

ORIGINAL RESEARCH

Open access

Interactive Explainable Fair Machine Learning: A Human-in-the-Loop Framework for Bias Detection and Mitigation

James Walker^{1*}, Olivia Harris¹, George Wilson²

Abstract

The expansion of machine-learning frameworks into sectors with significant end-user consequences has heightened concerns regarding the equity of trained models, particularly the necessity of preventing biases that disadvantage distinct demographic sub-groups. The majority of current bias-mitigation techniques adjust the relative weight of data observations during the pre-processing phase. Because fairness depends heavily on context, we support an interactive machine-learning strategy that empowers human supervisors to offer iterative guidance for model refinement. To this end, we adapt the explanatory interactive machine-learning framework Caipi for equitable machine learning. This modified approach, FairCaipi, integrates human-in-the-loop interventions regarding both model predictions and underlying explanations to enhance model equity. Empirical outcomes indicate that FairCaipi outpaces a state-of-the-art pre-processing bias mitigation strategy in both the equity metrics and the predictive accuracy of the resulting machine-learning framework. We show that FairCaipi serves to expose and minimize algorithmic bias in machine-learning architectures while simultaneously identifying underlying human bias.

Keywords Fair machine learning, Explanatory and interactive machine learning, FairCaipi, Bias-mitigation

*Correspondence:

James Walker

james.walker@ed.ac.uk

¹ Department of Interactive XAI and Fair Machine Learning, Faculty of Engineering, University of Edinburgh, Edinburgh, United Kingdom

² Department of Human-in-the-Loop Bias Mitigation, Faculty of Engineering, University of Leeds, Leeds, United Kingdom

Introduction

The identification of discriminatory outcomes in machine-learning deployment has discredited the widespread assumption that machine-learning (ML) systems provide objective determinations. As an illustration, the COMPAS tool (short for Correctional Offender Management Profiling for Alternative Sanctions) inaccurately attributes a higher probability of recidivism to Black defendants compared to white defendants, demonstrating an explicit racial bias [1]. Similarly, Microsoft's TayAI chatbot produced racist and anti-Semitic expressions [2]. Nonetheless, machine-learning frameworks are increasingly deployed in high-stakes environments, including automated hiring recruitment tools [3] and credit application reviews [4]. We contend that ML practitioners possess an ethical responsibility to ensure these algorithms do not perpetuate systemic biases present in source data, thereby systematically marginalizing vulnerable populations, irrespective of existing statutory obligations.

A variety of methodologies integrate equity—or bias reduction—as a secondary optimization goal, such as satisfying specific fairness constraints during the optimization phase [5-7] or designing completely novel classification architectures

[8, 9]. Celis et al. (2019) formulate a generalized classification algorithm that combines multiple distinct equity metrics into a linear group performance function used as an optimization target. Achieving fairness goals during the active optimization process represents a fundamentally different paradigm for bias reduction than adjusting data equity prior to training. For example, Reweighting alters the distribution of historically marginalized groups before a classification model is trained, ensuring it complies with one designated equity metric [10]. Strategies like Reweighting often yield superior performance over broader methods when measured against specific individual metrics, whereas the former class of techniques is advantageous when consolidating multiple disparate fairness constraints.

A shared characteristic among dominant bias-mitigation systems is their tendency to treat equity as a static, context-independent attribute. Crucially, definitions of equity are shaped by specific cultural frameworks, whether at a national or organizational level, emphasizing that the interpretation of fairness is deeply subjective [11]. Berman et al. (1985) analyze a scenario where employees at an Indian firm and a US firm were independently requested to allocate corporate perks among peers. While the Indian workforce viewed it as fair to allocate perks according to individual financial necessity, the US personnel favored an allocation model tied directly to individual productivity. Because fairness is deeply dependent on situational context, we argue that current paradigms relying on static fairness definitions must be broadened to incorporate interactive ML methods that accommodate the user's cultural assumptions regarding equity. To mirror these individualized views of fairness, users require the ability to guide the optimization pathways of ML models. In active learning [12], users iteratively annotate data points that optimize the information gain for the classification framework. Coactive learning modifies this active learning setup by allowing users to interactively adjust and correct the classifier's predictions [13].

Active and coactive learning frameworks assume that human operators completely comprehend the underlying decision-making systems of the problems modeled by ML, an assumption we consider overly optimistic. Because interpretations of fairness depend on context, different users possess highly individualized views of equitable decision-making. Explanatory and interactive ML (XIML) strengthens the interactive component of active learning by incorporating explanatory ML tools, thereby giving the operator visibility into how an ML model arrives at its conclusions [14]. Consequently, XIML opens up the model's inner reasoning using explanations and permits the operator to amend both the prediction and the underlying rationale. Angerschmid et al. (2022) [15] demonstrate that embedding an explanatory element within the decision-making pipeline improves the human perception of systemic fairness. Indeed, the goal of Caipi [14], a prominent XIML algorithm, is to utilize user guidance to steer the model toward a valid decision-making path from the viewpoint of a subject-matter expert. To facilitate this, Caipi has been integrated with a user interface designed for automated medical image analysis [16]. Other researchers focus on structural algorithmic changes; for example, Schramowski et al. (2020) [17] specifically optimize Caipi for deep learning architectures.

To illustrate the benefits of our interactive training model, let us look at a hypothetical scenario in the field of credit risk analysis, as depicted in **Figure 1**. The principal goals of credit risk analysis are evaluating the creditworthiness of a prospective borrower, minimizing corporate risk exposure, and sustaining a balanced loan asset pool. Initially, assume the applicant is a first-year philosophy student who intends to pursue an academic career path and whose current revenue is derived solely from part-time employment. This request is rejected automatically based on the applicant's youth, financial profile, and job classification. In this instance, the machine-learning framework's determination was precise and fair given the context of the student and the constraints of the banking institution—meaning the explanation is unbiased and no operator adjustment is required. Second, consider a scenario where the applicant holds a fixed-term research position in computer science. The model, focusing exclusively on the instability of a temporary employment contract, might initially deny the credit request. In this case, the subject-matter expert steps in because they can logically project a robust job market within the technology sector, subsequently correcting the classification label. Following this, the expert adds this updated case to the training set and retrain the algorithm to enhance its accuracy. Third, consider a female MSc student specializing in biology. The algorithm denies the application, citing the gender feature as a primary driver. The subject-matter expert agrees with the denial itself but disputes the explanation that the determination was driven by gender. Their analysis reveals that the explanation contains bias; specifically, female candidates are systematically treated as less creditworthy. Nevertheless, they conclude that the valid drivers for credit denial should be the applicant's lack of

permanent employment and insufficient personal capital to satisfy the bank's underwriting criteria. As a result, the expert steps in, correcting this explanation bias by introducing counterexamples into the dataset to refine the model.

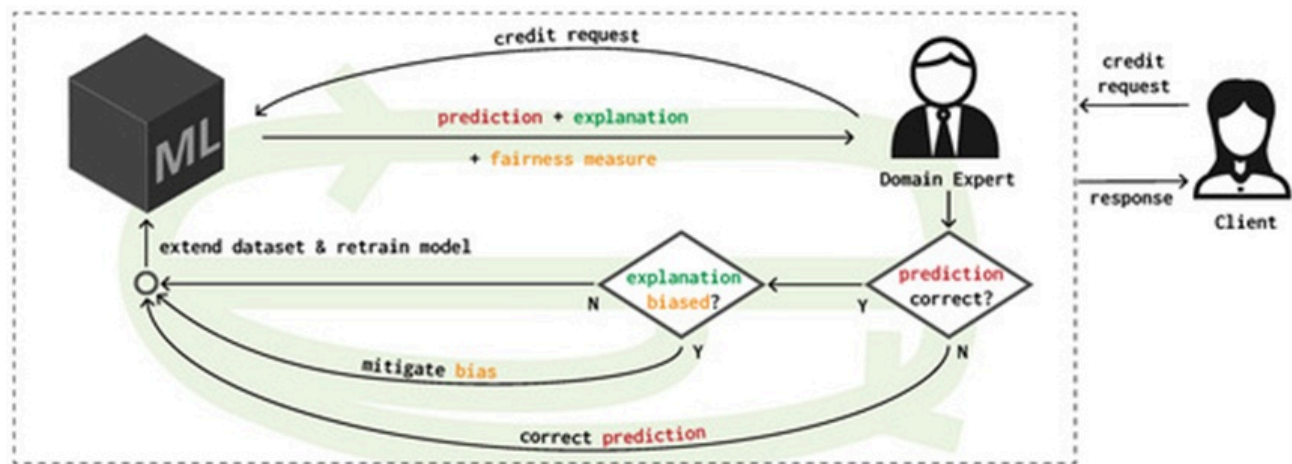


Figure 1. Interactive learning in the domain of lending.

We introduce FairCaipi, a XIML methodology constructed upon the Caipi algorithm. The core innovation of FairCaipi is that it permits the supervisor to engage directly with the model's internal reasoning from an equity perspective. Instead of enforcing predefined "correct" pathways, users offer iterative critiques to optimize the model's overall fairness. Distorted decision-making pathways are corrected via user interventions. Through FairCaipi, we deliver a framework capable of (i) revealing and (ii) minimizing algorithmic bias, while also (iii) identifying human bias during the model optimization process. This study outlines the underlying theory and offers a structured derivation of the FairCaipi pipeline. We validate our methodology using a simulation protocol and directly evaluate FairCaipi against Reweighting, a prominent pre-processing technique for bias reduction [10]. The research questions guiding our experimental assessment are:

- **(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 Reweighting strategy?

Approaches to Fairness in Machine Learning

Strategies designed to mitigate algorithmic bias are broadly categorized into three distinct operational intervals: pre-processing, in-processing, and post-processing frameworks [18]. For a structural breakdown, Table 1 organizes and clusters these contemporary methodologies. We review these options below to contextualize where FairCaipi fits within the broader discipline of equitable ML:

Pre-processing focuses on meeting specific equity benchmarks during the initial data assembly stage [19], a practice frequently seen in facial recognition developments [20]. More broadly applicable pre-processing schemes alter sample distributions by modifying case weights to foster greater algorithmic parity, such as Reweighting, which balances the ratios of marginalized and favored cohorts prior to training [10]. Other foundational examples include Disparate Impact Removal [21] and Optimized Pre-Processing [22].

Post-processing methods, conversely, manipulate the predictions generated by an already trained ML system to satisfy predefined equity rules. Certain techniques adjust probability thresholds to narrow the discrepancy in predictions between disadvantaged and favored cohorts, implementing either simplified linear setups [23] or complex classification systems

[24]. Reject Option Classification operates similarly by switching predicted outcomes for data instances located near the critical zones of decision boundaries [25].

In-processing methods share the closest technical alignment with FairCaipi, as they embed fairness targets directly into the model training phase by enforcing mathematical fairness metrics [5-8]. More intricate variants include custom cost functions that penalize errors differently depending on the group membership of an instance [26], a concept that has also been adapted for regression problems [27]. In-processing bias reduction has also been extended to niche domains, such as developing unbiased word embeddings for Natural Language Processing [28, 29], creating fair generative architectures like Generative Adversarial Networks [30] or Variational Autoencoders [31], ensuring unbiased visual recognition in deep learning pipelines [32, 33], and formulating fair causal structures like Bayesian networks [34-38].

Certain in-processing systems employ explainable AI techniques—such as counterfactual arguments [39], causal explanations [40], or Shapley values [41]—to expose underlying algorithmic bias. The use of Shapley values is highly relevant to our framework, as FairCaipi evaluates the model's inner reasoning using local explanations derived from Shapley values. Presently, human interaction with fairness in ML models is predominantly confined to visualization tools designed to expose model biases via user interfaces [42, 43]. A closely related methodology to FairCaipi is the XIML system developed by Nakao [44], which incorporates human feedback on explanations directly into model optimization to enhance fairness. Utilizing a mechanism for interactive user-explanation adjustments [45], it empowers users to modify a classifier's biased behavior across multiple protected characteristics simultaneously. In contrast, FairCaipi structures this interactive user-explanation dialogue within a continuous optimization loop.

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.

Before Model Training (Pre-Processing)	During Model Training (In-Processing)	After Model Training (Post-Processing)
Equitable data acquisition strategies [19, 20]	Incorporation of fairness-oriented constraints [5-8]	Adjustment of decision thresholds to improve fairness [23, 24]
Sampling techniques for reducing dataset bias [10, 21, 22]	Fairness-aware objective and cost functions [26, 27]	Modification of model outputs and predictions [25]
	Fair approaches in Natural Language Processing (NLP) [28, 29]	
	Fairness-enhanced generative modeling techniques [30, 31]	
	Bias-free image classification frameworks [32, 33]	
	Fair causal modeling methods [34-38]	
	Explanation-driven bias reduction approaches [39-41]	
	Interactive techniques for mitigating bias [42, 43]	
	Explainable and Interpretable Machine Learning (XIML) methods for bias mitigation [44]	

Materials and Methods

FairCaipi integrates the core technical concepts of bias mitigation and XIML. As a result, this section addresses both domains. It commences with a concise overview of the German Credit Risk dataset, which serves as a running example throughout this discussion and forms the basis of the simulation experiment validating FairCaipi. The German Credit Risk dataset exhibits a demonstrable gender bias. Because FairCaipi serves as an extension of the Caipi framework, the technical derivation of XIML will focus primarily on Caipi. We will modify Caipi to meet equity objectives, presenting a human-ML system capable of (i) exposing and (ii) minimizing machine bias, while also (iii) identifying human bias. The experimental configuration is detailed at the conclusion of this section. First, we outline the fundamental mathematical notation utilized throughout this article:

Notation.

A binary classification model $f: \mathcal{X}^n \rightarrow \mathcal{Y}$ represents a mapping function from a feature space $\mathcal{X}^n$ comprising n attributes to a target space $\mathcal{Y} = \{0, 1\}$. For simplicity, the superscript n is omitted hereafter, and the feature space is denoted as $\mathcal{X}$. A model inference is represented by $y = f(x)$. An individual data instance $x \in \mathcal{X}$ is denoted as a feature value vector $x = (x_1, x_2, \dots, x_n) \in \mathcal{X}$, where x_i represents a single attribute within x at index i . Furthermore, let $l: \mathcal{X} \rightarrow \mathcal{Y}$ represent a ground-truth labeling function that maps data instances to their true class labels.

Let $\mathcal{L} \subseteq \mathcal{X} \times \mathcal{Y}$ and $\mathcal{U} \subseteq \mathcal{X}$ denote the subsets of labeled and unlabeled data instances, respectively, where $\mathcal{X}$ and $\mathcal{Y}$ designate the instance attributes and corresponding labels within $\mathcal{L}$. Additionally, we employ $x(n)$ ($y(n)$) to represent the n -th feature (label) instance within $\mathcal{X}$ ($\mathcal{Y}$) when the reference set is clear from the context, or append a subscript such as $x(n)_{\mathcal{L}}$ ($y(n)_{\mathcal{L}}$) and $x(n)_{\mathcal{U}}$ to explicitly indicate the associated dataset. We assume the availability of a routine `Fit` to optimize and update the classification model.

Bias in the german credit risk dataset

The German Credit Risk dataset (<https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data>, accessed on 17 July 2023) displays a distinct gender bias, as illustrated in **Figure 2**. Within fairness literature, bias systematically advantages a privileged demographic group while disadvantaging an unprivileged demographic group [18]. Privileged and unprivileged segments are identified by the distribution of a favorable class label [18] relative to a defined protected attribute [46, 47]. Discrimination occurs when an unprivileged group faces systematic disadvantages stemming from a protected characteristic.

It is important to note that discrimination can manifest either directly or indirectly. Direct discrimination is tied explicitly to a protected attribute, whereas indirect discrimination arises from seemingly objective features that are highly correlated with the protected attribute [37]. To analyze and quantify this bias, the data is partitioned into four distinct categories [10]:

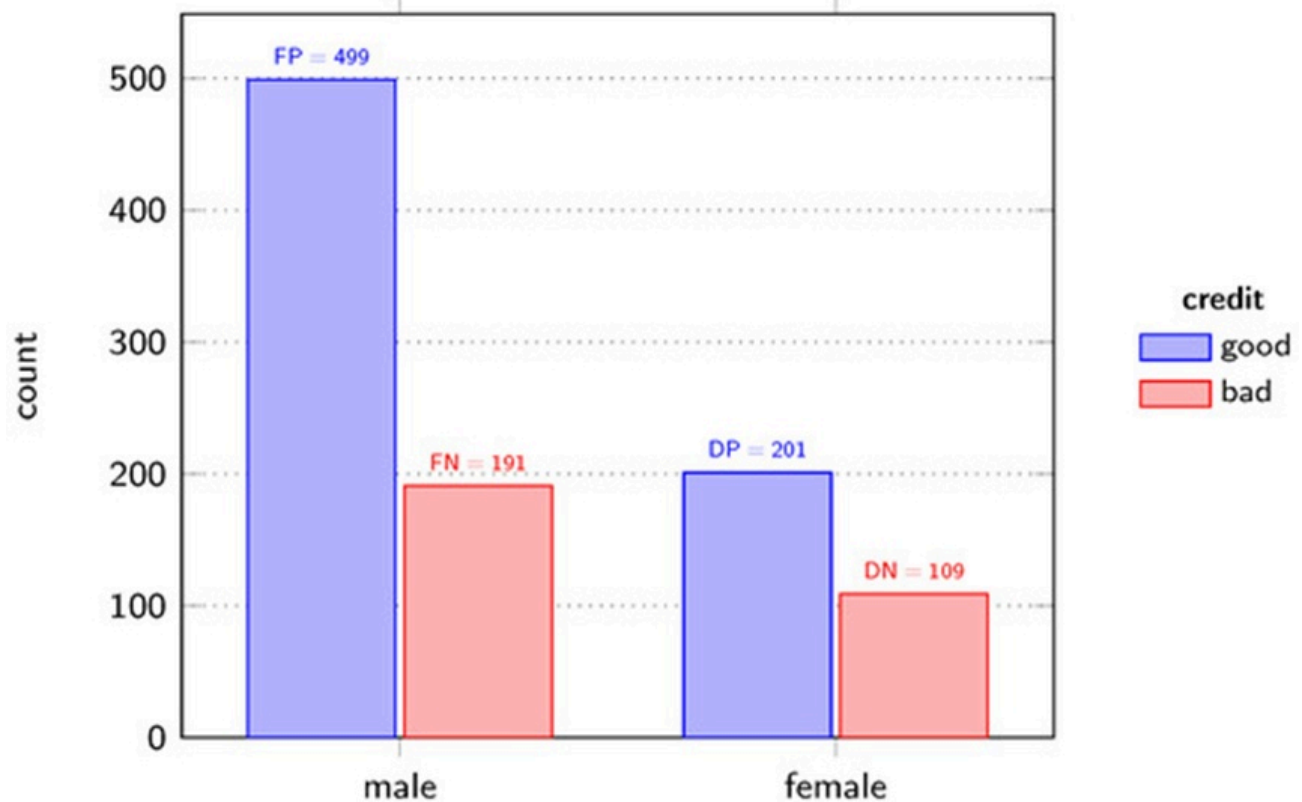


Figure 2. Visualization Favored group with Positive label (FP), Favored group with Negative label (FN), Deprived group with Positive label (DP), and Deprived group with Negative label (DN). This figure investigates the favorable label good credit risk regarding the protected feature gender and reveals a gender bias, as it is more likely to have a good credit risk for males than for females.

- **DP:** Deprived (unprivileged) group with Positive (favorable) label
- **DN:** Deprived (unprivileged) group with Negative (unfavorable) label
- **FP:** Favored (privileged) group with Positive (favorable) label
- **FN:** Favored (privileged) group with Negative (unfavorable) label

As shown in **Figure 2**, which illustrates the distribution of the favorable outcome *good credit risk* conditioned on the protected characteristic *gender*, a clear gender imbalance emerges: the favored group (*male*) exhibits a higher probability of receiving the positive label than the deprived group (*female*).

Up to this point, bias has been described primarily in a narrative manner, which prompts the question: How can it be mathematically quantified? Measures used to evaluate these imbalances are known as bias-detection metrics. The majority of bias-detection metrics evaluate differences in conditional performance across a classification model. For example, an equitable classifier demands that the model's recall—which, in this scenario, represents the true positive rate of identified *good credit risks*—remains consistent irrespective of the applicant's *gender* as a protected attribute.

Definition 1

(Protected Attribute). Let S represent the protected attribute, where $S=s$ and $S=\bar{s}$ denote the privileged and unprivileged cohorts, respectively.

Definition 2

(Favorable Label). Let $\hat{y} = (x)$ signify a prediction, where $\hat{y} = d$ represents the favorable outcome and $\hat{y} = d^-$ denotes the unfavorable outcome.

Below, we outline five distinct metrics for bias detection: Statistical Parity [48], Equalized Odds [23], Equalized Opportunity [23], False Positive Error Rate Balance [49], and Predictive Parity [49]. Based on the formulation of Statistical Parity (SP) [48], a classification model achieves fairness when the likelihood of being assigned an unfavorable label is distributed uniformly across both privileged and unprivileged cohorts, which is expressed as:

$$P(\hat{y} = d^- | S = s) \stackrel{!}{=} P(\hat{y} = d^- | S = s^-), \quad (1)$$

where P indicates the probability of the predicted outcome $\hat{y}$ conditioned upon the protected attribute S . When evaluating the likelihood of receiving an unfavorable label, denoted as $(\hat{y} = d^-)$, with respect to the privileged group $S=s$ and the unprivileged group $S=s^-$, we derive the conditional expressions $(\hat{y} = d^- | S=s)$ and $(\hat{y} = d^- | S=s^-)$, respectively.

It directly follows from the definition established in Eq. 1 that Statistical Parity is satisfied if the difference in conditional probabilities for receiving the unfavorable label between the privileged and unprivileged groups equals zero. The corresponding differential representation of (1) is formalized in Eq. 2, establishing that an unbiased classifier must produce a Statistical Parity estimate of $SP=0$, formulated as:

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

The primary distinction separating Statistical Parity from the alternative bias-detection metrics is its independence from ground truth data; it evaluates fairness strictly using the predicted outcomes conditioned on a designated protected attribute. Conversely, Equalized Odds (EqOdds) [23] incorporates the true label y into the conditional criteria for both sides of the expression:

$$P(\hat{y} = d^- | y = d^-, S = s) \stackrel{!}{=} P(\hat{y} = d^- | y = d^-, S = s^-). \quad (3)$$

Eq. 4 characterizes EqOdds as the mean value of two distinct differences. Each component of this summation represents a performance disparity between the privileged and unprivileged cohorts. Here, the conditional probabilities are substituted with standard performance metrics derived from binary classification. Accordingly, Equalized Odds accounts for both the false positive rate (fpr) and the true positive rate (tpr):

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

In practice, the core concept underlying (3) serves as the foundation for Equal Opportunity (EqOpp) [23], False Positive Error Rate Balance (FPERB) [49], and Predictive Parity (PP) [49], where EqOpp utilizes the tpr component in (5) and FPERB applies the fpr component in (6). Predictive Parity diverges slightly by utilizing the false discovery rate (fdr) in (7), which is computed via $fpr \times (fpr + tpr) - 1$.

$$EqOpp = tpr_{S=s^-} - tpr_{S=s} \quad (5)$$

$$FPERB = fpr_{S=s^-} - fpr_{S=s} \quad (6)$$

$$PP = \int dr_{s=g} = -\int dr_{s=g} \quad (7)$$

The proposed explanatory and interactive FairCaipi pipeline will introduce these aforementioned bias-detection metrics to the user. Its primary objective is to alert and instruct users regarding how modifications to the system alter the fairness of the classifier. Additionally, Statistical Parity (2) serves a critical function within the simulation study's benchmark evaluation, given that the Reweighting mechanism is explicitly optimized to minimize Statistical Parity discrepancies.

Explanatory interactive machine learning

The contemporary XIML framework known as Caipi [14] engages users through an iterative process that incorporates corrections for both predictions and explanations. Within the Caipi framework, prediction outcomes are categorized into three distinct states: Right for the Right Reasons (RRR), Right for the Wrong Reasons (RWR), or Wrong for the Wrong Reasons (W). The latter two inaccurate states necessitate user intervention.

While users modify the target label when addressing a W scenario, they supply corrective feedback for flawed rationale in an RWR scenario, prompting the generation of counterexamples. These counterexamples constitute novel, supplementary instances that isolate only the definitive features. They are designed to guide the underlying mechanics of the model toward a more accurate alignment with the user's domain perspective.

From a practical standpoint, leveraging appropriate data augmentation strategies allows a singular explanation adjustment to yield a multitude of varied counterexamples. Throughout the remainder of this study, we presume that a dedicated operation denoted as Gen produces these counterexamples based on user-driven explanation adjustments.

Definition 3

(Counterexample Generation). Let $\hat{y} = (x)$ represent a prediction, and assume that a user identifies the vector x^* as the decisive factor for $\hat{y}$. This implies that the attribution impact of every non-zero element within x^* surpasses a predefined threshold. The vector x^* is constructed from x such that the collection of non-zero elements in x^* forms a subset of the non-zero elements present in x . Let us define a mechanism Gen that accepts x , $\hat{y}$, x^* , and c as input parameters and generates the counterexample feature and target datasets $\mathcal{X}$ and $\mathcal{Y}$. For the target dataset $\mathcal{Y}$, the predicted label $\hat{y}$ is replicated c times. For the feature dataset $\mathcal{X}$, the original instance x is likewise replicated c times. For any feature component x_i not specified within x^* , the value of x_i undergoes perturbation, such as through a randomization process.

To capture local explanations from a classification model, we employ SHapley Additive exPlanations (SHAP) [50], whereas the traditional Caipi framework utilizes LIME. At this juncture, SHAP emerges as a more advantageous selection because the performance of LIME exhibits high sensitivity to segmentation algorithms [51]. In contrast, SHAP demonstrates robust reliability when applied to tabular datasets [50]. The SHAP explanation approach approximates an underlying model f for a specific input vector x via an explanatory model g . This explanatory model leverages a simplified input x' that is linked to the original input x through a mapping function $h(x') = x$. The formulation ensures that $(z') \approx f(h(x(z')))$ whenever $z' \approx x'$, where $z' \in \{0,1\}^M$ and M denotes the total number of simplified input features. The foundational basis of the SHAP technique rests on Shapley values, which quantify the significance of a feature x_i toward a prediction by computing the marginal impact of incorporating information about that feature. The exact contribution of a feature value x_i to a predicted outcome is designated as the SHAP value $\phi_i \in \mathbb{R}$, which is structured such that the summation of all individual feature attributions approximates the output (x) of the primary model.

The SHAP methodology is established upon three core properties: local accuracy, missingness, and consistency [50]. The simplified representation of a classification model (x) provided by SHAP, which incorporates a simplified feature space representation x' , is expressed as:

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

This formulation inherently satisfies the local accuracy property provided that $x = h_x(x')$ remains true [50]. SHAP approximates a target model f for a given input x by summing individual attributions in a manner that ensures the explanatory model $g(x')$ for the simplified input x' matches the output of the native model $f(x)$. In its standard configuration, attributions ϕ are combined linearly, where each ϕ signifies the importance of a distinct feature (or a specific subset combination of features) across a dimension size M . The baseline attribution ϕ_0 is derived as $\phi_0 = (h_x(\mathbf{0}))$.

Beyond maintaining local accuracy, SHAP adheres to the properties of missingness and consistency [50]. The missingness property dictates that feature values equal to zero (which represent missing features in this context) receive an attribution value of zero. Meanwhile, the consistency property ensures that any increase in a feature's true importance within f is reflected by an appropriately scaled attribution value ϕ .

By satisfying all three of these foundational principles, SHAP attributions are mathematically determined as:

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

where $z' \subseteq x'$ represents all vectors z' whose non-zero elements constitute a subset of the non-zero elements within x' , $|z'|$ signifies the count of non-zero entries in z' , $f_x(z') = f(h_x(z'))$ and $z' \setminus i$ indicates setting the specific index $z'_i = 0$. The right side of Eq. 9 embodies the classical theory of Shapley values, representing the performance difference for f_x when the i th feature of z' is included versus when it is excluded. Eq. 9 represents the unique solution derived from the simultaneous requirements of local accuracy, missingness, and consistency [50]. Young (1985) [52] originally demonstrated that Shapley values represent the unique configurations capable of satisfying local accuracy and missingness. While Young (1985) [52] utilized a supplementary axiom, Lundberg and Lee (2017) demonstrated that this additional axiom is non-essential. According to Lundberg and Lee (2017), the missingness property is not strictly essential for Shapley values on their own, yet it remains vital for additive feature attribution frameworks like SHAP.

Consider Figure 3 as an illustrative instance of a SHAP explanation for an arbitrary observation processed by a classification model with a decision threshold score of 0.5. Within this setup, scores below 0.5 indicate a favorable credit risk, while scores greater than or equal to 0.5 indicate an unfavorable credit risk. The baseline value for SHAP is positioned at approximately 0.27. Moving from this baseline, specific attributes exert a positive influence—meaning they possess a positive attribution value—which shifts the model's judgment toward an unfavorable credit risk designation. In this scenario, the primary drivers pushing the decision toward an unfavorable credit risk include having an active bank account status (i.e., not None) where the average monthly incoming payments over the preceding year fell below 200 German Mark, being female (sex=0), and not having maintained employment for a duration exceeding 4 years (employment=4+years=0). Conversely, the evaluated observation requested a credit amount of 2278 German Mark, which represents the singular feature assigned a negative attribution value, thereby driving the assessment toward a favorable credit risk.

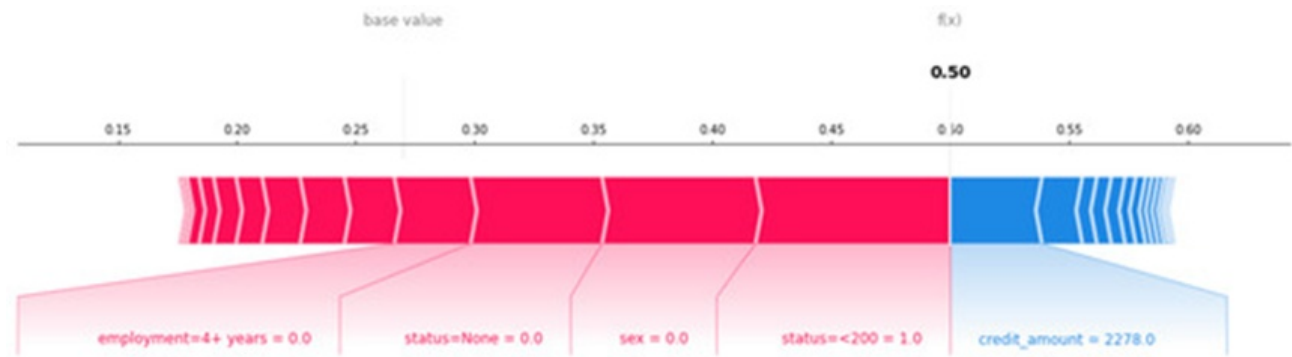


Figure 3. SHAP explanation for a test instance of a classification model trained on the German Credit Risk dataset. The classification score matches the decision boundary of 0.5. SHAP estimates attribution values for each feature. It mimics the classification model's outcome and builds a sum of each 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.forceplot.html>, accessed on 19 July 2023).

The CAIPI framework incorporates a local explanation protocol that processes an individual feature instance alongside a classification model to reveal the most influential features to the end-user. From a mathematical perspective, our explanation procedure Exp is constructed upon the SHAP framework.

Definition 4

(Local Explanation). Let ϕ denote the attribution values allocated to x , which highlight the relative importance of each feature x_i in x toward the final prediction $f(x)$ as formulated in Eq. 9. Additionally, let α represent a designated significance threshold. Let $e \subseteq x$ signify the specific set of features for which the condition $|\phi_i| > \alpha$ is satisfied. We assume an evaluation protocol Exp that accepts x , f , and α as inputs and yields the set e .

CAIPI [14] capitalizes on user feedback regarding both the model's prediction and explanation according to the state of the prediction outcome. During each consecutive iteration, it isolates the most informative instance from an unlabelled dataset—specifically targeting the instance whose prediction score lies closest to the decision boundary (set at 0.5 in this application) relative to a classifier trained on a restricted pool of labeled samples. We posit that this instance yields the highest information value because predictions located near the decision boundary are associated with elevated levels of uncertainty. Securing the true label for this specific instance maximizes the information gained by the classifier during its subsequent training iteration. The selection mechanism MII extracts the index corresponding to this most informative instance. For this operation, we assume direct access to both the predictive scores and the established decision boundary.

Definition 5

(Most Informative Instance). Let the routine MII accept a collection of predictions $\mathcal{Y}$ and a decision threshold β as input parameters. It outputs the specific index m representing the most informative instance, defined as the prediction score positioned closest to β .

The CAIPI routine demands human feedback at two distinct junctures: first, to assess the validity of the prediction, and second, to verify the accuracy of the explanation. In the latter scenario, users can guide the model toward a more desirable decision-making mechanism by adjusting the local explanation. We denote these points of user interaction via

the routine *Interact* and provide a summary of the Caipi framework within **Algorithm 1**. In every operational cycle, Caipi optimizes a model using the existing labeled data pool (line 2) and computes predictions across the remaining unlabeled instances (line 3) to isolate the most informative instance (line 4). The human supervisor evaluates the output and inputs the true label if the prediction proves incorrect (line 7). Alternatively, if the prediction is accurate, Caipi displays the corresponding local explanation, allowing the user to make adjustments if errors are detected (line 9). If the explanation is deemed valid, the instance is integrated into the labeled training set (line 12); otherwise, a set of counterexamples is generated (line 14). To prepare for the subsequent iteration, the current most informative instance is extracted and removed from the unlabelled data collection (line 15). Differing from the original presentation of the Caipi algorithm [14], we explicitly formalize each structural component. This explicit formalization will be utilized in the subsequent section to adapt the Caipi framework to fair ML.

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

```

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

Fair explanatory and interactive machine learning

FairCaipi modifies the foundational Caipi architecture to incorporate a fairness objective through two specific mechanisms: (i) it assesses the local explanation to identify discriminatory decision-making processes, and (ii) it consequently factors in protected attributes during the synthesis of counterexamples. With respect to modification (i), we utilize the groups DP, DN, FP, and FN introduced. Here, D represents the disadvantaged group and F represents the advantaged group, with each group being associated with either a beneficial positive outcome (P) or a detrimental negative outcome (N). We contend that a disproportionate representation of DN and FP indicates the presence of bias. This is because the former allocates the unfavorable outcome to the disadvantaged group, while the latter grants the favorable outcome to the advantaged group. Accordingly, we characterize an explanation—which represents a decision-making framework—as inequitable if membership in the disadvantaged group serves as a basis for receiving the unfavorable classification. In contrast, belonging to the advantaged group serves as a basis for receiving the favorable classification. With respect to modification (ii), our objective is to eliminate protected attributes from the decision-making

framework. This is accomplished by randomizing the protected attributes while holding all other features invariant during the generation of counterexamples. In this context, randomization implies that if being male is identified as a reason for a favorable credit risk evaluation, our generated counterexample will be an identical instance except that the gender attribute is altered to female. Let us formally define the concept 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 $\hat{y} = (x) \in \{d, \bar{d}\}$, and a protected attribute S with deprived group s^- and favored group s . We define the decision-making mechanism of f to be biased if it holds that

$$\begin{aligned} \hat{y} &= \bar{d} \text{ and } \exists e_i \\ &\in e, e_i \\ &= \bar{S} \\ &= \bar{s}, \text{ or } \hat{y} \\ &= d \text{ and } \exists e_i \\ &\in e, e_i \\ &= S \\ &= s. \end{aligned}$$

The bias-reduction strategy of FairCaipi is integrated into the counterexample generation routine GEN' (**Algorithm 2**). In this routine, we determine the specific configuration of the protected attribute that would sustain a bias relative to the prediction (line 2). For instance, if the model predicts a negative credit risk, the bias-sustaining configuration for the protected attribute of gender would be female. Subsequently, we construct a set containing all viable values of the protected attribute, excluding the bias-sustaining value (line 3). To synthesize a counterexample instance, we replicate each original feature but substitute the protected attribute with a randomly selected alternative (line 5), followed by appending the input label (line 6). The resulting counterexample dataset preserves all original feature correlations except for the association between the protected attribute and the target variable.

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:    $x' \leftarrow \text{sample}(s')$  if  $x_{i=S} = s^*$  else  $x_i$  for  $x_i \in x$ 
6:    $\mathcal{X}' \leftarrow \mathcal{X}' \cup \{x'\}; \mathcal{Y}' \leftarrow \mathcal{Y}' \cup \{y\}$ 
7: return  $\mathcal{X}', \mathcal{Y}'$ 
```

Algorithm 2. GEN' $[x, y, S, c]$ (Counterexample Generation)

In contrast to the standard Caipi framework, FairCaipi (**Algorithm 3**) accepts the protected attribute as an explicit input. Users continue to supply feedback regarding the model's explanations; however, their objective shifts from establishing correct mechanisms to mitigating bias. Consequently, in accordance with Definition 6, the interaction step in line 10 yields True if the explanation exhibits bias and False otherwise, where a detected bias triggers the production of bias-mitigating counterexamples (line 14).

FairCaipi is capable of (i) identifying and (ii) mitigating machine bias, as well as (iii) exposing human bias. It should be noted, however, that from a technical perspective, FairCaipi does not strictly necessitate human intervention provided that all components of **Algorithm 3** are accessible, which represents a realistic scenario when experiments are conducted in a pseudo-unlabeled manner. While this enables FairCaipi to achieve the first two capabilities, the third capability remains unfulfilled under these conditions. In practical applications, we deliberately seek to incorporate a human operator within

the optimization loop. The user possesses the authority to overrule a biased decision identified by FairCaipi (Definition 6). Such interventions may be influenced by human factors, resulting either in passive bias—where users neglect to rectify a discriminatory mechanism—or active bias, where users purposefully avoid making corrections or alter the mechanism in a manner that degrades the bias-detection metrics.

To address this, we enhance FairCaipi to display bias-detection metrics to the user at both the initialization and the conclusion of every iteration, specifically prior to and following the retraining of the model (line 2). Additionally, we introduce an interactive phase that allows the user to offer feedback, during which FairCaipi alerts the operator if executed or omitted corrections unfavorably impact the bias-detection metrics. Integrating FairCaipi into the previously outlined architecture (Figure 4) provides a critical advantage: it instructs users by establishing a direct correlation between human input and formal bias-detection metrics.

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

Output: Classification model f

```

1: for  $i \leftarrow 1 : n$  do
2:    $f \leftarrow \text{FIT}(\mathcal{L})$  and  $\text{ComputeFairnessMetrics}(f, S, \mathcal{L})$ 
       $\triangleright$  Compare fairness before and after refitting the model and present to the user
3:    $\hat{\mathcal{Y}} \leftarrow \{f(u) \mid u \in \mathcal{U}\}$ 
4:    $m \leftarrow \text{MII}(\hat{\mathcal{Y}}, 0.5)$ 
5:    $y^{(m)} \leftarrow \text{INTERACT}(x_{\mathcal{U}}^{(m)})$   $\triangleright$  Label retrieved from human annotator
6:   if  $\hat{y}^{(m)} \neq y^{(m)}$  then
7:      $\mathcal{L} \leftarrow \mathcal{L} \cup \{(x_{\mathcal{U}}^{(m)}, y^{(m)})\}$   $\triangleright$  Case W
8:   else
9:      $e \leftarrow \text{EXP}(x_{\mathcal{U}}^{(m)}, f, 0.005)$ 
10:     $b \leftarrow \text{INTERACT}(e)$   $\triangleright$  Bias information retrieved from human annotator
11:    if not  $b$  then
12:       $\mathcal{L} \leftarrow \mathcal{L} \cup \{(x_{\mathcal{U}}^{(m)}, \hat{y}^{(m)})\}$   $\triangleright$  Case RRR
13:    else
14:       $\mathcal{L} \leftarrow \mathcal{L} \cup \text{GEN}'(x_{\mathcal{U}}^{(m)}, \hat{y}^{(m)}, S, c) \cup \{(x_{\mathcal{U}}^{(m)}, \hat{y}^{(m)})\}$   $\triangleright$  Case RWR
15:    $\mathcal{U} \leftarrow \mathcal{U} \setminus x_{\mathcal{U}}^{(m)}$ 
16: return  $f$ 

```

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

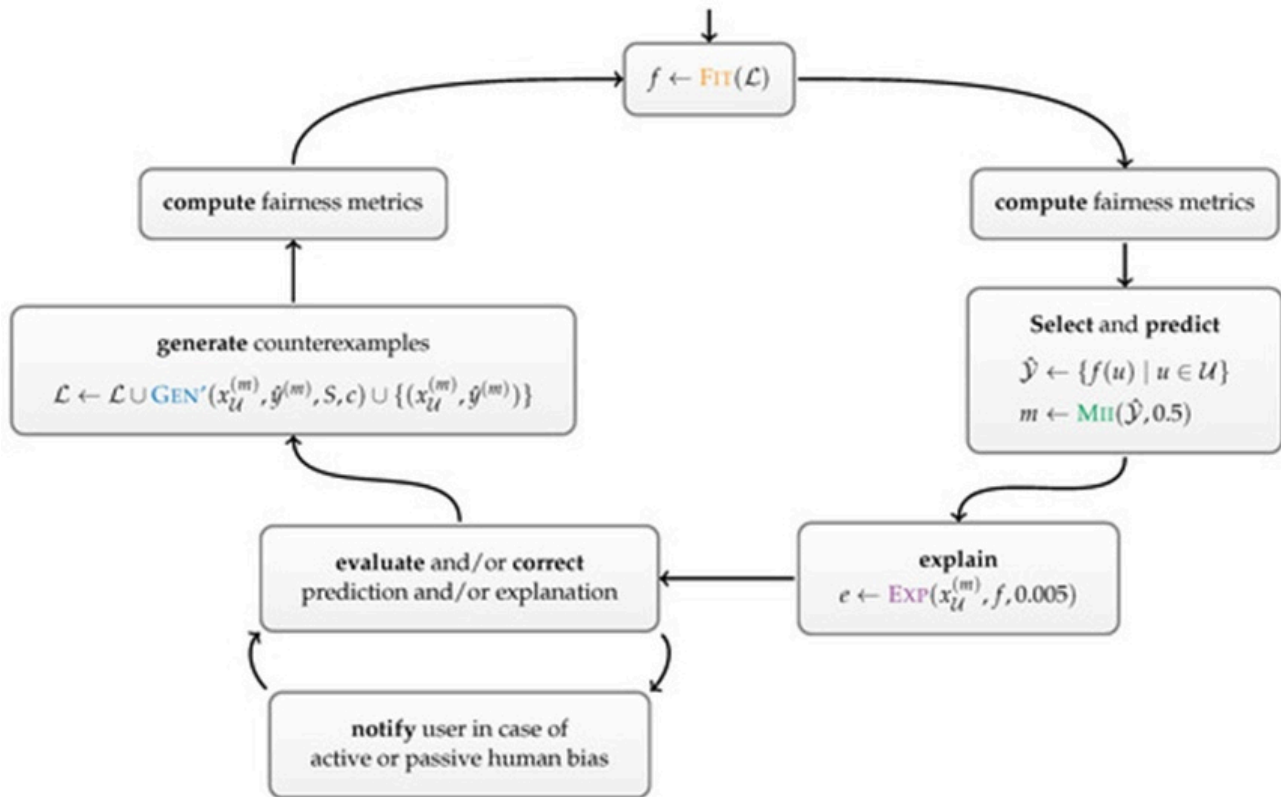


Figure 4. Human FairCaipi interaction. FairCaipi requests user input concerning both the prediction and explanation generated by a classification model. Displaying bias-detection metrics at the start and completion of each iteration allows operators to connect their modifications to the model's overall bias. Additionally, Caipi alerts users if performed or overlooked adjustments negatively impact these bias-detection metrics. When managed by users operating with equitable intentions, FairCaipi enhances the fairness of classification models.

Our methodology concentrates on enhancing model performance while minimizing the required volume of queries, interactions, and cumulative expenses. As stated in [14], active learning achieves a notably optimized sample complexity relative to passive learning owing to the deliberate choice of instances selected for labeling. Concerning its computational demands, our framework is affected by its model-agnostic architecture, thereby inheriting the complexities linked to training a model and executing predictions for particular instances using the underlying machine learning algorithm.

The primary elements of our framework, including FairCaipi and its concurrent operations—such as evaluating fairness metrics and determining SHAP values—retain a manageable level of complexity. Notably, while the calculation of SHAP values within the Exp procedure is generally intractable, it can be accurately approximated within polynomial time [50, 53]. The routines MII and GEN' designated for counterexample synthesis, along with the utilized set operations, possess a complexity that is at most polynomial. Consequently, we maintain that FairCaipi represents a computationally viable XIML framework.

Simulation study

We evaluate FairCaipi through a simulation experiment designed to minimize gender-based discrimination within the German Credit risk dataset (the codebase to replicate this simulation study and run FairCaipi interactively following the design in Figure 4 is hosted at: <https://github.com/emanuelsla/faircaipi>, accessed on 11 August 2023). Consequently, the gender variable serves as the sole protected attribute S . Furthermore, we maintain access to the true labels of the unlabeled instances via $\mathcal{Y} \setminus \mathcal{U} = \mathcal{I} \setminus \mathcal{U}$ and assess our explanations in line with Definition 6 utilizing an attribution threshold of 0.005. We bench-mark FairCaipi against Reweighting [10], a standard pre-processing

method that reallocates weights to training instances to satisfy Statistical Parity (1). Within the German Credit Risk dataset, we allocate 550 instances as labeled data and 150 instances as unlabeled data. The remaining 300 instances are reserved as test data. We train a Random Forest classifier employing balanced class weights. Initially, we obtain an accuracy of 75%. Establishing a relatively high baseline performance is critical for FairCaipi, given that its primary goal is to enhance the equity of decision-making frameworks. Prior research demonstrates that a smaller quantity of initially labeled instances is sufficient to achieve high predictive accuracy following Caipi optimization [16]. We execute FairCaipi for 100 iterations. During RWR iterations, we introduce a single counterexample consisting of the identical instance but with the alternative gender, thereby neutralizing the instance's gender effect.

Results and Discussion

Throughout the 100 FairCaipi iterations illustrated in [Figure 5](#), we detect a marginal trend toward a higher frequency of label adjustments relative to explanation adjustments. This indicates a higher proportion of type W iterations than type RWR iterations. Within the FairCaipi framework, roughly 35 out of 100 iterations involve modifications exclusively to the label, thereby affecting solely the predictive accuracy. Conversely, within 30 iterations, the SHAP explanation undergoes correction. These represent the specific iterations where the simulated operator attempts to adjust the model's internal framework to reduce the classifier's intrinsic bias. Finally, in approximately 35 FairCaipi iterations, the prediction is accurate and accompanied by an unbiased decision-making process, representing RRR instances.

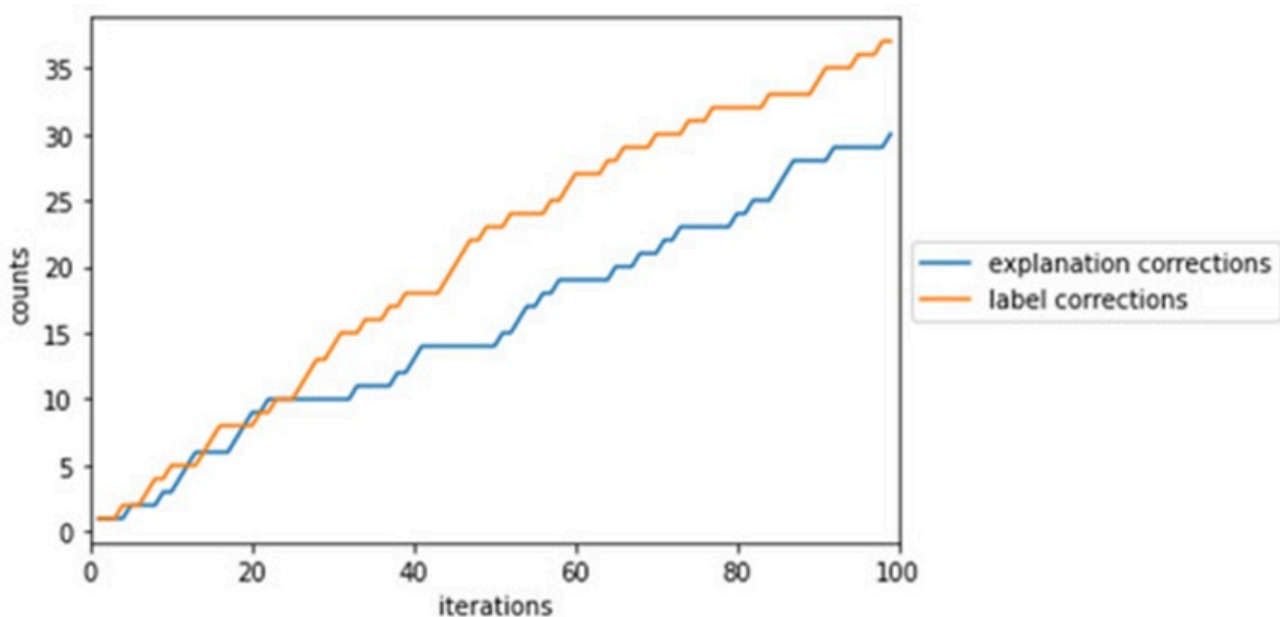


Figure 5. Number of label and explanation corrections within 100 FairCaipi iterations. The chart reveals a distinct interval where classification labels are primarily adjusted, occurring approximately between steps 20 and 30. Outside of this specific window, both forms of refinement are evenly dispersed throughout the FairCaipi optimization pipeline.

The outcomes regarding the bias-mitigation capabilities of FairCaipi are presented in [Table 2](#). The table contrasts multiple bias-detection metrics following 100 iterations of FairCaipi optimization, their optimal values recorded across the 100 iterations, the state-of-the-art sampling-based pre-processing method Reweighting, and the baseline Random Forest classifier operating without bias-mitigation adjustments. Every bias-detection metric achieves its optimal state at zero. We observe that FairCaipi delivers superior results across all bias-detection metrics, with the sole exception of Statistical Parity. This pattern is evident in the FairCaipi outcomes after 100 iterations and becomes even more pronounced when considering its optimal values. Reweighting, which modifies instance weights specifically to satisfy Statistical Parity,

noticeably outperforms the alternative approaches on this specific metric but yields negligible improvements for the remaining bias-detection metrics when compared to the baseline 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.

Fairness Metric	Reweighing Method	FAIRCAIPI (Optimized)	FAIRCAIPI	Baseline Model
Statistical Parity	0.0000	−0.0391	−0.0447	−0.0886
Equalized Odds	−0.1819	−0.0038	−0.0909	−0.1568
Equalized Opportunity	−0.1237	−0.0007	−0.0295	−0.0514
False Positive Error Rate Balance (FPERB)	−0.2401	0.0038	−0.1524	−0.2622
Predictive Parity	−0.0053	0.0008	−0.0026	−0.0322

The sequential progress of the monitored fairness indicators during the 100 FairCaipi optimization steps is charted in [Figure 6](#), where a value of zero denotes the ideal state for all indices. Each metric displays an overall inclination to move toward this ideal target, generally reaching its maximum effectiveness during its most optimal step. Most of these indicators exhibit minimal volatility and remain positioned near the ideal baseline throughout the entire runtime. Prominent departures from this pattern occur only within the Equalized Odds and False Positive Error Rate Balance dimensions. While the former shows a progressive and continuous trajectory toward its ideal value, the latter demonstrates wider fluctuations and begins to wander away from the optimal target after step 80.

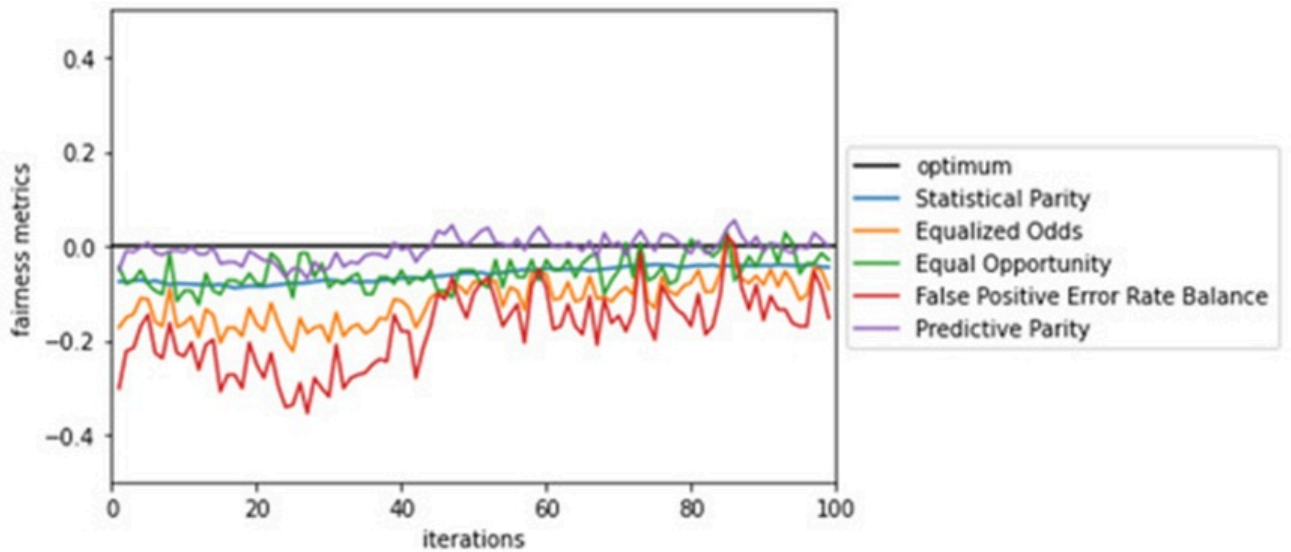


Figure 6. Bias-detection metrics during FairCaipi optimization. We compare the development of bias-detection metrics across 100 FairCaipi iterations. All metrics have their optimum at zero.

This constructive shift toward reduced algorithmic bias is further confirmed by **Figure 7**, which illustrates a steady decline in the volume of discriminatory classifications and biased explanations as the FairCaipi pipeline advances. The data plotted in **Figure 7** uncovers a few noteworthy phenomena: owing to the relative scales of the training and validation splits, biased decisions and explanations surface far more frequently within the training pool than in the test cohort. Furthermore, when an explanation is flagged as biased—indicating that the model is processing attributes unfairly—the corresponding classification is automatically deemed unfair as well. This dynamic does not work in reverse; a data point can easily belong to a marginalized demographic and receive an unfavorable decision for legitimate reasons completely divorced from the protected trait. For this reason, unfair predictions are consistently more numerous than unfair explanations.

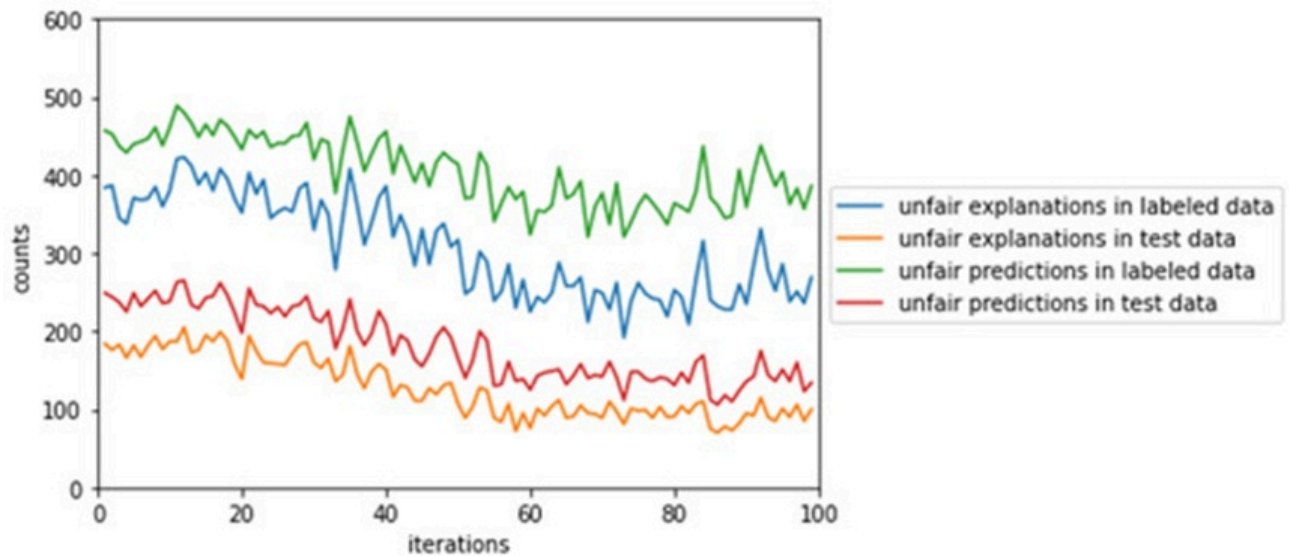


Figure 7. Unfair predictions and explanations during FairCaipi optimization. In each iteration, we calculate the number of unfair predictions and explanations on labeled and test data.

Efforts to suppress algorithmic bias must not come at the expense of the model's predictive accuracy. To monitor this trade-off, **Table 3** aggregates several standard metrics for binary evaluation, contrasting the accuracy, precision, recall, and F1-score of FairCaipi against both a baseline Random Forest and a Random Forest variant adjusted via Reweighting. FairCaipi achieves the top performance across every single metric. Even so, the data reveals a stark divergence when these metrics are split based on the target class. Although all evaluated models show strong capabilities when processing low-risk credit profiles, their overall performance drops noticeably when dealing with high-risk profiles. While the sequential test accuracy charted across the FairCaipi steps in **Figure 8** corroborates this analysis, the highly skewed nature of the underlying data distribution (originally shown in **Figure 2**) means that the tabular metrics in **Table 3** remain the primary authority. The timeline in **Figure 8** utilizes accuracy merely as a baseline indicator to observe performance shifts as the FairCaipi optimization unfolds. The resulting fluctuations are remarkably minor, which is the ideal outcome; FairCaipi is engineered specifically to dismantle bias within an already established model architecture rather than to serve as an optimization tool for raw predictive metrics.

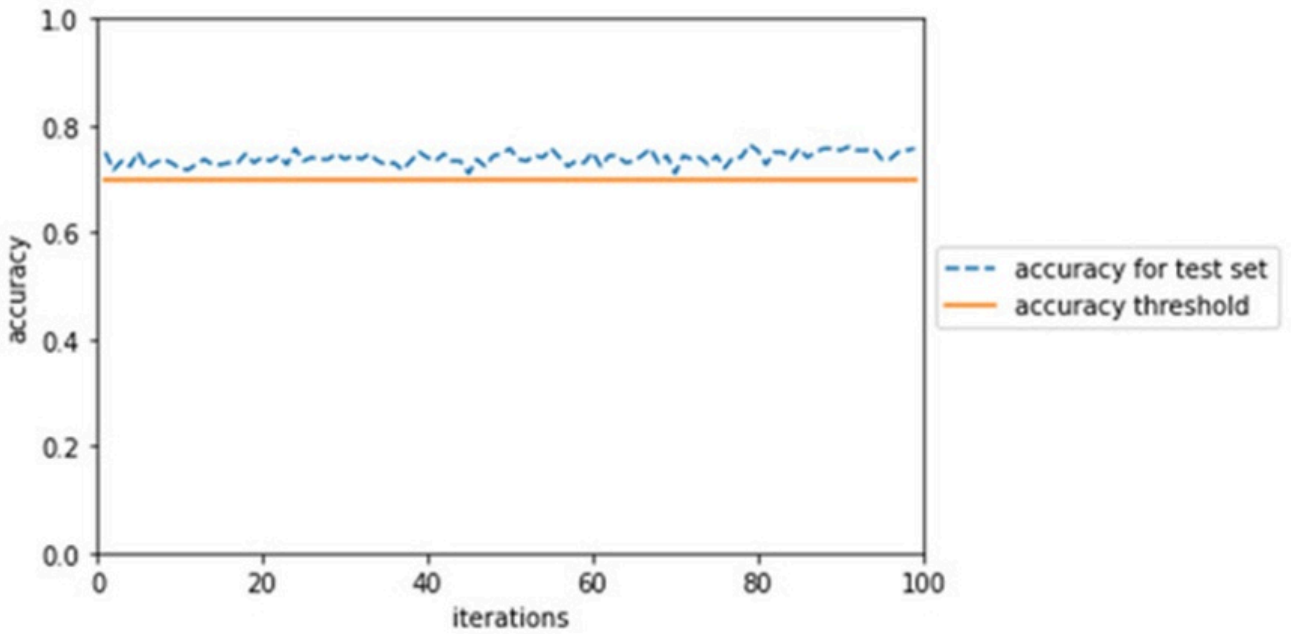


Figure 8. Predictive quality of FairCaipi during 100 iterations. Starting with a baseline accuracy of 75%, we calculate the test accuracy in each iteration. We add a 70% accuracy threshold mark.

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.

Performance Measure	Category	Reweighing Approach	FAIRCAIPI	Baseline Model
Accuracy	Overall	0.72	0.76	0.73
Precision	Good-risk class	0.75	0.77	0.76
	Bad-risk class	0.58	0.68	0.60
Recall	Good-risk class	0.89	0.92	0.89
	Bad-risk class	0.34	0.39	0.38
F1-Score	Good-risk class	0.82	0.84	0.82
	Bad-risk class	0.43	0.50	0.47

To wrap up this analysis, we address the core inquiries of this investigation:

- **(R1) Does modifying model explanations for fairness generate more equitable models?**

Indeed. The data plotted in [Figure 6](#) shows that fairness metrics steadily move toward their ideal baselines as explanation adjustments are introduced. Furthermore, discriminatory classifications consistently diminish over the 100 FairCaipi steps ([Figure 7](#)), driven by explanation adjustments executed during 30 of those steps ([Figure 5](#)).

- **(R2) Does modifying explanations for fairness yield more equitable explanations?**

Yes. **Figure 7** confirms a distinct downward trend in the volume of biased explanations when utilizing the FairCaipi framework.

- **(R3) Does the pursuit of equitable explanations degrade the predictive accuracy of the underlying model?**

No. **Figure 8** clearly shows that the model maintains highly stable performance with only minor variations throughout the FairCaipi routine. Furthermore, the data in **Table 3** reveals that FairCaipi actually provides a slight lift in overall predictive metrics compared to a standard Random Forest model, regardless of whether that baseline utilizes Reweighting.

- **(R4) Which approach performs better: FairCaipi or the standard Reweighting method?**

Reviewing the results in **Table 2**, FairCaipi proves to be the superior choice for mitigating bias across every evaluated metric except for Statistical Parity, which happens to be the singular mathematical objective that the Reweighting technique is explicitly engineered to maximize.

The empirical outcomes of this study indicate that FairCaipi functions as a highly competent strategy for bias mitigation. It successfully satisfies two critical requirements: first, it avoids undermining the predictive capability of the system—on the contrary, our assessments reveal a marginal improvement in overall predictive accuracy. Second, it yields effective bias reduction across multiple concurrent evaluation metrics, eventually outperforming a prominent state-of-the-art data pre-processing technique designed for fairness. Furthermore, in alignment with the core tenets of XIML, FairCaipi establishes an open and transparent decision-making architecture that directly integrates human oversight into the fairness pipeline. Given potential statutory mandates, this inclusion offers an undisputed ethical advantage, particularly since human operators do not view equity as a static or unchanging concept [11]. Our findings demonstrate that modifying a classification model's decision framework via iterative human input facilitates a comprehensive and well-rounded reduction in bias across diverse performance indices. Relying strictly on a single, fixed metric for optimization cannot guarantee widespread improvements in fairness.

We argue that FairCaipi possesses the capability to (i) identify and (ii) reduce machine-driven bias, while also (iii) uncovering human-driven bias. Although the first two functionalities have been verified through our experiments, we present a formalized framework for the third, which demands deeper empirical validation. Specialized user studies represent a viable path forward in this domain. Within such a setting, researchers could explore the specific question: *Is FairCaipi capable of diminishing human-driven bias?*

Generally, the current experimental framework faces certain constraints; for instance, we provide only a foundational proof of concept focused on a single dataset and a solitary form of discrimination. We do not explore how FairCaipi handles environments containing multiple protected variables—which represents a more frequent empirical scenario—nor do we evaluate its efficacy within multi-label classification or regression tasks. Nevertheless, the architecture of our simulation study mirrors established validation protocols used for leading bias-reduction strategies across the pre-processing [10, 21], in-processing [8], and post-processing [24] operational phases. A unifying element across all these cited investigations is their focus on mitigating gender-based discrimination, with a subset of these authors similarly utilizing the German Credit Risk dataset [8, 10, 21].

From an algorithmic perspective, FairCaipi exhibits certain limitations: even though operators receive alerts when their selections degrade fairness indices, the system operates under the assumption that users possess the necessary domain expertise to offer optimal guidance. However, a critical problem arises if discriminatory explanations become less apparent than they are in our baseline scenario—such as when complex, multivariate feature dependencies replicate the bias indirectly. Under those conditions, our expectations regarding user capability are likely overextended.

Additionally, in comparison to traditional Caipi implementations [14], we lack empirical data demonstrating how incremental quantities of counterexamples impact overall fairness. Applying our current, straightforward counterexample production routine would merely result in duplicate instances of the exact same counterexample. To resolve this, more advanced data generation pipelines leveraging statistical bootstrapping or generative modeling frameworks could be integrated.

To contextualize these insights within the broader literature: FairCaipi serves as an in-processing strategy for bias reduction, occupying a distinct position among XIML architectures focused on algorithmic fairness. Based on our review of the field, FairCaipi shares a close thematic relationship with a fairness-seeking framework that permits operators to interact directly with model explanations [44]. However, the fundamental distinction lies in FairCaipi's emphasis on a collaborative human-machine relationship where both entities benefit—the system minimizes algorithmic discrimination, while the user's personal biases are exposed. Existing XIML fairness frameworks rarely incorporate human operators to this degree. Conversely, those alternative systems might prove more practical for real-world deployments, as maintaining human involvement over 100 consecutive iterations can be exceptionally time-intensive. Nonetheless, FairCaipi broadens the operational spectrum of XIML strategies and fills a specific procedural requirement. Our experiments confirm that continuous human interaction serves as a productive engine for bias reduction and may offer a morally superior alternative in fairness-critical domains.

Conclusion

FairCaipi is an in-processing algorithmic framework for bias mitigation rooted in human-centered XIML. Continual human feedback is leveraged to guide the model's underlying decision-making architecture toward an equitable trajectory—specifically, an environment marked by a reduction in discriminatory classifications. Empirical testing demonstrates that formal fairness indicators show steady improvement throughout the FairCaipi optimization loop, while the predictive performance of the primary classification model remains stable. The system is fully capable of identifying and reducing machine-driven bias, and it can similarly expose human-driven bias.

In future research, we aim to expand the adaptability of our framework: we intend to address algorithmic bias across a broader range of classification environments and factor in multiple protected attributes simultaneously. This broader scope will encompass, for example, multi-label classification environments. This expansion will necessitate revisions to the core FairCaipi cycle, given that a single predictive instance can simultaneously correspond to multiple outcome classifications. In this context, we will investigate two primary research questions: *What form does Caipi user feedback take within a multi-label classification environment?* and *How can operator input in multi-label classification settings be successfully converted into functional counterexamples?*

In the most straightforward scenario, all existing protected attributes would undergo simultaneous randomization during the counterexample production phase. However, this introduces an important sub-question: *Must we systematically disentangle every single correlation among protected attributes, or is it sufficient to alter merely a subset of them to mitigate the classification model's overall bias?*

Furthermore, our subsequent efforts will prioritize the detection of multivariate correlations to uncover hidden bias. We anticipate that FairCaipi may struggle to resolve discriminatory decision frameworks when dealing with indirect bias, primarily because protected attributes might not surface directly within local feature attributions. We argue that methodology borrowed from causal statistics offers a promising remedy. Rather than displaying raw feature attribution values to users via frameworks like SHAP, we intend to map out the underlying dependencies using structures such as Bayesian networks. This approach will provide operators with a clearer understanding of how distinct features interact and influence one another, prompting the question: *How can user modifications to Bayesian networks be programmatically transformed into counterexamples?*

Ultimately, one of our primary long-term objectives is to instruct and inform users, even when systemic bias in decision-making is deeply hidden. This aspect of user instruction must be rigorously evaluated through targeted user experience studies. To facilitate this, we will engineer specialized user interfaces for FairCaipi, focusing heavily on clarity and operational simplicity as vital areas for future academic inquiry.

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

Acknowledgements

None

Conflict of interest

None

Financial support

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).

Ethics statement

None

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Angwin J, Larson J, Mattu S, Kirchner L. Machine bias. ProPublica. 2016 May 23.
- Wolf MJ, Miller KW, Grodzinsky FS. Why we should have seen that coming: comments on Microsoft's Tay "experiment," and wider implications. *ACM SIGCAS Comput Soc*. 2017;47(3):54-64.
<https://doi.org/10.1145/3144592.3144598>.
- Zhang L, Yencha C. Examining perceptions towards hiring algorithms. *Technol Soc*. 2022;68:101848.
<https://doi.org/10.1016/j.techsoc.2021.101848>.
- Mukerjee A, Biswas R, Deb K, Mathur AP. Multi-objective evolutionary algorithms for the risk-return trade-off in bank loan management. *Int Trans Oper Res*. 2002;9(5):583-97.
<https://doi.org/10.1111/1475-3995.00375>.
- Goel N, Yaghini M, Faltings B. Non-discriminatory machine learning through convex fairness criteria. *Proc AAAI Conf Artif Intell*. 2018;32(1):3029-36.
<https://doi.org/10.1609/aaai.v32i1.11573>.
- Menon AK, Williamson RC. The cost of fairness in binary classification. *Proc Mach Learn Res*. 2018;81:107-18.
- Kamishima T, Akaho S, Asoh H, Sakuma J. Fairness-aware classifier with prejudice remover regularizer. In: *Machine Learning and Knowledge Discovery in Databases*. 2012;7524:35-50.
https://doi.org/10.1007/978-3-642-33486-3_3.
- Celis LE, Huang L, Keswani V, Vishnoi NK. Classification with fairness constraints: a meta-algorithm with provable guarantees. *Proc Conf Fairness Account Transpar*. 2019:319-28.
<https://doi.org/10.1145/3287560.3287586>.
- Kearns M, Neel S, Roth A, Wu ZS. Preventing fairness gerrymandering: auditing and learning for subgroup fairness. *Proc Mach Learn Res*. 2018;80:2564-72.
- Kamiran F, Calders T. Data preprocessing techniques for classification without discrimination. *Knowl Inf Syst*. 2012;33(1):1-33.
<https://doi.org/10.1007/s10115-011-0463-8>.
- Berman JJ, Murphy-Berman V, Singh P. Cross-cultural similarities and differences in perceptions of fairness. *J Cross Cult Psychol*. 1985;16(1):55-67.
<https://doi.org/10.1177/0022002185016001005>.
- Settles B. *Active learning*. San Rafael (CA): Morgan & Claypool Publishers; 2012.
<https://doi.org/10.2200/S00429ED1V01Y201207AIM018>.
- Shivaswamy P, Joachims T. Coactive learning. *J Artif Intell Res*. 2015;53:1-40.
- Teso S, Kersting K. Explanatory interactive machine learning. *Proc AAAI/ACM Conf AI Ethics Soc*. 2019:239-45.
<https://doi.org/10.1145/3306618.3314293>.

Angerschmid A, Zhou J, Theuermann K, Chen F, Holzinger A. Fairness and explanation in AI-informed decision making. *Mach Learn Knowl Extr.* 2022;4(2):556-79.
<https://doi.org/10.3390/make4020026>.

Slany E, Ott Y, Scheele S, Paulus J, Schmid U. CAIPI in practice: towards explainable interactive medical image classification. In: *Artificial Intelligence Applications and Innovations*. 2022;652:389-400.
https://doi.org/10.1007/978-3-031-08341-9_31.

Schramowski P, Stammer W, Teso S, Brugger A, Herbert F, Shao X, et al. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nat Mach Intell.* 2020;2:476-86.
<https://doi.org/10.1038/s42256-020-0212-3>.

Bellamy RKE, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, et al. AI fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. *IBM J Res Dev.* 2019;63(4/5):4:1-4:15.
<https://doi.org/10.1147/JRD.2019.2942287>.

Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model cards for model reporting. *Proc Conf Fairness Account Transpar.* 2019:220-9.
<https://doi.org/10.1145/3287560.3287596>.

Kärkkäinen K, Joo J. FairFace: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021:1548-58.

Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S. Certifying and removing disparate impact. In: *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015:259-68.
<https://doi.org/10.1145/2783258.2783311>.

Calmon FP, Wei D, Vinzamuri B, Ramamurthy KN, Varshney KR. Optimized pre-processing for discrimination prevention. In: *Advances in Neural Information Processing Systems* 30. 2017:3992-4001.

Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. In: *Advances in Neural Information Processing Systems* 29. 2016:3315-23.

Pleiss G, Raghavan M, Wu F, Kleinberg JM, Weinberger KQ. On fairness and calibration. In: *Advances in Neural Information Processing Systems* 30. 2017:5680-9.

Kamiran F, Karim A, Zhang X. Decision theory for discrimination-aware classification. In: *Proceedings of the 12th IEEE International Conference on Data Mining*. 2012:924-9.
<https://doi.org/10.1109/ICDM.2012.45>.

Agarwal A, Beygelzimer A, Dudík M, Langford J, Wallach HM. A reductions approach to fair classification. *Proc Mach Learn Res.* 2018;80:60-9.

Agarwal A, Dudík M, Wu ZS. Fair regression: quantitative definitions and reduction-based algorithms. *Proc Mach Learn Res.* 2019;97:120-9.

Bolukbasi T, Chang K, Zou JY, Saligrama V, Kalai AT. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: *Advances in Neural Information Processing Systems* 29. 2016:4349-57.

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*. 2019:1:629-34.
<https://doi.org/10.18653/v1/N19-1064>.

Xu D, Yuan S, Zhang L, Wu X. FairGAN: fairness-aware generative adversarial networks. In: Proceedings of the IEEE International Conference on Big Data. 2018:570-5.
<https://doi.org/10.1109/BigData.2018.8622525>.

Louizos C, Swersky K, Li Y, Welling M, Zemel RS. The variational fair autoencoder. In: Proceedings of the 4th International Conference on Learning Representations. 2016.

Wang Z, Qinami K, Karakozis IC, Genova K, Nair P, Hata K, et al. Towards fairness in visual recognition: effective strategies for bias mitigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:8916-25.
<https://doi.org/10.1109/CVPR42600.2020.00894>.

Gong S, Liu X, Jain AK. Mitigating face recognition bias via group adaptive classifier. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021:3414-24.
<https://doi.org/10.1109/CVPR46437.2021.00342>.

Nabi R, Shpitser I. Fair inference on outcomes. *Proc AAAI Conf Artif Intell*. 2018;32(1):1931-40.
<https://doi.org/10.1609/aaai.v32i1.11553>.

Loftus JR, Russell C, Kusner MJ, Silva R. Causal reasoning for algorithmic fairness. *arXiv*. 2018;arXiv:1805.05859.
<https://doi.org/10.48550/arXiv.1805.05859>.

Kilbertus N, Rojas-Carulla M, Parascandolo G, Hardt M, Janzing D, Schölkopf B. Avoiding discrimination through causal reasoning. In: *Advances in Neural Information Processing Systems* 30. 2017:656-66.

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. 2017:3929-35.
<https://doi.org/10.24963/ijcai.2017/549>.

Zhang L, Wu Y, Wu X. Causal modeling-based discrimination discovery and removal: criteria, bounds, and algorithms. *IEEE Trans Knowl Data Eng*. 2019;31(11):2035-50.
<https://doi.org/10.1109/TKDE.2018.2872988>.

Sharma S, Henderson J, Ghosh J. CERTIFAI: counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models. In: Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency. 2020:285-94.
<https://doi.org/10.1145/3375627.3375812>.

Zhang J, Bareinboim E. Fairness in decision-making—the causal explanation formula. *Proc AAAI Conf Artif Intell*. 2018;32(1):2037-45.
<https://doi.org/10.1609/aaai.v32i1.11564>.

Begley T, Schwedes T, Frye C, Feige I. Explainability for fair machine learning. *arXiv*. 2020;arXiv:2010.07389.

Cabrera AA, Epperson W, Hohman F, Kahng M, Morgenstern J, Chau DH. FAIRVIS: visual analytics for discovering intersectional bias in machine learning. In: Proceedings of the IEEE Conference on Visual Analytics Science and Technology. 2019:46-56.
<https://doi.org/10.1109/VAST47406.2019.8986948>.

Ahn Y, Lin Y. FairSight: visual analytics for fairness in decision making. *IEEE Trans Vis Comput Graph*. 2020;26(1):1086-95.
<https://doi.org/10.1109/TVCG.2019.2934262>.

Nakao Y, Stumpf S, Ahmed S, Naseer A, Strappelli L. Towards involving end-users in interactive human-in-the-loop AI fairness. *ACM Trans Interact Intell Syst*. 2022;12(3):18.
<https://doi.org/10.1145/3514258>.

Kulesza T, Burnett MM, Wong W, Stumpf S. Principles of explanatory debugging to personalize interactive machine learning. In: Proceedings of the 20th International Conference on Intelligent User Interfaces. 2015:126-37.
<https://doi.org/10.1145/2678025.2701399>.

Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *ACM Comput Surv*. 2021;54(6):115:1-115:35.
<https://doi.org/10.1145/3457607>.

Chen J, Kallus N, Mao X, Svacha G, Udell M. Fairness under unawareness: assessing disparity when protected class is unobserved. *Proc Conf Fairness Account Transpar*. 2019:339-48.
<https://doi.org/10.1145/3287560.3287594>.

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel RS. Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. 2012:214-26.
<https://doi.org/10.1145/2090236.2090255>.

Chouldechova A. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data*. 2017;5(2):153-63.
<https://doi.org/10.1089/big.2016.0047>.

Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems* 30. 2017:4765-74.

Schallner L, Rabold J, Scholz O, Schmid U. Effect of superpixel aggregation on explanations in LIME—a case study with biological data. In: *Machine Learning and Knowledge Discovery in Databases International Workshops of ECML PKDD 2019*. 2019;1167:147-58.
https://doi.org/10.1007/978-3-030-43823-4_13.

Young HP. Monotonic solutions of cooperative games. *Int J Game Theory*. 1985;14(2):65-72.
<https://doi.org/10.1007/BF01769885>.

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(63):1-58.