Auditor Fairness Evaluation via Learning Latent Assessment Models from Elicited Human Feedback - Rebuttal Draft

Anonymous authors
Paper under double-blind review

Abstract

Algorithmic fairness literature presents numerous mathematical notions and metrics, and also points to a tradeoff between them while satisficing some/all of them simultaneously. Furthermore, the contextual nature of fairness notions makes it difficult to automate bias evaluation in diverse algorithmic systems. Therefore, in this paper, we propose a novel model called latent assessment model (LAM) to characterize binary feedback provided by human auditors, by assuming that the auditor compares the classifier's output to his/her own intrinsic judgment for each input. We prove that individual and/or group fairness notions are guaranteed as long as the auditor's intrinsic judgments inherently satisfy the fairness notion at hand, and are relatively similar to the classifier's evaluations. We also demonstrate this relationship between LAM and traditional fairness notions on three well-known datasets, namely COMPAS, German credit and Adult Census Income datasets. Furthermore, we also derive the minimum number of feedback samples needed to obtain probably approximately correct (PAC) learning guarantees to estimate LAM for black-box classifiers. Moreover, we propose a novel multi-attribute reputation measure to evaluate auditor's preference towards various fairness notions as well as sensitive groups. These guarantees are also validated using standard machine learning algorithms, which are trained on real binary feedback elicited from 400 human auditors regarding COMPAS.

1 Introduction

Several machine learning (ML) based systems have recently been reported as being discriminatory with respect to the sensitive attributes (e.g. race and gender) in a variety of application domains, such as recommender systems in criminal justice (Angwin et al., 2016), e-commerce services in online markets (Fisman & Luca, 2016), and life insurance premiums (Waxman, 2018). Most of these applications have outcomes that depend on subjective human judgements which vary drastically across different people. Typically, these outcome labels are obtained on a case-by-case basis from different types of individuals (typically subject matter experts who have rich experience in their respective fields.). However, experts also exhibit biases especially when it comes to subjective judgements. For instance, in criminal justice applications, one may argue that a 'just' outcome for a parole application by a Black applicant is to deny the bail since he/she has a high chance to recidivate back to crime since the applicant lives in an unsafe neighborhood. On the other hand, the same applicant could be presented with a bail if he/she has no prior counts and has a clean record in the past. This idea of justice where everyone should be treated fairly based on their own features/merit, and independent of how others are treated, is called *non-comparative fairness* (Feinberg, 1974). In a practical data collection process, outcome labels are elicited from a non-comparative fairness standpoint since each person is presented with one case at a time (for example, readers may refer to (Goel & Faltings, 2019)). However, these judgements are usually subjective and even experts may not always arrive at a consensus on one correct verdict. Since the labelers suffer from inherent biases, their labels are often subject to inquisition (Cooper et al., 2021). This dearth of correct labels invalidates the quantification of ML performance using *accuracy*, which measures the alignment of classifier's outcome with the correct label found in training data.

On the contrary, algorithmic fairness literature do not question the fidelity of training data and discusses the quantification of biases and tradeoffs with accuracy using *comparative fairness* notions (Mehrabi et al., 2021).

Comparative fairness notions primarily focus on issues such as equality, similarity and proportionality, which questions whether an individual is treated/serviced 'justly' in comparison to that received by others. Biases measured by comparative notions can be broadly classified into two types, namely *disparate treatment*, where different groups of people are treated differently; and *disparate impact*, where different groups of people obtain different proportions of outcomes. Individual fairness (Dwork et al., 2012) quantifies disparate treatment via comparing the similarity between two individuals based on a task-specific distance metric. On the other hand, group fairness notions (Caton & Haas, 2020; Chouldechova & Roth, 2018; Hardt et al., 2016b) quantify disparate impact by comparing the differences in outcome proportions (e.g. true positive, false positive rates, predictive positive value) across different sensitive groups. In addition to measuring comparative fairness in ML systems, various mitigation techniques have been proposed to alleviate these biases in the literature. These mechanisms are typically categorized into three types: (i) *pre-processing* techniques that modifies the input to the ML model (Kamiran & Calders, 2012), (ii) *in-processing* techniques that retrain the ML model (Kamishima et al., 2012; Zafar et al., 2017), and (iii) *post-processing* techniques that modify the output labels (Hardt et al., 2016b; Corbett-Davies et al., 2017). Each of these approaches either maximize accuracy (or some loss function) under prescribed comparative fairness constraints, or maximize comparative fairness under tolerable constraints on accuracy loss. Finally, a fundamental tradeoff between comparative fairness and accuracy has also been investigated based on which, a compromise in accuracy was proposed to obtain fairness improvements in the system. However, comparative fairness notions cannot be completely automated and demands human feedback to identify appropriate fairness notions for system evaluation. Heterogeneous stakeholders compete against each other to enforce their preferred fairness notions, all of which cannot be satiated simultaneously due to a fundamental tradeoff that exists between comparative fairness notions (Chouldechova, 2017; Kleinberg et al., 2016). This dogfight between various stakeholders regarding the selection of an appropriate comparative fairness metric is also context-dependent (Binns, 2018), due to differences in protected groups across applications. Therefore, several researchers have attempted to explore human perception of fairness to identify the appropriate comparative fairness notion. For more details, the reader may refer to Section 1.2.

The inability of comparative notions to enable a reliable algorithmic fairness analysis, combined with the desertion of non-comparative fairness perspective in the past, leads to a major predicament in the design of fair ML algorithms. To the best of our knowledge, this paper makes the first attempt to address this gap via modeling human auditors from a non-comparative fairness perspective, and develop a scalable method to evaluate high-dimensional ML algorithms. Furthermore, this paper also assumes that the auditor provides a binary feedback regarding system fairness since the auditor's intrinsic evaluation of labels is a subconscious phenomenon and is practically difficult to quantify in most applications.

1.1 Technical Contributions

The main contributions of the paper are four-fold. Firstly, a novel *latent assessment model* (LAM) was proposed to represent binary human feedback regarding the fairness of the system, from a noncomparative perspective. Unlike most of the past literature on human perception of fairness, the proposed model allows human auditors to provide their feedback at will, without enforcing any specific fairness notion artificially (Gillen et al., 2018). Secondly, we emphasize that the availability of non-comparative fairness does not necessarily dispense with the need of comparative fairness. We prove that a system/entity satisfies individual and/or group fairness notions if the auditor exhibits LAM in his/her fairness evaluations. We also show that converse holds true in the case of individual fairness. Since both the system and auditor rules are hidden, we compute PAC learning guarantees on algorithmic auditing based on binary feedback obtained from human auditors. Third, we introduce a novel multi-attribute reputation measure to evaluate auditor's inherent biases based on various comparative fairness notions as well as sensitive attributes. Lastly, we validate the relationships with comparative fairness notions on three real datasets, namely COMPAS, Adult Income Census and German credit datasets. Using the feedback data of 400 crowd workers collected by (Dressel & Farid, 2018), we compare various learning frameworks such as logistic regression, support vector machines (SVM) and decision trees to estimate auditor's intrinsic judgements and their feedback. We also measure the reputation of the crowd workers to evaluate auditor's preference towards various comparative fairness notions.

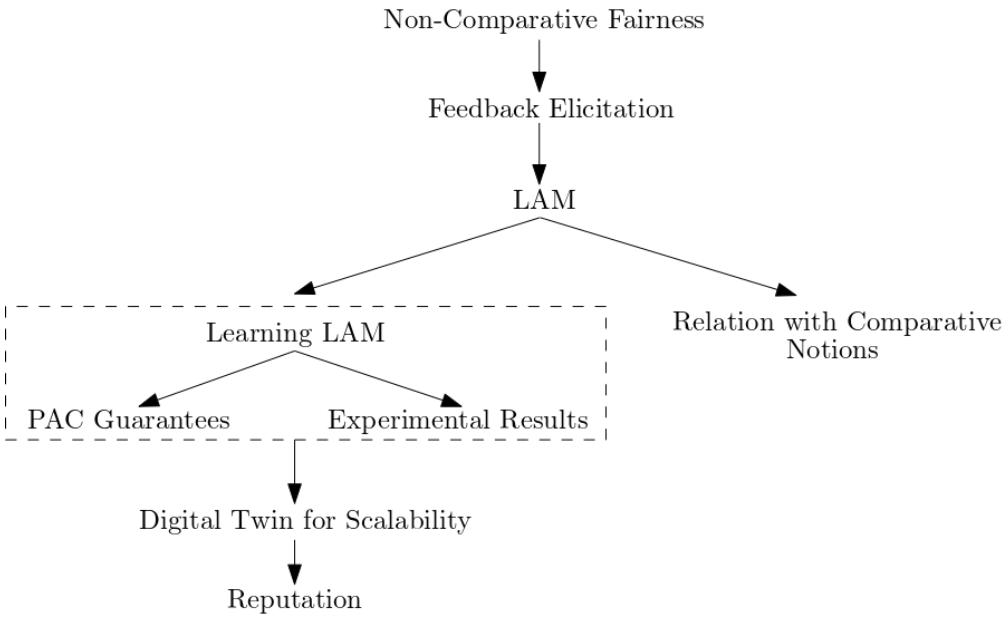


Figure 1: A Map of Technical Contributions

1.2 Related Work on Human Perception of Fairness

In the past, several researchers have attempted to model human perception of fairness, but have always tried to fit their revealed feedback to one of the existing fairness notions. For instance, task-based similarity metric used in individual fairness notion (Dwork et al., 2012) were estimated by (Jung et al., 2019) based on feedback elicited from auditors regarding how a given pair of individuals have been treated. Similarly, the work of (Gillen et al., 2018) assumes the existence of an auditor who is capable of identifying unfairness given pair of individuals when the underlying similarity metric is Mahalanobis distance. Saxena investigated how people perceive the fitness of three different individual fairness notions in the context of loan decisions (Saxena et al., 2019). The three notions are: (1) treat similar individuals similarly (Dwork et al., 2012), (2) Never favor a worse individual over a better one (Joseph et al., 2016), and (3) the probability of approval should be equal to the reward being the greatest (Liu et al., 2017). They show that people exhibit a preference for the last fairness definition.

From the perspective of group fairness notions, an experiment performed by (Srivastava et al., 2019) asks participants to choose among two different models to identify which notion of fairness (demographic parity or equalized odds) best captures people's perception in the context of both risk assessment and medical applications. Likewise, another team surveyed 502 workers on Amazon's Mturk platform and observed a preference towards *equal opportunity* in (Harrison et al., 2020). Note that both papers asked participants to reveal their analysis concerning a specific fairness notion in the context of given sensitive attributes (e.g. race, gender) which was clearly pointed out as a limitation of their work. On the contrary, in this paper, we impose no such restrictions on the auditor in constructing their feedback with respect to satiating any specific fairness notion. Instead, we assume that the expert auditor employs an intrinsic fair decision rule (which is unknown) to evaluate a given data tuple. Dressel and Farid in (Dressel & Farid, 2018) showed that COMPAS is as accurate and fair as that of untrained human auditors regarding predicting recidivism scores. On the other hand, (Yaghini et al., 2021) proposed a novel fairness notion, equality of opportunity (EOP), which requires that the distribution of utility should be the same for individuals with similar desert. Based on eliciting human judgments, they learned the proposed EOP notion in terms of criminal risk assessment context. Results show that EOP performs better than existing notions of algorithmic fairness in terms of equalizing utility distribution across groups. Another interesting work is by (Grgic-Hlaca et al., 2018), who discovered that people's fairness concerns are typically multi-dimensional (relevance, reliability, and volitionality), especially when binary feedback was elicited. This means that modeling human feedback should consider several factors beyond social discrimination. A major drawback of these approaches is that

the demographics of the participants involved in the experiments (Yaghini et al., 2021; Grgic-Hlaca et al., 2018; Harrison et al., 2020; Saxena et al., 2019) are not evenly distributed. For instance, the conducted experiments ask how models treated Caucasians and African-Americans. However, non-Caucasian participants were insufficient to assess the relationship between the participant's own demographics and the group that was disadvantaged by the model. Moreover, the participants are presented with multiple questions in the existing literature which cannot be scaled for larger decision-based models (Yaghini et al., 2021).

2 Preliminaries: Traditional Fairness Notions

In most practical systems, two types of discrimination exist: *(i) disparate treatment*, where an individual is intentionally treated differently based on his/her membership in a protected class; and *(ii) disparate impact*, where members of a protected class are more negatively impacted than others. However, algorithmic fairness literature has studied a different set of fairness notions (ref. (Caton & Haas, 2020; Chouldechova & Roth, 2018; Mehrabi et al., 2021; Pessach & Shmueli, 2020)). Let $f(\cdot)$ be a predictor which predicts an outcome $\hat{y} = f(\boldsymbol{x})$ where $\boldsymbol{x}$ is the multi-attribute variable and y be the true label.

2.1 Group Fairness Notions

The notion of group fairness seek for parity of some statistical measure across all the protected attributes present in the data. Different versions of group-conditional metrics led to different group definitions of fairness. Let A be the set of protected attributes where, $a \in A$ is the privileged group and $a' \in A$ is the underprivileged group.

Statistical Parity: This measures seeks to compute the probability difference of individuals who are predicted to be positive across different sensitive groups. Formally, it can be defined as followed.

$$\mathbb{P}[\hat{y} = 1 \mid A = a] - \mathbb{P}[\hat{y} = 1 \mid A = a'] \leq \delta \tag{1}$$

Since statistical parity aims for equal proportions of positive outcomes, the ideal value of probability difference should be 0. Specifically, if the difference is greater than 0, privileged group is benefited. Whereas, if the difference is less than 0, the underprivileged group is benefited. A major disadvantage of statistical parity is when the base rates (ratio of actual positive outcomes) are significantly different for various groups.

Equal Opportunity: To overcome the drawbacks in statistical parity, (Hardt et al., 2016a) introduced the notion of equalized odds which computes the difference between the true positive rates (TPR) of two protected groups.

$$\mathbb{P}[\hat{y} = 1 \mid y = 1, A = a] - \mathbb{P}[\hat{y} = 1 \mid y = 1, A = a'] \leq \delta. \tag{2}$$

Smaller differences between groups indicate better fairness. Since this notion considers the true label y, it assumes that the base rates of the two groups are representative and were not obtained in a biased manner.

Calibration: The measures computed the difference between positive predictive value of two groups. Positive predictive value represents the probability of an individual with a positive prediction actually experiencing a positive outcome. This notion is mathematically formulated as follows.

$$\mathbb{P}[y = 1 \mid \hat{y} = 1, A = a] - \mathbb{P}[y = 1 \mid \hat{y} = 1, A = a'] \leq \delta. \tag{3}$$

Although in some cases equal calibration may be the desired measure, it has been shown that it is incompatible with equalized odds (Pleiss et al., 2017) and is insufficient to ensure accuracy (Corbett-Davies & Goel, 2018).

Equal Accuracy: This requires similar accuracy across groups (Berk et al., 2018).

$$\mathbb{P}[y = \hat{y} \mid A = a] - \mathbb{P}[y = \hat{y} \mid A = a'] \leq \delta. \tag{4}$$

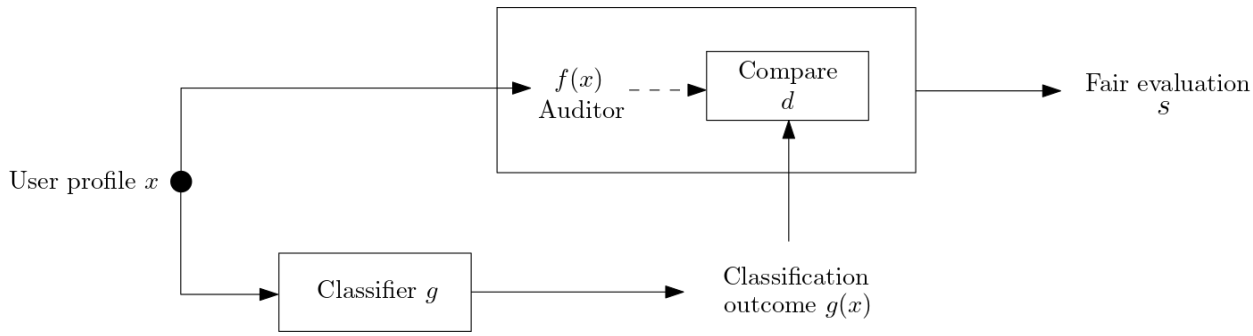


Figure 2: Latent Assessment Model of the Expert Auditor

2.2 Individual Fairness

Individual fairness assessments, rather than measuring discrimination across different sensitive groups, consider fairness for each individual, with the assumption that similar individuals should be treated as similarly as feasible (Dwork et al., 2012). Formally, given any two individuals $x_i, x_j \in \mathcal{X}$ and for some $\kappa, \delta \geq 0$, the predictor f is (κ, δ)-individually fair if

$$d\Big(f(x_i), f(x_j)\Big) \leq \delta, \quad \text{if } \mathcal{D}\left(x_i, x_j\right) \leq \kappa. \tag{5}$$

Note that the above definition contains a slack in the bound presented in Equation equation 5. Since the original definition of individual fairness directly compares the two similarlity metrics $d(f(x_i), f(x_j))$ and $\mathcal{D}(x_i, x_j)$, it becomes very challenging to find appropriate distance metrics. Furthermore, this relaxed definition enables us to evaluate a relationship between our fairness model and individual fairness. Such relaxations have also been considered in the past for rigorous analysis Yona & Rothblum (2018).

Unfortunately, in most practical applications, the similarity metric $\mathcal{D}(x_i, x_j)$ is task-specific and is typically unknown, which causes a severe restraint on our ability to ensure individual fairness. As a solution, metric learning was proposed to discern a task-specific similarity metric by evaluating the relative distance between human judgements (Ilvento, 2019) for any given pair of inputs. On the other hand, (Mukherjee et al., 2020) utilizes Mahalanobis distance as a fair metric and proposed EXPLORE, an algorithm to learn similarity between individuals from pairs of comparable and incomparable samples. It learns similarity such that the logistic regression predicts "comparable" when the fair distance is short, and "incomparable" when the fair distance is large. Interested readers can refer to (Fleisher, 2021) which discussed various inefficiencies of individual fairness in detail.

3 Latent Assessment Model for Non-Comparative Fairness

Consider an expert auditor who is presented with a data tuple $(x, \hat{y})$, where $x \in \mathcal{X}$ is the input given to ML model g and $\hat{y} = g(x) \in \mathcal{Y}$ is the output label as shown in Figure 2. Let f denote the expert auditor's decision rule, $y = f(x)$ be the intrinsic subjective evaluation for the input x, and $s(x, y, \hat{y})$ represent the auditor's binary feedback regarding the input-output pair (x, y) when the auditor uses an intrinsic evaluation $\hat{y} = f(x)$. In this paper, we model auditor's judgments from a non-comparative perspective as follows:

Definition 1 (ϵ-Latent Assessment Model)**.** *An auditor is said to satisfy ϵ-LAM if there exists a tuple $(\mathcal{X}, \mathcal{Y}, d, f, \epsilon)$ such that the auditor compares the system's output $g(x)$ with an intrinsic judgment $f(x)$ using a distance metric d and reveal his/her binary feedback as*

$$s\big(x, y = g(x), \hat{y} = f(x)\big) = \begin{cases} 1, & \textit{if } \; d\big(g(x), f(x)\big) \geq \epsilon, \\ 0, & \textit{otherwise.} \end{cases} \tag{6}$$

Furthermore, we say the auditor f is ϵ-LAM with respect to the system g if $d\big(g(x), f(x)\big) < \epsilon$ for all $x \in \mathcal{X}$.

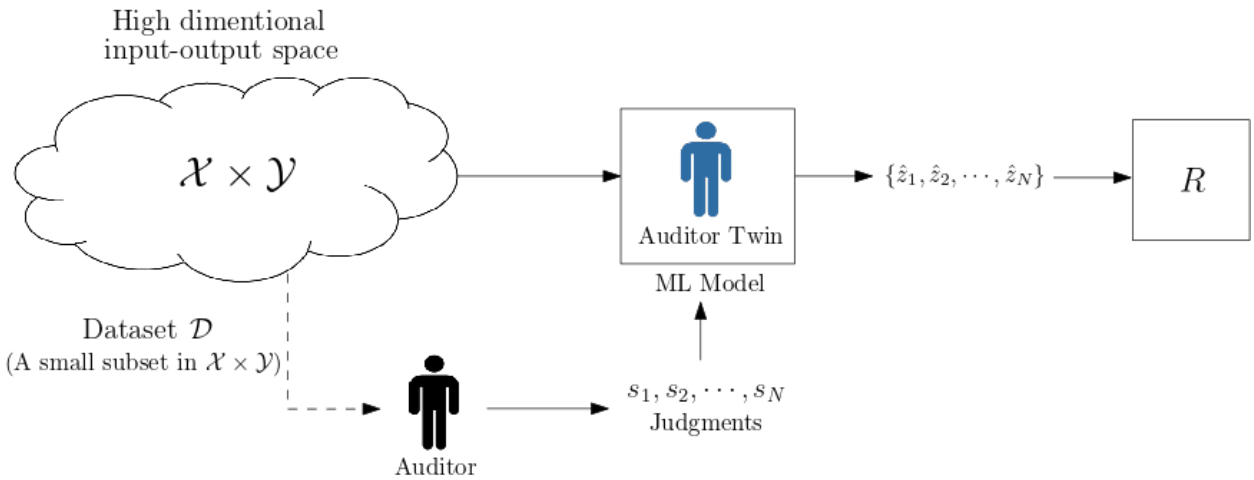


Figure 3: Auditor Evaluation Framework based on ϵ-LAM

In other words, if the distance between the output label $g(x)$ and the subjective evaluation $f(x)$ is greater than ϵ, the auditor will deem $g(x)$ as *unfair*. For the sake of illustration, consider an individual who committed felony and has multiple prior offences, received low recidivism score from a risk assessment tool. The expert auditor evaluates the individual intrinsically and may decide that he/she should receive a higher recidivism score. Then, the auditor judges the tool's output as unfair. In this paper, we assume the fair relation f employed by the expert auditor is unknown. Therefore, we need to learn the proposed ϵ-LAM using statistical learning techniques. This will be discussed in Sections 4.1 and 4.2 in greater detail. Although ϵ-LAM comprises of three unknowns in practice, namely d, f and ϵ, we assume that the distance metric d used by the auditor is known.

Remark 1. *In the case of binary classification, the Hamming distance between g and f, denoted as $d_H(g(x), f(x))$, takes a binary value of 0 or 1. Therefore, for any $\epsilon \in (0, 1)$, ϵ-LAM model in Definition 1, $s = g(x) \oplus f(x)$, where $\oplus$ denotes the XOR operation.*

Note that auditors exhibit different types of biases in the real-world. Examples include confirmation bias, hindsight bias, anchoring bias, racial and gender bias. Every auditor is susceptible to biases depending on their prior experiences and/or knowledge about various social groups within the community. Consequently, it is not reasonable to accept a given auditor's feedback without evaluating their inherent biases. Our proposed model, LAM, captures auditor biases through the function f in equation 6. However, in order to evaluate auditor biases, we investigate the relationship between our proposed LAM and traditional fairness notions. By doing so, we can estimate any given auditor's performance in terms of multiple fairness notions.

Since there are multiple fairness perspectives which are not necessarily aligned with each other, we propose a novel auditor evaluation framework in Figure 3 to quantify his/her performance across fairness notions with respect to diverse sensitive groups. Since auditor's intrinsic evaluation is unknown, we learn and estimate auditor's responses using standard ML algorithms - logistic regression, decision tree, and SVM. We assume that the auditor is presented a dataset $\mathcal{D}$ of size N, for feedback elicitation. The auditor evaluates each data tuple $i \in \mathcal{D}$, and presents a binary feedback $s_i \in \{0, 1\}$. Based on the feedback elicited from the auditor, we can predict auditor's intrinsic labels $\{\hat{z}_i = \hat{f}(x_i)\}_{i \in \mathcal{D}}$ via learning the auditor's intrinsic rule $\hat{f}$. However, if the original system is a binary classifier, the problem of finding the auditor's intrinsic evaluations is very straightforward. Specifically, for any given input x, the proposed $\epsilon - LAM$ in Definition 1 reduces to $s_i = g(x_i) \oplus f(x_i)$ in a binary classification setting. Since the XOR function is reversible, we have $\hat{z}_i = z_i = f(x_i) = y_i \oplus s_i$, where $y_i = g(x_i)$ is the true label in the data tuple.

We measure the overall bias in the system from a noncomparative fairness perspective as defined below.

Definition 2 (System Bias)**.** *The* true system bias *from a noncomparative fairness perspective is defined formally as the average feedback obtained directly from the original auditor, i.e.* $\mathbb{E}_x[s]$.

On the other hand, the empirical system bias *from a noncomparative fairness perspective is given by the empirical estimate of the true bias, i.e.* $\mathbb{E}_x[\hat{s}]$*, where* $\hat{s}$ *is computed using the estimated auditor rule* $\hat{f}$ *as well as the estimated system's outcome* $\hat{g}$*.*

Note that both f and g need to be learned from data since the paper assumes that the auditor does not have access to the original ML system g and are also incapable of expressing their intrinsic evaluations f formally in functional form.

4 LAM-Based Auditor Twin for Analyzing High-Dimensional Systems

Most ML systems map high-dimensional input spaces to low-dimensional outcomes. For example, in criminal justice or banking applications, input space spans across the entire population of individuals serviced by the ML system, while the outcome space is finite (e.g. binary decisions such as approval/disapproval of bails or loans.). In such a scenario, feedback elicitation from human auditors is extremely challenging due to the large number of feasible input-output pairs, which cannot be evaluated by any one human auditor. Therefore, there is a need for human-system teaming for a practically scalable fairness evaluation, wherein human auditors give feedback for a small subset of input-output pairs and the system learns their LAM from elicited feedback and scales it across all other possibilities for a wholesome evaluation. This section focuses on finding the minimum number of feedback samples (each sample comprising of an input-output pair and binary feedback from the human auditor) needed to effectively learn auditor's LAM with prescribed guarantees from a probably approximately correct (PAC) learning perspective.

4.1 PAC Learning Guarantees

Definition 3 (PAC Learnability)**.** *We say that the classifier* f *is PAC-learnable (from a realizable sense) if there exists* $N > 0$*,* $\epsilon > 0$*,* $\delta > 0$*, and an algorithm* $\mathcal{A}$ *which receives* $n \geq N$ *i.i.d. samples from distribution* D *as input, and outputs an estimated classifier* $\hat{f}_n$ *with at least* $1-\delta$ *probability such that* $d(f, \hat{f}_n) \leq \epsilon$*.*

Note that the auditor's intrinsic rule f is not directly revealed, while eliciting his/her feedback regarding biases in the system-of-interest. Therefore, we need to compute the intrinsic rule $\hat{f}$ in order to reproduce auditor's judgements for other input possibilities. At the same time, the classifier is typically unavailable to the bias-evaluation platform, i.e., g is also unknown to the bias-evaluation platform. In other words, a practical bias-evaluation platform is expected to learn (f, g) to obtain a model $(\hat{f}, \hat{g})$, and identify an appropriate fairness notion for the given context so that the bias evaluation platform can algorithmically evaluate bias in a system with a large input space. In the following theorem, we provide guarantees for the algorithmic $\epsilon - LAM$ evaluations, based on estimated rules $\hat{f}$ and $\hat{g}$.

Theorem 1. *Let* N *denote the minimum number of samples needed to guarantee* $\epsilon - LAM$ *empirically, i.e.* $\mathbb{P}\left(d(\hat{g}_n, \hat{f}_n) < \epsilon_{lam}\right) > 1 - \delta_{lam}$ *for some* $\epsilon_{lam} > 0$ *and* $\delta_{lam} > 0$*. Then, for any auditor's intrinsic rule* f *and classifier* g*, there exists some* $0 < N_f, N_g < N$*,* $\epsilon_g, \epsilon_f, \epsilon, \delta, \delta_g, \delta_f > 0$ *such that* $\epsilon_g + \epsilon_f + \epsilon_{lam} < \epsilon$ *and* $\delta_g + \delta_f + \delta_{lam} < 2 + \delta$*, and an algorithm* $\mathcal{A}$ *that receives i.i.d. samples* $\{(x_i, y_i, z_i)\}_{i=1}^n$ *as input, and outputs rules* $\hat{f}_n$ *and* $\hat{g}_n$ *with a probability of* $d(f(x), g(x)) < \epsilon$ *being at least* $1-\delta$*, only when*

$$n \geq N \triangleq \min_{\epsilon_1, \epsilon_2, \delta_1, \delta_2} (\max\{N_g, N_f\}), \tag{7}$$

where N_f *and* N_g *satisfy PAC learning bounds for* f *and* g*.*

Proof. Our goal is to ensure that

$$\mathbb{P}(d(f(x), g(x)) < \epsilon) \geq 1 - \lambda \tag{8}$$

for some small $\epsilon > 0$ and $\delta > 0$. Assuming that there exists some $0 < \epsilon_g < \epsilon$ and $0 < \epsilon_f < \epsilon$, we obtain an upper bound to LHS in the above equation using triangle inequality, as shown below:

$$\begin{aligned} \mathbb{P}\Big(d(g(x), f(x)) < \epsilon\Big) \quad \leq \quad & \mathbb{P}\Big(d(g(x), \hat{g}_n(x)) < \epsilon_g\Big) \; + \; \mathbb{P}\Big(d(f(x), \hat{f}_n(x)) < \epsilon_f\Big) \\ & + \mathbb{P}\Big(d(\hat{f}_n(x), \hat{g}_n(x)) < \tau\Big) \end{aligned} \tag{9}$$

where $\tau = \epsilon - \epsilon_g - \epsilon_f$.

Note that the first and the second probability terms correspond to PAC learnability guarantees of g and f respectively. Let $N_g(\epsilon_g, \delta_g)$ and $N_f(\epsilon_f, \delta_f)$ denote the minimum samples needed to guarantee PAC learnability at g and f respectively. In other words, the maximum of the two numbers will guarantee PAC learnability of both g and f, i.e. $N(\epsilon_g, \epsilon_f, \delta_g, \delta_f) = \max\{N_g(\epsilon_g, \delta_g), N_f(\epsilon_f, \delta_f)\}$. However, for the valid inequality in Equation equation 9, it is also essential to split δ between PAC learning bounds for g and f, as well as the probability corresponding to empirical $\epsilon - LAM$ $\mathbb{P}\Big(d(\hat{f}_n(x), \hat{g}_n(x)) < \epsilon_{lam}\Big)$. In other words, the split is valid only when

$$\begin{aligned} &\epsilon_g + \epsilon_f + \epsilon_{lam} < \epsilon, \text{ and} \\ &(1 - \delta_g) + (1 - \delta_f) + (1 - \delta_{lam}) > 1 - \delta. \end{aligned} \tag{10}$$

Then, N can be minimized by choosing an appropriate split ϵ_g, ϵ_f, δ_g and δ_f to obtain the necessary guarantee stated in Equation equation 8. □

Though we computed the minimum number of samples required, there should exist an algorithm $\mathcal{A}$ that receives i.i.d samples as input to estimate the intrinsic fair rule f of the expert auditor. Therefore, we validate our findings using different ML models to learn and the predict auditor responses using standard ML algorithms.

4.2 Learning Audit Twins using Elicited Feedback from Real Human Auditors

In contrast to simulated responses, we analyze real human feedback data elicited by (Dressel & Farid, 2018) using different fairness notions and also learn crowd worker's intrinsic rule f using various machine learning algorithms (e.g. logistic regression, decision trees, and support vector machines). This data acquisition experiment consists of a short description of the defendant (gender, age, race, and previous criminal history) is provided to the human auditor. A total of 1000 defendant descriptions are used that are drawn randomly from the original ProPublica's COMPAS dataset. Furthermore, these descriptions were divided into 20 subsets of 50 each. The experiment consisted of 400 different crowd workers and each one of them was randomly assigned to see one of these 20 subsets. The participants predicted whether a particular individual would recidivate within 2 years of their most recent crime. The original data consists whether a crowd worker predicted correctly or not compared to the original classification in the COMPAS dataset. We preprocessed this dataset and obtained the true prediction given by the crowd workers.

In Section 8.1, we learn that the choice of learning algorithm varies from one auditor to another. Therefore, in this experiment, we train each auditor using all three algorithms and choose the one which has the highest accuracy in predicting his/her responses. Figure 4 highlights the number of crowd workers whose responses are predicted using ML algorithms across different accuracy levels. It can be observed that logistic regression performs well by predicting the responses of 249 crowd workers with accuracy greater than 80% compared to other models. Based on predicted responses from the best learning algorithms, we analyze the fairness of crowd workers for various notions.

In terms of *individual fairness*, Figure 5 shows that only a few crowd workers are individually fair (0 violated clusters). The majority of the crowd workers violate 5% to 20% of the clusters present in the subset given to them. If observed carefully, the distribution of crowd workers satisfying individual fairness resembles gamma distribution when $\alpha = 12.75$ and $\beta = 0.0187$. Furthermore, we evaluate the crowd workers based on different *group fairness* notions (statistical parity, equalized odds, calibration, and equal accuracy) as defined in Section 2.1 given a probabilistic threshold δ. For each crowd worker, we assess their performance based on all the four group fairness notions by varying the probabilistic threshold δ from 0 to 0.10. Figure 6 shows the cumulative number of workers who comply with different group fairness notions increases as the threshold δ increases with respect to both race and gender. Only a few crowd workers comply with group fairness notions when $\delta = 0$ i.e. they exhibit absolutely no bias in terms of respective notion. However, the majority of the crowd workers' judgments are unfair (when $\delta > 0.1$) with respect to race and gender across all three notions. An interesting result that can be observed here is that the crowd workers are satisfying multiple group fairness notions simultaneously. To demonstrate this, we evaluate the crowd workers for combination

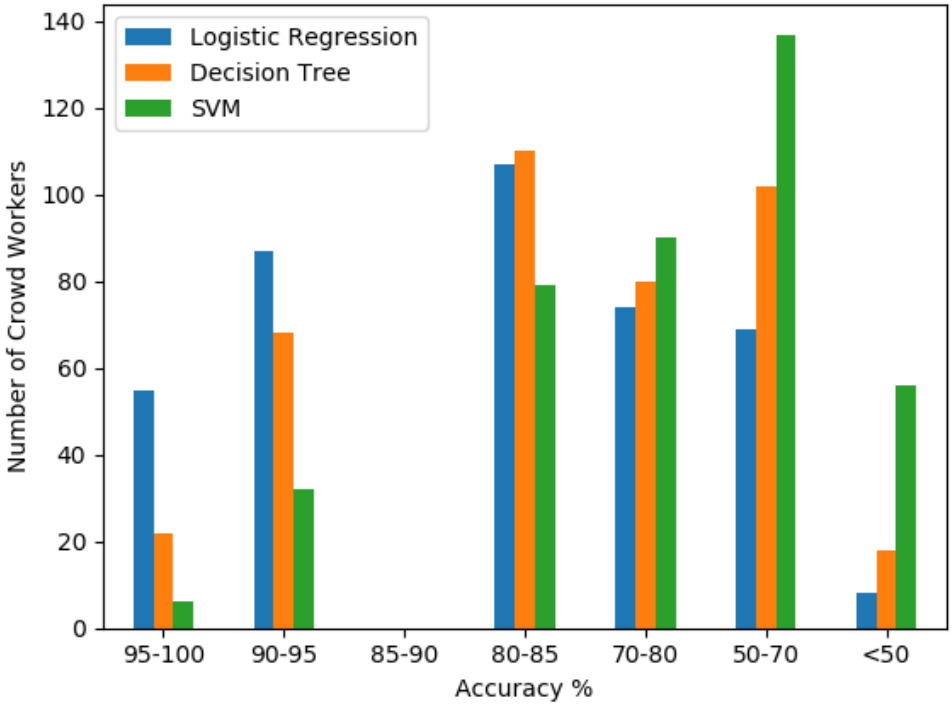


Figure 4: The performance of different ML algorithms to predict crowd workers' responses across various accuracy levels

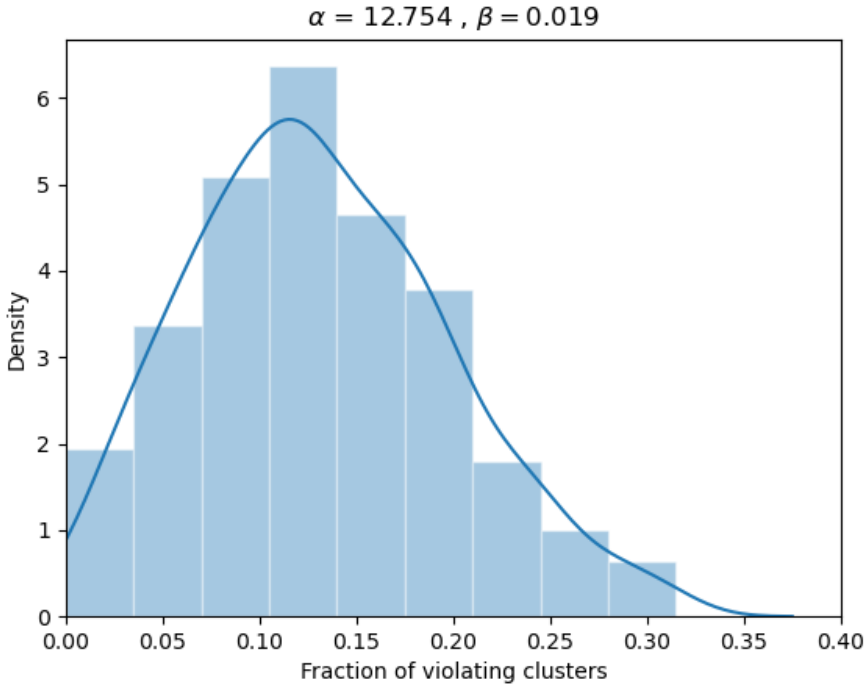


Figure 5: The distribution of crowd workers based on the fraction of clusters violating individual fairness.

of group fairness notions (e.g. statistical parity and equalized odds, equalized and calibration) for both the sensitive attributes. Figure 7 shows that the number of crowd workers who satisfy multiple group fairness

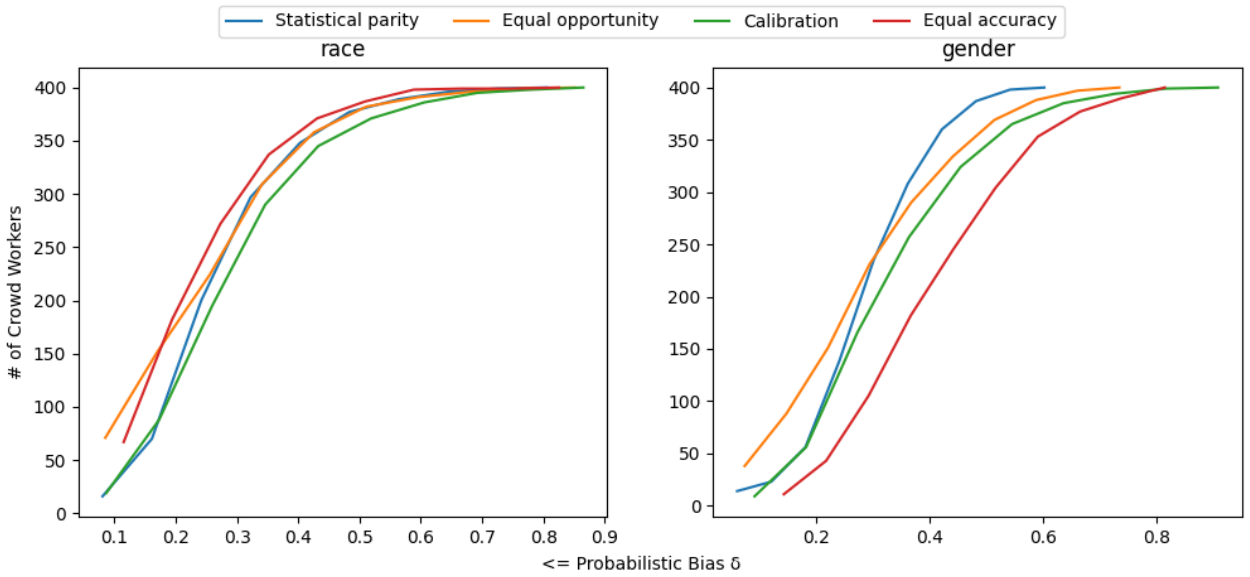


Figure 6: The cumulative number of crowd workers who comply with different group fairness notions with respect to race and gender.

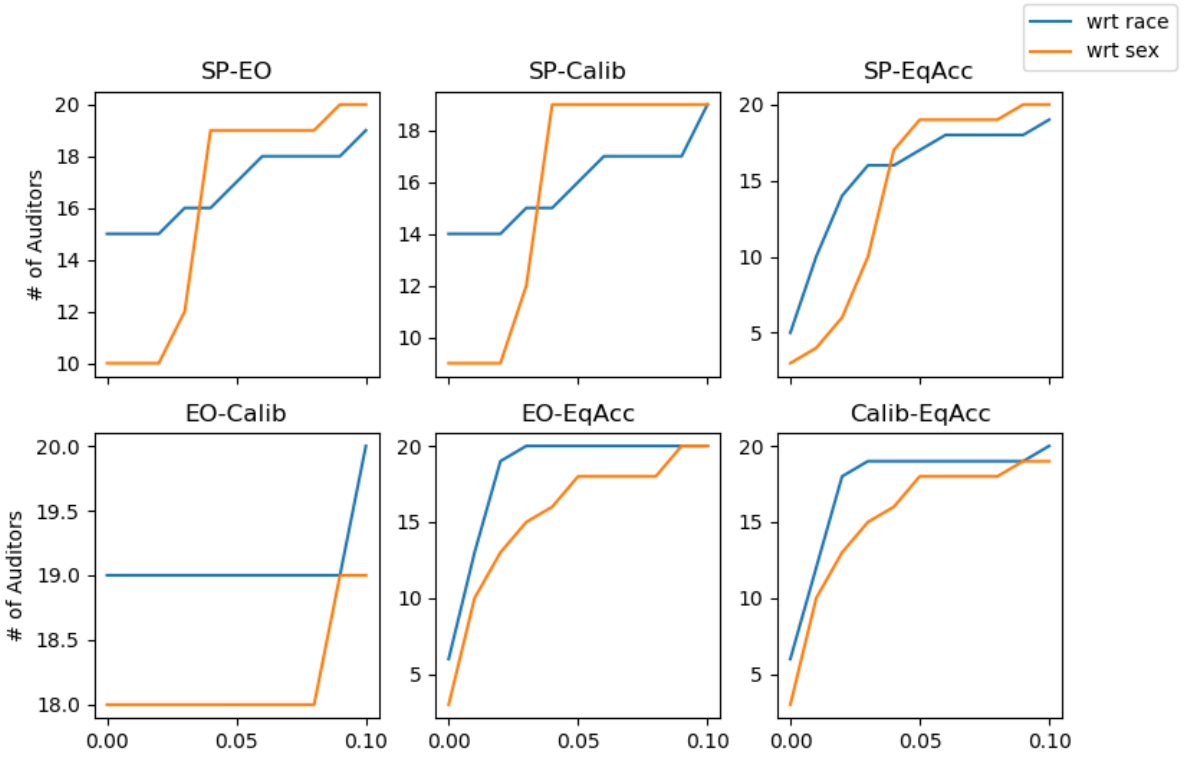


Figure 7: Number of crowd workers who satisfy multiple group fairness notions simultaneously increases with threshold δ.

notions simultaneously increases as the threshold δ increases with respect to both race and gender. Thus, it is not a trivial task to decide on what fairness notion is appropriate based on human judgments since different evaluators have different preferences on fairness notions.

5 Relationship Between Non-Comparative and Comparative Fairness Notions

Although comparative and non-comparative fairness notions are fundamentally different from each other, this section discusses their mutual relationship formally.

5.1 Relation with Individual Fairness

In the following proposition, we show how the system g can be evaluated based on the notion of (κ, δ)-individual fairness, when the auditor f is ϵ-LAM with respect to the system g.

Proposition 1. *A system g is $(\kappa, 2\epsilon + \delta)$-individually fair, if its auditor f satisfies ϵ-LAM and is (κ, δ)-individually fair.*

Proof. Given (x_i, y_i) and $(x_j, y_j) \in \mathcal{X} \times \mathcal{Y}$ such that $\mathcal{D}(x_i, x_j) \leq \kappa$ (the two individuals are κ-similar), then f is (κ, δ)-individually fair if $d\Big(f(x_i), f(x_j)\Big) < \delta$. However, note that if g is ϵ-LAM with respect to f, then $d\Big(g(x_i), f(x_i)\Big) < \epsilon$ and $d\Big(g(x_j), f(x_j)\Big) < \epsilon$. Therefore, by applying a chain of triangle inequalities, we obtain

$$\begin{aligned} d\Big(g(x_i), g(x_j)\Big) &\leq d\Big(g(x_i), f(x_i)\Big) + d\Big(f(x_i), f(x_j)\Big) + d\Big(f(x_j), g(x_j)\Big) \\ &< 2\epsilon + \delta, \end{aligned} \tag{11}$$

for all x_i, x_j such that $\mathcal{D}(x_i, x_j) \leq \kappa$. □

Remark 2. *In the case of binary classifiers, if the auditor f is κ-individually fair, we have $f(x_i) = f(x_j)$ for all x_i, x_j whenever $\mathcal{D}(x_i, x_j) \leq \kappa$. Furthermore, if the auditor f is ϵ-LAM with respect to the system g for any $\epsilon \in (0, 1)$, we have $g(x) = f(x)$ for all $x \in \mathcal{X}$. Combining the above two properties, we get $g(x_i) = f(x_j) = f(x_i) = g(x_j)$ for all x_i, x_j such that $\mathcal{D}(x_i, x_j) \leq \kappa$.*

We illustrate this result using the following example from the banking domain. Consider two individuals who are looking to apply for a loan. A banking system would evaluate both the applications via collecting information such as gender, race, address, credit history, collateral, and his/her ability to pay back. At the same time, consider an auditor who makes fairness judgements based on the rule: "If he/she has cleared all the debts and possesses reasonably valued collateral, the loan must be granted". Given that the auditor treats any two similar individuals similarly, the auditor satisfies individual fairness. Hence, if the evaluation of the banking system is relatively similar to the auditor's fair relation, from Proposition 1, the banking system is also individually fair.

Proposition 2. *If an auditor f is not (κ, δ)-individually fair, but satisfies ϵ-LAM with respect to the system g, then the system is not $(\kappa, \delta - 2\epsilon)$-individually fair as well.*

Proof. If f is not individually fair, then for some input pair (x_1, x_2) such that $\mathcal{D}(x_1, x_2) < \kappa$, we have $d\left(f(x_1), f(x_2)\right) > \delta$ for all $\kappa, \delta \in \mathbb{R}$. However, note that if g is ϵ-LAM with respect to f, then $d\left(g(x_1), f(x_1)\right) < \epsilon$ and $d\left(g(x_2), f(x_2)\right) < \epsilon$. Therefore, by applying a chain of triangle inequalities, we have

$$d\left(f(x_1), f(x_2)\right) \leq d\left(g(x_1), f(x_1)\right) + d\left(g(x_2), f(x_2)\right) + d\left(g(x_1), g(x_2)\right) \tag{12}$$

Substituting the bounds of $d\left(g(x_2), f(x_2)\right)$ and $d\left(g(x_1), g(x_2)\right)$ we get

$$\begin{aligned} 2\epsilon + d\left(g(x_1), g(x_2)\right) &> d\left(g(x_1), f(x_1)\right) + d\left(g(x_2), f(x_2)\right) + d\left(g(x_1), g(x_2)\right) \\ &\geq d\left(f(x_1), f(x_2)\right) \\ &> \delta \end{aligned} \tag{13}$$

for all $\delta \in \mathbb{R}$. Therefore, we also have $d\left(g(x_1), g(x_2)\right) > \delta - 2\epsilon$. □

Remark 3. *For binary classification, Proposition 2 can be reduced as follows. Note that if f is not κ-individually fair, we have $f(x_i) \neq f(x_j)$ even though $\mathcal{D}(x_i, x_j) \leq \kappa$. Furthermore, if g is $\epsilon - LAM$ with respect to f for any $\epsilon \in (0, 1)$, we have $g(x) = f(x)$ for all $x \in \mathcal{X}$. Combining the above two properties, we get $g(x_i) = f(x_j) \neq f(x_i) = g(x_j)$ for all x_i, x_j such that $\mathcal{D}(x_i, x_j) \leq \kappa$.*

Consider the earlier example of banking where, there are two individuals, A and B, who possess the same degree of merit. Imagine that the bank approves A's loan application and denies B. This outcome remains the same as per the auditor's fair relation. Imagine further that neither A nor B merits the outcome. Though both banking's evaluation and auditor's judgements seem to be similar, they violate the precept, "treat similar individuals similarly". Moreover, A is treated in a way that A does not merit. Hence, we can assert that banking evaluation does not satisfy individual fairness.

5.2 Relation with Group Fairness Notions

As discussed earlier, group fairness notions compare certain probabilistic measure across two protected groups. In the remaining section, we will focus on the relationship between group fairness notions and our proposed LAM. For the sake of convenience, let us denote $p_{x,y}(g,a) = \mathbb{P}[g(x) = y \mid A = a]$.

Proposition 3. *Given that the probability distributions are M-Lipschitz continuous over all possible f and g functions, g satisfies $(2M\epsilon + \delta)$-statistical parity, if g is ϵ-LAM with respect to f, and f satisfies δ-statistical parity.*

Proof. Given the set of protected attributes $\mathcal{A}$, since f satisfies δ-statistical parity, we have $||p_{x,y}(f,a) - p_{x,y}(f,a')|| < \delta$ for all $a, a' \in \mathcal{A}$. Then, we have

$$\begin{aligned} p_{x,y}(g,a) - p_{x,y}(g,a') \quad = \quad & [p_{x,y}(g,a) - p_{x,y}(f,a)] \; + \; [p_{x,y}(g,a') - p_{x,y}(f,a')] \\ & + \; [p_{x,y}(f,a) - p_{x,y}(f,a')] \end{aligned} \tag{14}$$

Assuming M-Lipschitz continuity over all $f(x)$, $g(x)$, we have $||p_{x,y}(g,a) - p_{x,y}(f,a)|| < M \cdot \epsilon$, since $d(g(x), f(x)) < \epsilon$. Combining all the inequalities, we have

$$||p_{x,y}(g,a) - p_{x,y}(g,a')|| < 2M\epsilon + \delta. \tag{15}$$

□

Again, consider the earlier example of loan approvals to illustrate the above proposition. Consider that there exists two groups which are classified based income - low and high. The banking system builds a credit model based purely. Moreover, the system may decide to use different requirement levels - low interest or default to low income group, so that the percentage of people getting a loan in low-income group is equal to the percentage of people getting a loan in high-income group. Now, suppose an auditor presents fair judgements based on the rule: "If Group A has a FICO credit score of 550 and cleared all the debts, the loan must be granted. If Group B has a FICO score of 700 and has valuable collateral, grant the loan". Note that, the auditor's fair relation is somewhat similar to that of the bank's policy. Since the bank's policy is known to be statistically fair, the auditor is also unbiased from a group fairness perspective.

Similarly, the following three propositions identify the relationship between our proposed $\epsilon - LAM$ and three other group fairness notions, namely equal opportunity, calibration, and equal accuracy.

Proposition 4. *Given that the probability distributions are M-Lipschitz continuous over all possible f and g functions, g satisfies $(2M\epsilon + \delta)$-equal opportunity, if g is ϵ-LAM with respect to f, and f satisfies δ-equal opportunity.*

Proof. The proof is similar to that of Proposition 3. Therefore, for the sake of brevity, the proof is not included. □

Proposition 5. *Given that the probability distributions are M-Lipschitz continuous over all possible f and g functions, g satisfies $(2M\epsilon + \delta)$-calibration, if g is ϵ-LAM with respect to f, and f satisfies δ-calibration.*

Proof. The proof is similar to that of Proposition 3. Therefore, for the sake of brevity, the proof is not included. □

Proposition 6. *Given that the probability distributions are M-Lipschitz continuous over all possible f and g functions, g satisfies $(2M\epsilon + \delta)$-equal accuracy, if g is ϵ-LAM with respect to f, and f satisfies δ-equal accuracy.*

Proof. The proof is similar to that of Proposition 3. Therefore, for the sake of brevity, the proof is not included. □

6 Comprehensive Reputation of Human Auditors

Given the auditor's intrinsic labels $\{\hat{z}_1, \cdots, \hat{z}_K\}$, we can compute the auditor's performance for a given sensitive/protected group in terms of various fairness notions such as statistical parity (ref. Equation equation 1), equal opportunity (ref. Equation equation 2), calibration (ref. Equation equation 3) and individual fairness (ref. Section 2.2). Furthermore, note that this multi-dimensional fairness evaluation is different for different sensitive groups. For example, in the United States, protected groups are typically defined based on race, gender, religion or any combination of these attributes. Such multi-attribute fairness evaluations across different sensitive/protected groups naturally steers us towards defining a multi-attribute reputation matrix

$$R(\hat{z}_1, \cdots, \hat{z}_K) = \begin{bmatrix} r_{1,1}(\hat{z}_1, \cdots, \hat{z}_K) & \cdots & r_{1,L}(\hat{z}_1, \cdots, \hat{z}_K) \\ \vdots & \ddots & \vdots \\ r_{M,1}(\hat{z}_1, \cdots, \hat{z}_K) & \cdots & r_{M,L}(\hat{z}_1, \cdots, \hat{z}_K) \end{bmatrix}, \tag{16}$$

where M is the total number of fairness notions and L is the total number of sensitive groups. For example, if we are evaluating the auditor based on *statistical parity (SP)* with respect to the sensitive attribute *race*, then $r_{SP,\ race} = \mathbb{P}[f(x) = 1 \mid race = a] - \mathbb{P}[f(x) = 1 \mid race = a']$.

Although we propose a multi-dimensional fairness evaluation in a matrix format R, it is necessary to represent auditor's biases as a one-dimensional score signifying his/her overall performance with respect to various fairness notions. In crowdsourcing literature, auditors' reliability is measured by comparing their responses with majority vote (Jamaludeen et al., 2019) or ground truth (Le et al., 2010). Since majority vote assumes that every auditor has same expertise, it cannot be applied to our framework as the biases of the auditors vary from one another. Moreover, due to the absence of ground truth, it is not possible to compare auditor responses. Therefore, we apply Frobenius norm of the reputation matrix R to compute the auditor's scalar reputation score as follows.

$$||R||_F = \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{N} |r_{i,j}|^2} \tag{17}$$

We choose this reputation score due to the following axiomatic properties:

- ***Perfect Fairness:*** A utopian auditor satisfies all the fairness notions, i.e. every entry in R becomes zero. Consequently, $||R||_F \to 0$.

- ***Lipschitz-Boundedness:*** Consider any deviation Δ from R. Then, we have $||R + \Delta||_F \leq ||R||_F + ||\Delta||_F$, due to triangle inequality. In other words, the Frobenius norm based score satisfies Lipschitz property, since

  $$\frac{||R + \Delta||_F - ||R||_F}{||\Delta||_F} \leq 1.$$

  Lipschitz property is a particularly important since there is a bound to the change in score, even though the auditor exhibits dynamic preferences regarding fairness notions.

- ***Equal Treatment of Fairness Notions:*** Frobenius norm of the matrix R can also be represented as follows.

  $$||R||_F = \sqrt{tr(R^T R)} \tag{18}$$

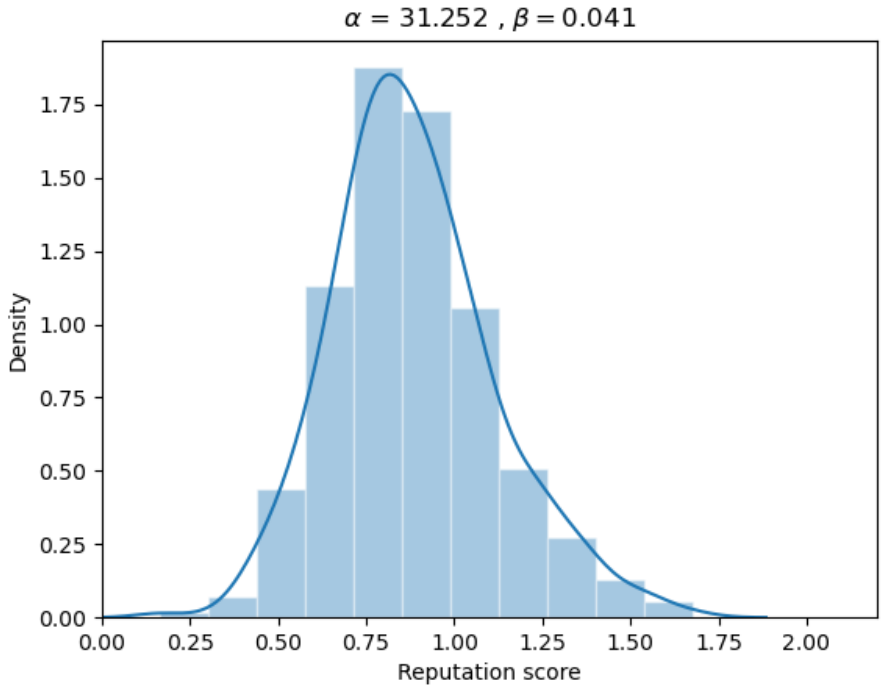


Figure 8: The reputation score distribution of 400 crowd workers.

Let R, UR, and RV be the reputation matrices of three different auditors. Then, we have

$$||UR||_F = ||RV||_F = \sqrt{tr(R^T R)} = ||R||_F. \tag{19}$$

In other words, the three auditors with reputation matrices which differ by orthogonal transformations have the reputation score.

For the sake of illustration, consider two auditors: (i) one who complies with statistical parity, but not with calibration, and (ii) another who satisfies calibration but not statistical parity. Assuming that both (i) and (ii) treats all the remaining fairness notions identically, the Frobenius norm of their reputation matrices would be same. In other words, our reputation score treats all fairness notions equally.

6.0.1 Reputation Evaluation of Real Human Auditors

In this section, we compute the reputation of 400 crowd workers based on different fairness notions - statistical parity (SP), equalized odds (EO), calibration (C), equal accuracy (EA) with respect to race and gender, , and individual fairness (IF). The fairness of the workers is computed using the methodology discussed in Section 7.1. Upon computing the fairness of the worker based on different notions, we represent his/her reputation in a matrix format as shown in the Figure 9. The ideal value of every fairness measure is 0. Note that ϕ is used since individual fairness is independent of sensitive attributes. We then compute the reputation scores of the crowd workers using Frobenius norm of the matrices. Figure 8 demonstrates the distribution of crowd workers based on their reputation scores. The minimum and maximum reputation scores are 0.16 and 1.67 respectively. It can observed that majority of them obtained a reputation score between 0.5 and 1.0. However, an ideal crowd worker would obtain a lower reputation between 0 and 0.25. No crowd worker is absolutely fair with respect to every fairness notion ($||R||_F = 0$). Moreover, the distribution of crowd workers comply with gamma distribution when $\alpha = 31.25$ and $\beta = 0.04$.

$$R = \begin{array}{c} \\ \text{SP} \\ \text{EO} \\ \text{C} \\ \text{EA} \\ \text{IF} \end{array} \begin{array}{c} \begin{array}{ccc} \text{gender} & \text{race} & \phi \end{array} \\ \begin{bmatrix} 0.15 & 0.01 & 0 \\ 0.03 & 0.01 & 0 \\ 0 & 0.2 & 0 \\ 0.01 & 0.01 & 0 \\ 0 & 0 & 0.06 \end{bmatrix} \end{array}$$

Figure 9: Reputation matrix of a random crowd worker

7 Evaluation Methodology and Metrics

In this section, we discuss different methodologies and metrics used to evaluated our proposed ϵ-LAM based on simulation as well as real human audit data.

7.1 Demonstrating Interplay Between LAM and Fairness Notions

We validate our theoretical findings using the following datasets, each of which are pre-processed as follows:

1. ***ProPublica's COMPAS dataset (Larson et al., 2016):*** In this paper, we perform same preprocessing as the original analysis of ProPublica. The races in the dataset are only restricted to African-American, Caucasian, and other. We consider *females* and *Caucasians* as privileged groups. The feature *two year recidivism* (most likely or least likely) is considered as the output full and 0 (least likely) as the favourable outcome. Since the feature *age* is continuous, we create different age groups (e.g. 25-45 or > 45) and rename the features as *age category*. Similar grouping is also performed on *priors count* as well.

2. ***German credit data (Merz & Murphy, 1996):*** In this dataset, we consider credit history, savings, employment, sex, and age as input features. Moreover, we categorize *age* into two groups: young (< 26) and old ($>=26$). We assume that males and older individuals as privileged groups and 1 (good credit risk) as favourable outcome.

3. ***Adult income dataset (Kohavi & Becker, 1994):*** The objective is to predict whether the income of an individual is $> \$50K$ or $< \$50K$. The input features include age, sex, race, and education. In the pre-processing phase, the continuous feature *age* is transformed into different groups of ages (0-10, 11-20, and so on). For the feature *race*, we limited the labels to binary by mapping 'White' to 1 and all other races to 0. We have 32561 data tuples in total.

Since we have no access to the true labels while evaluating the auditor, we learn and estimate his/her responses using standard ML algorithms - logistic regression, decision tree, and support vector machine. The responses, along with the input, will be split into train (75%) and test (25%) sets. The fairness of the auditor is computed based on predicted labels and is averaged across 25 random train-test splits.

Recall that, individual fairness notion relies on a similarity metric $\mathcal{D}$ between two individuals. Inspired from prior work (Zemel et al., 2013; Lahoti et al., 2019; John et al., 2020), we construct clusters of similar individuals based on context-dependent and non-sensitive attributes from the respective dataset. Clusters are formed based on the metric Mahalanobis distance which is given as $\mathcal{D} = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T C^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j)}$, where $\boldsymbol{x}_i, \boldsymbol{x}_j$ are observations/rows in a dataset and C is positive semi-definite covariance matrix. In practice, we utilize the given dataset as a covariance matrix. For COMPAS, we utilize *juvenile counts* (both misdemeanor and felony), *prior offences* and *charge degree* to construct clusters (ref. Table 1 for example of 5 such clusters). Similarly, for German credit dataset, we consider *credit history*, *employment* and *savings* attributes. Lastly, for the Adult income dataset, we use *education years* and *age* to measure similarity. Upon constructing the clusters, we validate whether the respective entity (classifier or the auditor) assigns similar output labels for every individual in a particular cluster. An entity is said to violate a cluster when it produces different output

Sex	Age	Race	Prior Offenses	Charge Degree
Female	25-45	Caucasian	1 to 3	M
Male	25-45	Other	0	F
Female	Greater than 45	African-American	0	F
Male	25 - 45	Other	More than 3	M
Male	Greater than 45	Other	1 to 3	M

Table 1: Example of 5 different clusters present in COMPAS dataset

Dataset	**No. of Test Samples**	**Total No. of Clusters**	**No. of Clusters Violated**	**Auditor Evaluation**
COMPAS	1543	40	17	Individually fair
German credit	250	22	14	Individually fair
Adult income	8141	61	48	Individually fair

Table 2: Individual fairness evaluation of real-world classifiers and the auditor twin.

labels for individuals in that cluster. If an entity is not individually fair, this experiment would output the number of clusters violating individual fairness. In other words, the ideal value of this experiment is 0.

We now describe how different group fairness notions (statistical parity, equalized odds, and calibration) are measured. As defined in Section 2.1, the conditional probability differences are computed between the unprivileged group to the privileged group with respect to the favorable outcome. The ideal value of this difference is 0 for all four measures. In other words, if the probability difference is > 0, the underprivileged group benefits. Whereas, if the probability difference is < 0, the privileged group benefits. However, many group fairness notions rely on both true labels and predicted labels, e.g. equal opportunity and calibration. While evaluating a real-world classifier for group fairness, the auditor's judgments are true labels y, and the classifier's outputs are predicted labels $\hat{y}$.

We construct arbitrary intrinsic decision rules of the auditor for COMPAS, German credit, and Adult income datasets as follows. For COMPAS dataset, we consider *number of prior offences* and *degree of the offence* (felony or misdemeanor) to construct the fair relation. Note that, the binary feature *two year recidivism* (most likely or least likely) is viewed as the output feature.

$$f_1(x) = \begin{cases} 1, & \text{if } x.\text{priors-count} \in [1,3] \text{ and } x.\text{charge-degree} = \text{Felony} \\ & \qquad\qquad \text{OR} \\ & \text{if } x.\text{priors-count} > 3 \text{ and } x.\text{charge-degree} = \text{Misdemeanor} \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

Similarly, for German credit dataset, the features *savings, credit history* and *employment* are considered while designing the auditor's relation.

$$f_2(x) = \begin{cases} 1, & \text{if } x.\text{savings} > 500,\ x.\text{credit-history} = \text{Paid, and } x.\text{employment} > 2 \text{ years} \\ 2, & \text{otherwise.} \end{cases} \tag{21}$$

Since the task of the Adult income dataset is to predict whether yearly income of an individual is > 50K or $<= 50$K, we consider the feature *education* in the auditor's fair relation as shown below.

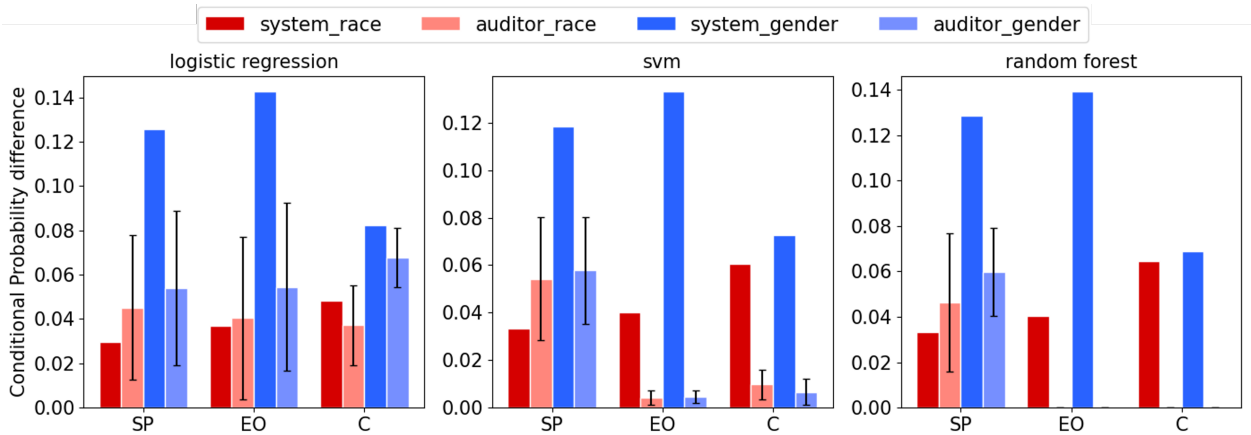
$$f_3(x) = \begin{cases} 1, & \text{if } x.\text{education} \in [\text{Bachelors, Masters, School Professor, Doctorate}] \\ 0, & \text{otherwise.} \end{cases} \tag{22}$$

8 Results and Discussion

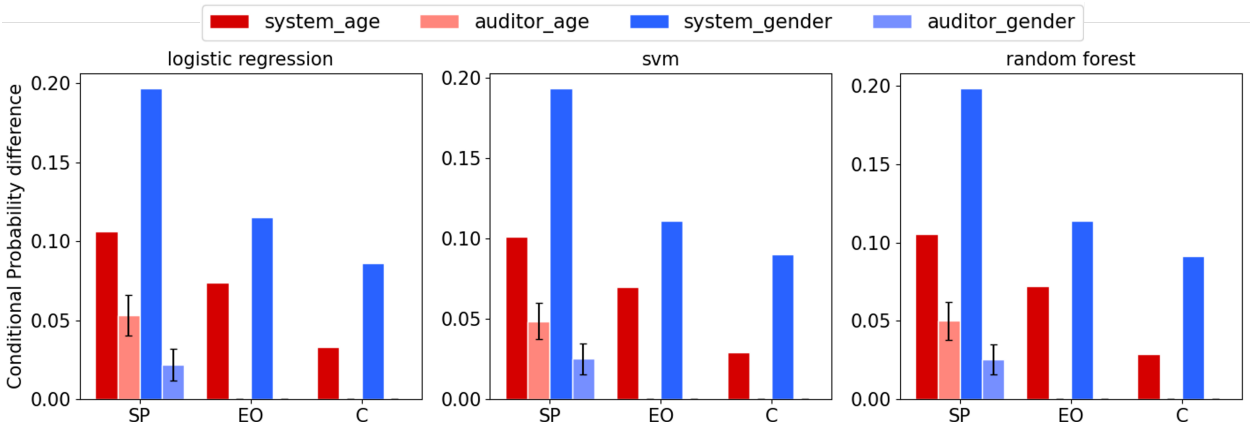
8.1 Simulating Human Assessments on Real Datasets

First, we evaluate *individual fairness* of COMPAS, German credit, and Adult income datasets, and compare the datasets with that of the auditor's simulated responses using the methodology discussed in Section 7.1

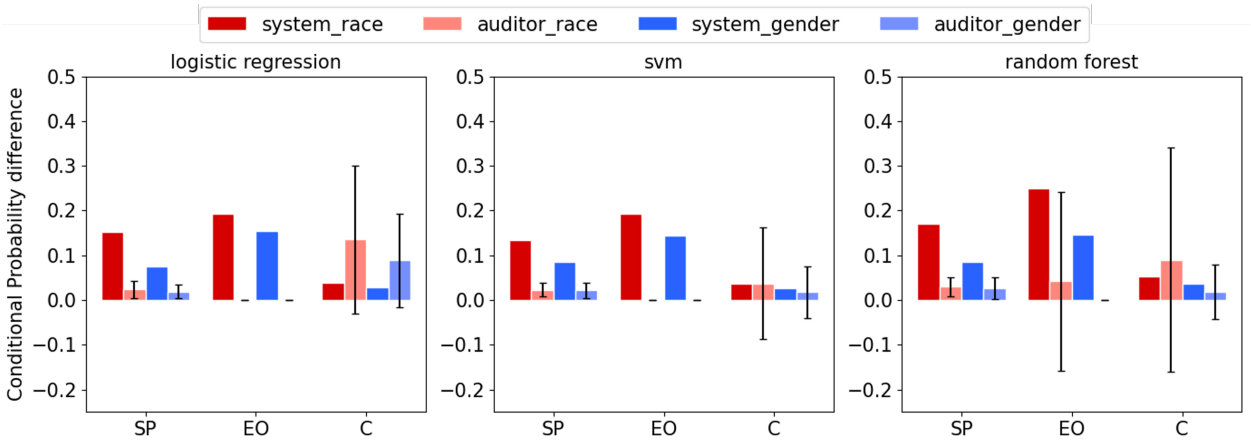
in Table 2. While none of the classifiers satisfied individual fairness, the simulated auditor is individually fair across all the three datasets regardless of the learning algorithm. Compared to the three classifiers, COMPAS is fairer in terms of individual fairness by satisfying 57.5% of the clusters. On the contrary, the German credit and the Adult income datasets satisfy only 36% and 21% of clusters respectively. Note that, this experiment is restricted to the features used to construct the similarity metric and the might vary with a different set of features.



(a) COMPAS



(b) Adult income



(c) German credit

Figure 10: Evaluation of real-world classifiers and the auditor twin for group fairness notions - statistical parity (SP), equal opportunity (EO) and calibration (C), with respect to different protected attributes.

However, the fairness of the auditor twin is heteroskedastic in terms of *group fairness* notions. In Figure 10, we demonstrate how LAM works in the case of three real datasets, namely COMPAS, Adult income, and German credit as discussed in Section 7.1. Though all three learning algorithms predicted auditor responses with almost equal accuracy ($\sim 96\%$), the fairness of the auditor twin is quite different from one another. In the context of the COMPAS dataset, we observe that the auditor twin's judgments are fair when trained using SVM and random forest. For the specific auditor rule chosen in our simulation experiment, the auditor twin is fairer compared to COMPAS in terms of equal opportunity and calibration with respect to both race and gender (ref. Figures 10a). Interestingly, the fairness of the auditor twin remains the same for the Adult income dataset across all three learning algorithms in contrast to the COMPAS dataset. Moreover, the auditor twin is fairer compared to the Adult income dataset in terms of both race and gender across all notions. On the other hand, for the German credit dataset, the auditor twin's judgments are fair when trained using SVM. The fairness of the auditor twin in terms of statistical parity remains the same across all three learning algorithms.

Based on the obtained results, it is essential to choose the appropriate learning model to replicate the auditor's responses not only in terms of accuracy but also in fairness. In addition, the choice of the learning algorithm relies upon the application context as well. We can also observe that the auditor twin satisfies both individual fairness and certain group fairness (with minimal probability difference) notions simultaneously. For instance, the auditor twin is absolutely fair in terms of equal opportunity and calibration for COMPAS and Adult income (when the learning algorithm is random forest). At the same time, the auditor twin is individually fair across all three datasets. Moreover, the auditor twin's judgments comply with multiple group fairness notion (e.g. equal opportunity and calibration with respect to random forest) for probabilistic bias $\delta = 0$ simultaneously. However, there's no guarantee that human auditors would exhibit biases similar to the shown results in reality. In fact, to support this claim, we analyze real human feedback data based on different fairness notions and show that not all human auditors perceive fairness in the same manner and can exhibit unfair judgments.

9 Conclusion and Future Work

We developed a novel latent assessment model to characterize human auditor feedback and demonstrated its relationship with traditional fairness notions both theoretically and on real datasets. We obtained PAC learning guarantees on learning auditor's intrinsic fairness assessments, and demonstrated the learning performance of three learning algorithms on a real human feedback dataset. Consequently, this paper enabled us to accomplish two important challenges in the design of a crowd-auditing platform: (i) we can learn/mimic auditor's intrinsic evaluations using little elicited feedback and automate the evaluation on the remaining possibilities especially in high-dimensional learning algorithms, and (ii) we can also evaluate auditor biases with respect to diverse traditional fairness notions. In addition, we use the relationship between LAM and traditional fairness notions to identify reliable auditors for feedback elicitation based on their reputation scores.

In future, we will address all the other challenges in the design of crowd-auditing platforms. Since feedback elicitation is an expensive process, we will improve our LAM model to account for feedback for data bundles, as opposed to our current feedback model for singleton data tuples. Furthermore, we will also investigate appropriate fusion rules to aggregate feedback collected from multiple auditor with heterogeneous opinions based on their reputation.

References

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine Bias. *ProPublica*, May 23 2016.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2018.

Reuben Binns. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*, pp. 149–159. PMLR, 2018.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020.

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

Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

A Feder Cooper, Ellen Abrams, and Na Na. Emergent unfairness in algorithmic fairness-accuracy trade-off research. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 46–54, 2021.

Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pp. 797–806, 2017.

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness Through Awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226. ACM, 2012.

Joel Feinberg. Noncomparative justice. *The philosophical review*, 83(3):297–338, 1974.

Ray Fisman and Michael Luca. Fixing Discrimination in Online Marketplaces. In *Harvard Business Review*, December 2016.

Will Fleisher. What's fair about individual fairness? In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 480–490, 2021.

Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. Online learning with an unknown fairness metric. In *Advances in Neural Information Processing Systems*, pp. 2600–2609, 2018.

Naman Goel and Boi Faltings. Crowdsourcing with fairness, diversity and budget constraints. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 297–304, 2019.

Nina Grgic-Hlaca, Elissa M Redmiles, Krishna P Gummadi, and Adrian Weller. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 World Wide Web Conference*, pp. 903–912, 2018.

M. Hardt, E. Price, and N. Srebro. Equality of Opportunity in Supervised Learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 3315–3323. Curran Associates, Inc., 2016a.

Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, pp. 3315–3323, 2016b.

Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 392–402, 2020.

Christina Ilvento. Metric learning for individual fairness. *arXiv preprint arXiv:1906.00250*, 2019.

Noor Jamaludeen, Vishnu Unnikrishnan, Maya S Sekeran, Majed Ali, Le Anh Trang, and Myra Spiliopoulou. Assessing the reliability of crowdsourced labels via twitter. In *LWDA*, pp. 115–126, 2019.

Philips George John, Deepak Vijaykeerthy, and Diptikalyan Saha. Verifying individual fairness in machine learning models. In *Conference on Uncertainty in Artificial Intelligence*, pp. 749–758. PMLR, 2020.

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. *Advances in neural information processing systems*, 29, 2016.

Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, Logan Stapleton, and Zhiwei Steven Wu. Eliciting and enforcing subjective individual fairness. *arXiv preprint arXiv:1905.10660*, 2019.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 35–50. Springer, 2012.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Ronny Kohavi and Barry Becker. UCI machine learning repository, 1994. URL `http://archive.ics.uci.edu/ml`.

Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. ifair: Learning individually fair data representations for algorithmic decision making. In *2019 ieee 35th international conference on data engineering (icde)*, pp. 1334–1345. IEEE, 2019.

Jeff Larson, Julia Angwin, Lauren Kirchner, and Surya Mattu. How we analyzed the compas recidivism algorithm. *ProPublica*, May 23 2016.

John Le, Andy Edmonds, Vaughn Hester, and Lukas Biewald. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 workshop on crowdsourcing for search evaluation*, volume 2126, pp. 22–32, 2010.

Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C Parkes. Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875*, 2017.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

C. J. Merz and P. Murphy. UCI machine learning repository, 1996. URL `http://archive.ics.uci.edu/ml`.

Debarghya Mukherjee, Mikhail Yurochkin, Moulinath Banerjee, and Yuekai Sun. Two simple ways to learn individual fairness metrics from data. In *International Conference on Machine Learning*, pp. 7097–7107. PMLR, 2020.

Dana Pessach and Erez Shmueli. Algorithmic fairness. *arXiv preprint arXiv:2001.09784*, 2020.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pp. 5680–5689, 2017.

Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 99–106, 2019.

Megha Srivastava, Hoda Heidari, and Andreas Krause. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2459–2468, 2019.

Andrew Waxman. BankThink AI Can Help Banks Make Better Decisions, But it Doesn't Remove Bias. *American Banker*, June 05 2018.

Mohammad Yaghini, Andreas Krause, and Hoda Heidari. A human-in-the-loop framework to construct context-aware mathematical notions of outcome fairness. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 1023–1033, 2021.

Gal Yona and Guy Rothblum. Probably approximately metric-fair learning. In *International Conference on Machine Learning*, pp. 5680–5688. PMLR, 2018.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 1171–1180. International World Wide Web Conferences Steering Committee, 2017.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 325–333, 17–19 Jun 2013.