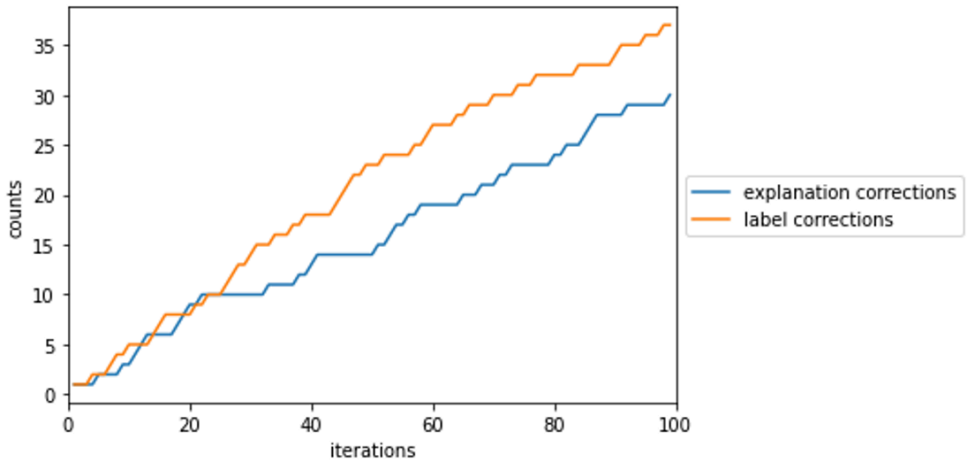

Figure 5. Number of label and explanation corrections within 100 FAIRCAIPI iterations. We observe a phase with predominantly label corrections around iteration 20 to 30. Before and afterward, both correction types are proportionally distributed across the FAIRCAIPI iterations.

The results for the bias-mitigation property of FAIRCAIPI are shown in Table 2. There, we compare several bias-detection metrics after 100 iterations of the FAIRCAIPI optimization, their optimal value during 100 iterations, the state-of-the-art sampling-based pre-processing procedure Reweighing, and the default Random Forest classifier without bias-mitigation extensions. All bias-detection metrics have their optimum at zero. We observe that FAIRCAIPI is superior for every bias-detection metric, except Statistical Parity. This holds for the FAIRCAIPI results after 100 iterations and is even amplified by taking its optima into account. Reweighing, which adds weights to instances to satisfy Statistical Parity, clearly outperforms the others regarding Statistical Parity but offers only minor improvements for other bias detection metrics compared to the default Random Forest classifier.

Table 2. Bias evaluation. We compare bias metrics for a Random Forest classifier trained on the German Credit Risk dataset. **Default** values result from a plain Random Forest model, **Reweighing** includes a Statistical parity-optimized sampling procedure prior to training. The **FAIRCAIPI** column references the end result (after 100 iterations), **FAIRCAIPI (opt.)** the optimum values.

Metric	Default	FAIRCAIPI	FAIRCAIPI (Opt.)	Reweighing
Statistical Parity	−0.0886	−0.0447	−0.0391	0.0000
Equalized Odds	−0.1568	−0.0909	−0.0038	−0.1819
Equalized Opportunity	−0.0514	−0.0295	−0.0007	−0.1237
FPERB	−0.2622	−0.1524	0.0038	−0.2401
Predictive Parity	−0.0322	−0.0026	0.0008	−0.0053

Figure 6 visualizes the development of the investigated bias-detection metrics over the course of 100 FAIRCAIPI iterations, all of which have their optimum at zero. Each metric tends to move towards its optimum and reaches it approximately in its best iteration. Most metrics have a smaller amplitude and are close to the optimum throughout the entire optimization cycle. Exceptions are Equalized Odds and False Positive Error Rate Balance. Although the former steadily converges toward its optimum, the latter has an even higher amplitude, and it tends to diverge from the optimum value again after iteration 80.

The bias reduction trend can also be observed in Figure 7, where the number of unfair predictions and explanations tend to decrease during FAIRCAIPI optimization. Figure 7 provides some interesting insights: Due to the size of the labeled and test data, unfair predictions and explanations occur more frequently in the labeled data than in the test data. Moreover, if the explanations are unfair—predictions are made for biased reasons—the prediction is also classified as unfair. This does not hold the other way around, as instances can be part of the deprived group and receive the unfavored label, but out of fair reasons—not caused by the protected attribute. Hence, unfair predictions occur more often than unfair explanations.

Bias mitigation should not negatively affect the predictive performance of the classifier. Therefore, we summarize several performance metrics for binary classification in Table 3, where we compare accuracy, precision, recall, and F1-score for FAIRCAIPI, a standard Random Forest classifier, and a Random Forest classification pre-processed with Reweighing. FAIRCAIPI is superior for each performance metric. However, we observe a large discrepancy in the performance metrics when they are conditioned on the target. Although all classifiers perform comparatively well for good credit risk, their performance suffers for bad credit risk. We underpin the prior comparison and visualize the test accuracy of each FAIRCAIPI iteration (Figure 8) but we stress the highly imbalanced setting (Figure 2) and, therefore, emphasize the results in Table 3. We only use the accuracy metric in Figure 8 as a proxy to visualize performance changes over the course of FAIRCAIPI optimization. We see minor changes, which is the desirable behavior, as FAIRCAIPI is designed to reduce the bias of an established decision-making mechanism rather than optimize the predictive quality. Let us finalize this section and answer our research questions:

(R1) Does the correction of explanations for fairness lead to fairer models?
Yes, according to Figure 6, the bias detection metrics converge to their optimum as explanation corrections are added. Furthermore, unfair predictions decrease over the course of 100 FairCaipi *iterations (Figure 7), where explanations are corrected in 30 iterations (Figure 5).*

(R2) Does correcting explanations for fairness lead to fairer explanations?
Yes, Figure 7 reveals that the number of unfair explanations decreases with FairCaipi.

(R3) Does correcting for fair explanations have a negative impact on the predictive performance of the model?
No, Figure 8 clearly illustrates solely minor performance changes during FairCaipi *optimization. Table 3 even indicates a slight increase in predictive quality with* FairCaipi *compared to Random Forest classification with and without Reweighing.*

(R4) Which is superior, FairCaipi or the state-of-the-art Reweighing strategy?
Considering Table 2, FairCaipi *is the superior bias-mitigation strategy for every metric except Statistical Parity, which is the optimization goal of Reweighing.*

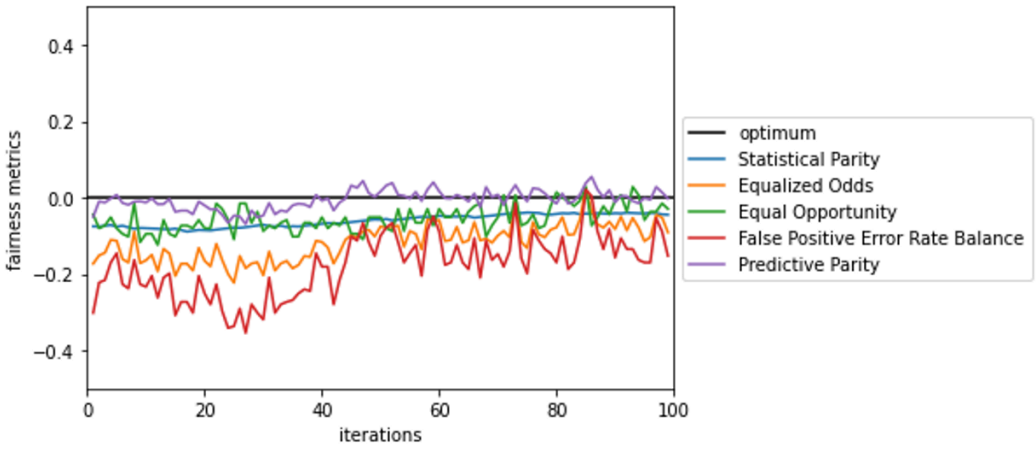

Figure 6. Bias-detection metrics during FairCaipi optimization. We compare the development of bias-detection metrics across 100 FairCaipi iterations. All metrics have their optimum at zero.

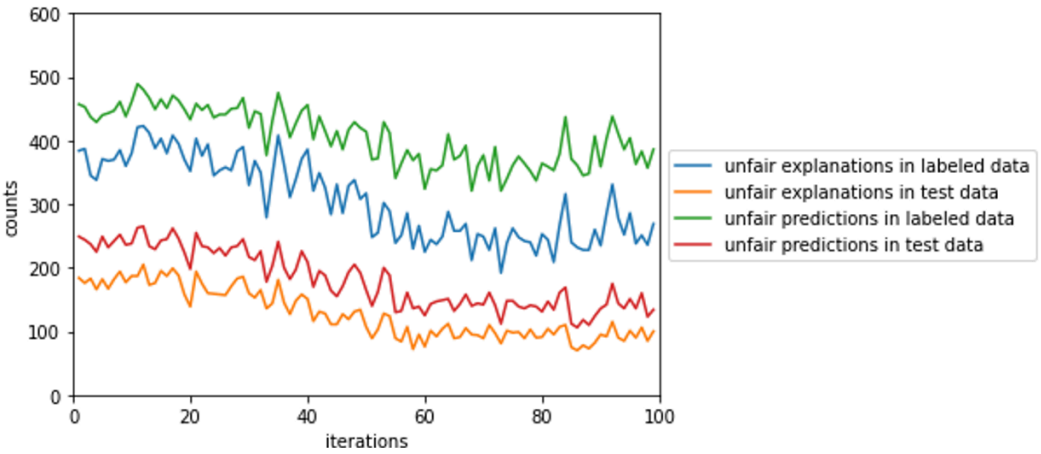

Figure 7. Unfair predictions and explanations during FairCaipi optimization. In each iteration, we calculate the number of unfair predictions and explanations on labeled and test data.

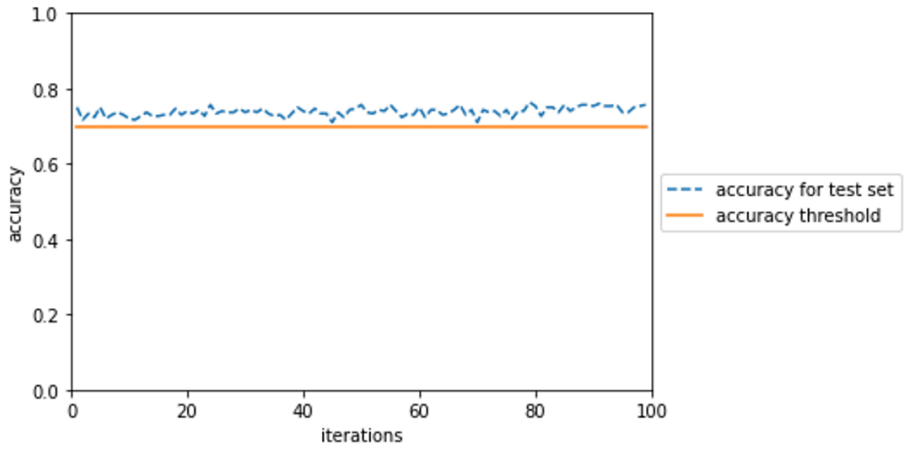

Figure 8. Predictive quality of FAIRCAIPI during 100 iterations. Starting with a baseline accuracy of 75%, we calculate the test accuracy in each iteration. We add a 70% accuracy threshold mark.

Table 3. Evaluation of predictive quality. We compare several performance metrics for a Random Forest classifier trained on the German Credit Risk dataset. The **Default** column references a plain Random Forest model without modifications, **Reweighing** includes a bias-mitigation pre-processing sampling strategy, and **FAIRCAIPI** contains the results after 100 iterations.

Metric	Subset	Default	FAIRCAIPI	Reweighing
Accuracy	-	0.73	0.76	0.72
Precision	good risk	0.76	0.77	0.75
	bad risk	0.60	0.68	0.58
Recall	good risk	0.89	0.92	0.89
	bad risk	0.38	0.39	0.34
F1-score	good risk	0.82	0.84	0.82
	bad risk	0.47	0.5	0.43

5. Discussion

Our experimental results show that FAIRCAIPI is a suitable bias-mitigation strategy. It satisfies two desirable properties: First, it does not negatively affect the predictive performance—on the contrary, according to our findings, it slightly increases predictive performance. Second, it mitigates bias successfully, taking several bias-detection metrics into account, and even outperforms a state-of-the-art bias-mitigation pre-processing procedure. Moreover, in the spirit of XIML, FAIRCAIPI ensures a transparent decision-making mechanism and is capable of directly involving humans in bias mitigation, which, despite legal requirements that might exist, is undoubtedly an ethical benefit. Humans do not treat fairness as a stationary concept [11]. According to our findings, the adaptation of the decision-making mechanism of a classification model by human feedback yields an overall well-suited bias mitigation, taking several bias-detection metrics into account. Optimizing for a stationary metric alone does not guarantee overall fairness improvements.

We argue that FAIRCAIPI is capable of (i) discovering and (ii) reducing machine bias and (iii) detecting human bias. Although the first two arguments are investigated experimentally, we propose a formal architecture for the third. This requires further investigation. User studies are applicable here. In this context, we could ask the question: Is FAIRCAIPI able to reduce human bias? In general, the experiments face limitations, e.g., we only present a proof of concept on a single dataset, investigating a single bias. We do not address how FAIRCAIPI performs with multiple protected variables, which is empirically more often the case, or in the context of multi-label classification or regression. Nevertheless, the setup of our simulation study is in line with evaluations of selected state-of-the-art bis mitigation strategies at the pre-processing [10,21], in-processing [8], and post-processing [24] stages. All aforementioned papers have in common that they mitigate

gender bias. A subset of them uses the German Credit Risk data set [8,10,21]. In general, FAIRCAIPI has some algorithmic shortcomings: Even if users are notified when their decision negatively affects bias-detection metrics, FAIRCAIPI assumes sufficient knowledge to provide optimal feedback. However, what happens when unfair explanations are less obvious than in our vanilla case, e.g., when multivariate correlations of features reproduce bias? Then, our user assumption is probably too strong. Compared to traditional CAIPI [14], we also lack experimental evidence on how an increasing number of counterexamples affects fairness. Using our simple counterexample generation procedure would imply repeating the identical counterexample multiple times. More sophisticated counterexample generators using statistical bootstrap methods or generative approaches are applicable here.

Let us conclude this section and place our findings into the existing research: FAIRCAIPI is a bias-mitigation in-processing method that is located in a specific niche of XIML methods with a fairness objective. According to our literature review, FAIRCAIPI is closely related to a bias-mitigation method that lets users interact with explanations [44]. However, the major difference is that FAIRCAIPI aims for a human–machine partnership, where both parties profit—machine bias is mitigated, and the user's bias is detected. Existing XIML bias-mitigation procedures involve users more rarely. In contrast, they may be more applicable to practical problems because involving users over 100 iterations may be time-consuming. Nevertheless, FAIRCAIPI extends the spectrum of XIML procedures and occupies a specific niche.Experiments show that frequent user interaction is also fruitful for bias mitigation and may be morally superior in the context of fairness.

6. Conclusions

FAIRCAIPI is an in-processing bias-mitigation algorithm that is based on XIML that involves users. Iterative user feedback is used to prone the model's decision-making mechanism into a *fairer* direction—a direction with less biased decision making. Experiments show that bias detection metrics improve during FAIRCAIPI optimization while the predictive quality of the classification model remains stable. FAIRCAIPI can detect and mitigate machine bias. Furthermore, it also detects human bias.

For future work, we plan to generalize our framework: We will mitigate bias in arbitrary classification settings and consider multiple protected attributes at once. Therefore, arbitrary classification includes, for instance, multi-label classification. This will force us to adjust our FAIRCAIPI cycle because a predictive result can be assigned to multiple prediction outcome cases at the same time. In this regard, we will ask two research questions: How does CAIPI user feedback look like in the context of multi-label classification? In addition, how can user feedback in multi-label classification settings be converted into counterexamples? In the simplest case, multiple protected attributes will all be subject to randomization in the counterexample generation step. However, do we need to resolve each correlation between protected attributes, or does it suffice to change a subset of them to mitigate the overall bias of the classification model? In addition, our research will focus on multivariate correlations for bias detection. We assume that FAIRCAIPI is not able to resolve biased decision-making mechanisms in the case of indirect bias because protected attributes might not be directly covered by local explanations. We believe that methods from causal statistics are a possible solution. Instead of presenting attribution values to users, like SHAP does, we will visualize underlying correlations, for instance, by Bayesian networks, to give the user awareness about how features influence each other. This raises the question: How can user feedback to Bayesian networks be transformed into counterexamples? One of our major future research goals is to educate users, even when biased decision making is less obvious. User education needs to be evaluated in dedicated user studies. We will develop appropriate FAIRCAIPI user interfaces. Their clarity and simplicity also offer potential future research directions.

Author Contributions: Conceptualization L.H., E.S., S.S. and U.S.; Methodology L.H., E.S., S.S. and U.S.; Implementation and data curation L.H.; Writing—original draft L.H.; Writing—review and editing E.S., S.S. and U.S.; Supervision U.S.; Funding acquisition and project administration U.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research is being funded by the German Ministry of Education and Research *Project Human-Centered AI in the Chemical Industry, hKI-Chemie* (grant number 01IS21023A) and the Bavarian Ministry of Economy, Development, and Industry, Germany *Project HIX* (grant number DIK0330).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This work exclusively uses public data sets, which are referenced appropriately. All code is available under: https://github.com/emanuelsla/faircaipi (accessed on 11 August 2023).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ML	Machine Learning
XIML	Explanatory and Interactive Machine Learning
LIME	Local Interpretable Model-agnostic Explanations
SHAP	SHapley Additive exPlanations
DP	Deprived (protected) group with Positive (favorable) label
DN	Deprived (protected) group with Negative (unfavorable) label
FP	Favored (unprotected) group with Positive (favorable) label
FN	Favored (unprotected) group with Negative (unfavorable) label
FPERB	False Positive Error Rate Balance
RRR	Right for the Right Reasons
RWR	Right for the Wrong Reasons
W	Wrong

References

1. Angwin, J.; Larson, J.; Mattu, S.; Kirchner, L. Machine Bias. In *Ethics and Data Analytics*; Martin, K., Ed.; Auerbach Publications: New York, NY, USA, 2022; Chapter 6.11.
2. Wolf, M.J.; Miller, K.W.; Grodzinsky, F.S. Why we should have seen that coming: Comments on Microsoft's tay "experiment," and wider implications. *SIGCAS Comput. Soc.* **2017**, *47*, 54–64. [CrossRef]
3. Zhang, L.; Yencha, C. Examining perceptions towards hiring algorithms. *Technol. Soc.* **2022**, *68*, 101848. [CrossRef]
4. Mukerjee, A.; Biswas, R.; Deb, K.; Mathur, A.P. Multi–objective Evolutionary Algorithms for the Risk–return Trade–off in Bank Loan Management. *Int. Trans. Oper. Res.* **2002**, *9*, 583–597. [CrossRef]
5. Goel, N.; Yaghini, M.; Faltings, B. Non-Discriminatory Machine Learning Through Convex Fairness Criteria. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, LA, USA, 2–7 February 2018; McIlraith, S.A., Weinberger, K.Q., Eds.; AAAI Press: Washington, DC, USA, 2018; pp. 3029–3036.
6. Menon, A.K.; Williamson, R.C. The cost of fairness in binary classification. In Proceedings of the Conference on Fairness, Accountability and Transparency, FAT 2018, New York, NY, USA, 23–24 February 2018; Friedler, S.A., Wilson, C., Eds.; PMLR: Cambridge, MA, USA, 2018; Volume 81, pp. 107–118.
7. Kamishima, T.; Akaho, S.; Asoh, H.; Sakuma, J. Fairness-Aware Classifier with Prejudice Remover Regularizer. In Proceedings of the Machine Learning and Knowledge Discovery in Databases-European Conference, ECML PKDD 2012, Bristol, UK, 24–28 September 2012; Proceedings, Part II; Flach, P.A., Bie, T.D., Cristianini, N., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7524, pp. 35–50. [CrossRef]
8. Celis, L.E.; Huang, L.; Keswani, V.; Vishnoi, N.K. Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. In Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, 29–31 January 2019; Boyd, D., Morgenstern, J.H., Eds.; ACM: New York, NY, USA, 2019; pp. 319–328. [CrossRef]
9. Kearns, M.J.; Neel, S.; Roth, A.; Wu, Z.S. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, 10–15 July 2018; Dy, J.G., Krause, A., Eds.; PMLR: Cambridge, MA, USA, 2018, Volume 80, pp. 2569–2577.

10. Kamiran, F.; Calders, T. Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.* **2011**, *33*, 1–33. [CrossRef]
11. Berman, J.J.; Murphy-Berman, V.; Singh, P. Cross-Cultural Similarities and Differences in Perceptions of Fairness. *J.-Cross-Cult. Psychol.* **1985**, *16*, 55–67. [CrossRef]
12. Settles, B. *Active Learning*; Synthesis Lectures on Artificial Intelligence and Machine Learning, Springer: Berlin/Heidelberg, Germany, 2012. [CrossRef]
13. Shivaswamy, P.; Joachims, T. Coactive Learning. *J. Artif. Intell. Res.* **2015**, *53*, 1–40. [CrossRef]
14. Teso, S.; Kersting, K. Explanatory Interactive Machine Learning. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019, Honolulu, HI, USA, 27–28 January 2019; Conitzer, V., Hadfield, G.K., Vallor, S., Eds.; ACM: New York, NY, USA, 2019; pp. 239–245. [CrossRef]
15. Angerschmid, A.; Zhou, J.; Theuermann, K.; Chen, F.; Holzinger, A. Fairness and Explanation in AI-Informed Decision Making. *Mach. Learn. Knowl. Extr.* **2022**, *4*, 556–579.
16. Slany, E.; Ott, Y.; Scheele, S.; Paulus, J.; Schmid, U. CAIPI in Practice: Towards Explainable Interactive Medical Image Classification. In Proceedings of the Artificial Intelligence Applications and Innovations. AIAI 2022 IFIP WG 12.5 International Workshops-MHDW 2022, 5G-PINE 2022, AIBMG 2022, ML@HC 2022, and AIBEI 2022, Hersonissos, Crete, Greece, 17–20 June 2022; Proceedings; Maglogiannis, I., Iliadis, L., MacIntyre, J., Cortez, P., Eds.; Springer: Berlin/Heidelberg, Germany, 2022; Volume 652, pp. 389–400. [CrossRef]
17. Schramowski, P.; Stammer, W.; Teso, S.; Brugger, A.; Herbert, F.; Shao, X.; Luigs, H.; Mahlein, A.; Kersting, K. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nat. Mach. Intell.* **2020**, *2*, 476–486.
18. Bellamy, R.K.E.; Dey, K.; Hind, M.; Hoffman, S.C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilovic, A.; et al. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *arXiv* **2018**, arXiv:1810.01943. Available online: https://arxiv.org/abs/1810.01943 (accessed on 17 July 2023).
19. Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I.D.; Gebru, T. Model Cards for Model Reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, 29–31 January 2019; Boyd, D., Morgenstern, J.H., Eds. ACM: New York, NY, USA, 2019; pp. 220–229. [CrossRef]
20. Kärkkäinen, K.; Joo, J. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, 3–8 January 2021; pp. 1547–1557. [CrossRef]
21. Feldman, M.; Friedler, S.A.; Moeller, J.; Scheidegger, C.; Venkatasubramanian, S. Certifying and Removing Disparate Impact. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, 10–13 August 2015; Cao, L., Zhang, C., Joachims, T., Webb, G.I., Margineantu, D.D., Williams, G., Eds.; ACM: New York, NY, USA, 2015; pp. 259–268. [CrossRef]
22. Calmon, F.P.; Wei, D.; Vinzamuri, B.; Ramamurthy, K.N.; Varshney, K.R. Optimized Pre-Processing for Discrimination Prevention. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R., Eds.; Neural Information Processing Systems Foundation, Inc. (NeurIPS): San Diego, CA, USA, 2017; pp. 3992–4001.
23. Hardt, M.; Price, E.; Srebro, N. Equality of Opportunity in Supervised Learning. In Proceedings of the Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, Barcelona, Spain, 5–10 December 2016; Lee, D.D., Sugiyama, M., von Luxburg, U., Guyon, I., Garnett, R., Eds.; Neural Information Processing Systems Foundation, Inc. (NeurIPS): San Diego, CA, USA, 2016; pp. 3315–3323.
24. Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.M.; Weinberger, K.Q. On Fairness and Calibration. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R., Eds.; Neural Information Processing Systems Foundation, Inc. (NeurIPS): San Diego, CA, USA,2017; pp. 5680–5689.
25. Kamiran, F.; Karim, A.; Zhang, X. Decision Theory for Discrimination-Aware Classification. In Proceedings of the 12th IEEE International Conference on Data Mining, ICDM 2012, Brussels, Belgium, 10–13 December 2012; Zaki, M.J., Siebes, A., Yu, J.X., Goethals, B., Webb, G.I., Wu, X., Eds.; IEEE Computer Society: Washington, DC, USA, 2012; pp. 924–929. [CrossRef]
26. Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; Wallach, H.M. A Reductions Approach to Fair Classification. In Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, 10–15 July 2018; Dy, J.G., Krause, A., Eds.; PMLR: Cambridge, MA, USA, 2018; Volume 80, pp. 60–69.
27. Agarwal, A.; Dudík, M.; Wu, Z.S. Fair Regression: Quantitative Definitions and Reduction-Based Algorithms. In Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Long Beach, CA, USA, 9–15 June 2019; Chaudhuri, K., Salakhutdinov, R., Eds.; PMLR: Cambridge, MA, USA, 2019; Volume 97, pp. 120–129.
28. Bolukbasi, T.; Chang, K.; Zou, J.Y.; Saligrama, V.; Kalai, A.T. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In Proceedings of the Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, Barcelona, Spain, 5–10 December 2016; Lee, D.D., Sugiyama, M., von Luxburg, U., Guyon, I., Garnett, R., Eds.; Neural Information Processing Systems Foundation, Inc. (NeurIPS): San Diego, CA, USA, 2016; pp. 4349–4357.

29. Zhao, J.; Wang, T.; Yatskar, M.; Cotterell, R.; Ordonez, V.; Chang, K. Gender Bias in Contextualized Word Embeddings. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019; Long and Short Papers; Burstein, J.,Doran, C., Solorio, T., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2019; Volume 1, pp. 629–634. [CrossRef]
30. Xu, D.; Yuan, S.; Zhang, L.; Wu, X. FairGAN: Fairness-aware Generative Adversarial Networks. In Proceedings of the IEEE International Conference on Big Data (IEEE BigData 2018), Seattle, WA, USA, 10–13 December 2018; Abe, N., Liu, H., Pu, C., Hu, X., Ahmed, N.K., Qiao, M., Song, Y., Kossmann, D., Liu, B., Lee, K., et al., Eds.; IEEE: Piscataway, NJ, USA, 2018; pp. 570–575. [CrossRef]
31. Louizos, C.; Swersky, K.; Li, Y.; Welling, M.; Zemel, R.S. The Variational Fair Autoencoder. In Proceedings of the 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, 2–4 May 2016; Conference Track Proceedings; Bengio, Y., LeCun, Y., Eds.; 2016. Available online: http://arxiv.org/abs/1511.00830 (accessed on 17 July 2023).
32. Wang, Z.; Qinami, K.; Karakozis, I.C.; Genova, K.; Nair, P.; Hata, K.; Russakovsky, O. Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020; Computer Vision Foundation/IEEE: Piscataway, NJ, USA, 2020; pp. 8916–8925. [CrossRef]
33. Gong, S.; Liu, X.; Jain, A.K. Mitigating Face Recognition Bias via Group Adaptive Classifier. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, 19–25 June 2021; Computer Vision Foundation/IEEE: Piscataway, NJ, USA, 2021; pp. 3414–3424. [CrossRef]
34. Nabi, R.; Shpitser, I. Fair Inference on Outcomes. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, LA, USA, 2–7 February 2018; McIlraith, S.A., Weinberger, K.Q., Eds.; AAAI Press: Washington, DC, USA, 2018; pp. 1931–1940.
35. Loftus, J.R.; Russell, C.; Kusner, M.J.; Silva, R. Causal Reasoning for Algorithmic Fairness. *arXiv* **2018**, arXiv:1805.05859. Available online: https://arxiv.org/abs/1805.05859 (accessed on 17 July 2023).
36. Kilbertus, N.; Rojas-Carulla, M.; Parascandolo, G.; Hardt, M.; Janzing, D.; Schölkopf, B. Avoiding Discrimination through Causal Reasoning. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R., Eds.; Neural Information Processing Systems Foundation, Inc. (NeurIPS): San Diego, CA, USA, 2017; pp. 656–666.
37. Zhang, L.; Wu, Y.; Wu, X. A Causal Framework for Discovering and Removing Direct and Indirect Discrimination. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, 19–25 August 2017; Sierra, C., Ed.; 2017; pp. 3929–3935. [CrossRef]
38. Zhang, L.; Wu, Y.; Wu, X. Causal Modeling-Based Discrimination Discovery and Removal: Criteria, Bounds, and Algorithms. *IEEE Trans. Knowl. Data Eng.* **2019**, *31*, 2035–2050. [CrossRef]
39. Sharma, S.; Henderson, J.; Ghosh, J. CERTIFAI: Counterfactual Explanations for Robustness, Transparency, Interpretability, and Fairness of Artificial Intelligence models. *arXiv* **2019**, arXiv:1905.07857. Available online: https://arxiv.org/abs/1905.07857 (accessed on 17 July 2023).
40. Zhang, J.; Bareinboim, E. Fairness in Decision-Making-The Causal Explanation Formula. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, LA, USA, 2–7 February 2018; McIlraith, S.A., Weinberger, K.Q., Eds.; AAAI Press: Washington, DC, USA, 2018; pp. 2037–2045.
41. Begley, T.; Schwedes, T.; Frye, C.; Feige, I. Explainability for fair machine learning. *arXiv* **2020**, arXiv:2010.07389. Available online: https://arxiv.org/abs/2010.07389 (accessed on 17 July 2023).
42. Cabrera, Á.A.; Epperson, W.; Hohman, F.; Kahng, M.; Morgenstern, J.; Chau, D.H. FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning. In Proceedings of the 14th IEEE Conference on Visual Analytics Science and Technology, IEEE VAST 2019, Vancouver, BC, Canada, 20–25 October 2019; Chang, R., Keim, D.A., Maciejewski, R., Eds.; IEEE: Piscataway, NJ, USA, 2019; pp. 46–56. [CrossRef]
43. Ahn, Y.; Lin, Y. FairSight: Visual Analytics for Fairness in Decision Making. *IEEE Trans. Vis. Comput. Graph.* **2020**, *26*, 1086–1095. [CrossRef] [PubMed]
44. Nakao, Y.; Stumpf, S.; Ahmed, S.; Naseer, A.; Strappelli, L. Towards Involving End-users in Interactive Human-in-the-loop AI Fairness. *arXiv* **2022**, arXiv:2204.10464.
45. Kulesza, T.; Burnett, M.M.; Wong, W.; Stumpf, S. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In Proceedings of the 20th International Conference on Intelligent User Interfaces, IUI 2015, Atlanta, GA, USA, 29 March–1 April 2015; Brdiczka, O., Chau, P., Carenini, G., Pan, S., Kristensson, P.O., Eds.; ACM: New York, NY, USA, 2015; pp. 126–137. [CrossRef]
46. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* **2022**, *54*, 115:1–115:35. [CrossRef]

47. Chen, J.; Kallus, N.; Mao, X.; Svacha, G.; Udell, M. Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved. In Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, 29–31 January 2019; Boyd, D., Morgenstern, J.H., Eds.; ACM: New York, NY, USA, 2019; pp. 339–348. [CrossRef]
48. Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; Zemel, R.S. Fairness through awareness. In Proceedings of the Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, 8–10 January 2012; Goldwasser, S., Ed. ACM: New York, NY, USA, 2012; pp. 214–226. [CrossRef]
49. Chouldechova, A. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* **2017**, *5*, 153–163. [CrossRef] [PubMed]
50. Lundberg, S.M.; Lee, S. A Unified Approach to Interpreting Model Predictions. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R., Eds.; Neural Information Processing Systems Foundation, Inc. (NeurIPS): San Diego, CA, USA, 2017; pp. 4765–4774.
51. Schallner, L.; Rabold, J.; Scholz, O.; Schmid, U. Effect of Superpixel Aggregation on Explanations in LIME— A Case Study with Biological Data. In Proceedings of the Machine Learning and Knowledge Discovery in Databases-International Workshops of ECML PKDD 2019, Würzburg, Germany, 16–20 September 2019; Proceedings, Part I; Cellier, P., Driessens, K., Eds.; Springer: Berlin/Heidelberg, Germany, 2019; Volume 1167, pp. 147–158. [CrossRef]
52. Young, H.P. Monotonic solutions of cooperative games. *Int. J. Game Theory* **1985**, *14*, 65–72.
53. Arenas, M.; Barcelo, P.; Bertossi, L.; Monet, M. On the Complexity of SHAP-Score-Based Explanations: Tractability via Knowledge Compilation and Non-Approximability Results. *J. Mach. Learn. Res.* **2023**, *24*, 1–58.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.