On the Identification of Fair Auditors to Evaluate Recommender Systems based on a Novel Non-Comparative Fairness Notion

MUKUND TELUKUNTA, Missouri University of Science and Technology

VENKATA SRIRAM SIDDHARDH NADENDLA, Missouri University of Science and Technology

Decision-support systems are information systems that offer support to people's decisions in various applications such as judiciary, real-estate and banking sectors. Lately, these support systems have been found to be discriminatory in the context of many practical deployments. In an attempt to evaluate and mitigate these biases, algorithmic fairness literature has been nurtured using notions of comparative justice, which relies primarily on comparing two/more individuals or groups within the society that is supported by such systems. However, such a fairness notion is not very useful in the identification of fair auditors who are hired to evaluate latent biases within decision-support systems. As a solution, we introduce a paradigm shift in algorithmic fairness via proposing a new fairness notion based on the principle of non-comparative justice. Assuming that the auditor makes fairness evaluations based on some (potentially unknown) desired properties of the decision-support system, the proposed fairness notion compares the system's outcome with that of the auditor's desired outcome. We show that the proposed fairness notion also provides guarantees in terms of comparative fairness notions by proving that any system can be deemed fair from the perspective of comparative fairness (e.g. individual fairness and statistical parity) if it is non-comparatively fair with respect to an auditor who has been deemed fair with respect to the same fairness notions. We also show that the converse holds true in the context of individual fairness. A brief discussion is also presented regarding how our fairness notion can be used to identify fair and reliable auditors, and how we can use them to quantify biases in decision-support systems.

1 Introduction

In recent years, the rapid advancements in the fields of artificial intelligence (AI) and machine learning (ML) have resulted in the proliferation of algorithmic decision making in many practical applications such as decision-support systems for judges whether or not to release a prisoner on parole [2], decisions for banks regarding granting or denying loans [18], and product recommendations by e-commerce websites [26]. Although these algorithms are highly efficient, they are found to be biased and unfair which substantially affect human lives. For example, A study by ProPublica in [1] has shown that the recidivism scores computed by the COMPAS algorithm are found to be biased in terms of both race and gender. Similarly, racial discrimination is found in e-commerce services in online markets [8], as well as in life insurance premiums [29]. In order to reduce the biases in such recommender systems, researchers have proposed different notions of fairness, including individual fairness, statistical parity, equality of opportunity and other similar notions [4, 13, 19, 24, 32]. *Statistical parity* states that the prediction outcomes across different groups are ensured to have equal probabilities. Unlike some of the other formalizations of fairness, statistical parity is independent of the "ground truth" (i.e. correctness in the data). Even though such fairness notion can be achieved without making any assumptions on data, it does not ensure fairness to individuals or subgroups across different groups. For instance, if there are disproportional outcomes concerning non-protected attributes, statistical parity may show discrimination

against qualified individuals in the group due to incorrect predictions. On the contrary, *individual fairness* [6] states that similar individuals should be treated similarly concerning a particular task. Though the notion focuses on individuals rather than groups, it is hard to determine a metric to compute similarity between two individuals. However, Dwork pointed out in [6] that the establishment of metric on individuals already exists in many decision support systems such as experience and skills for a job, grades for college admission, and FICO scores for loan applications. Dwork emphasized that similarity metric can be constructed appropriately if the two individuals belong to the same group. On the other hand, human judgements are necessary if the two individuals belong two different groups.

Traditional algorithmic fairness notions have exhibited several drawbacks in the identification and mitigation of biases. Therefore, researchers have developed mechanisms where reliable audit feedback is collected from trusted human auditors to evaluate systems. For example, in [16], Kearns *et al.* proposed equality of opportunity and statistical parity notions when there are a large number of protected groups. Their work focuses on developing algorithms for learning classifiers that are fair with respect to large number of groups based on a formulation of two-player zero-sum game between a learner and an auditor. Learner strategy corresponds to classifiers that minimize the sum of prediction error and the Auditor is responsible for rectifying the fairness violation of the Learner. A similar work by Zhang and Neil [33] also designed audit algorithms for classifiers to identify discrimination in subgroups that have not been pre-defined. However, these approaches raise questions such as: what features should be sensitive? How will you define the infinite class of groups?. Another recent work by Kim *et al.* in [17] aims to compute a distance metric in the notion of individual fairness using an expert auditor. Auditor returns an unbiased estimator to compute the distance between two random individuals based on an unknown fairness metric. Similarly, Gillen *et al.* in [10] assumes the existence of an auditor who is capable of identifying the fairness violations made in an online setting based on an unknown metric. Jung *et al.* in [15] study an offline learning problem with subjective individual fairness which is benefited by human experts. Proposed algorithm obtains a feedback from human experts by asking them questions of the form: "should this pair of individuals be treated similarly or not?". On the other hand, Raji *et al.* in [23] suggests that auditing the algorithms can be achieved through internal organization development cycle which could help tackling the ethical issues raised in a company. Their framework is developed by a small team of auditors in a large company who present data and model documentation along with metrics to facilitate auditing in specific contexts.

1.1 Contributions

From a social science perspective, justice can be categorized into two principles: comparative and non-comparative justice [7, 21, 22]. The principle of *comparative justice* states that relatively similar cases are treated similarly and dissimilar cases are treated differently. The characteristic feature of comparative justice lies in deriving a comparison or contrast between the way in which some system has treated two individuals or groups. In contrast, the principle of *non-comparative justice* mandates that each person be treated precisely as he/she deserves or merits regardless of the way in which anyone else treated. Non-comparative justice arguments do not need any comparison or contrast between two or more individuals. Although many algorithmic and data-driven approaches have been proposed to improve different fairness notions as discussed, these approaches were executed based on the *principle of comparative justice*. Both individual fairness and statistical parity, either compare two individuals or two groups of individuals. This, along with strong criticism from the social scientists [21, 22]. To date, algorithmic fairness has only focused on the principle of comparative justice resulting in an impasse in the fairness literature, thereby raising a question as to what does fairness means from the perspective of *non-comparative justice*. Therefore, in this paper, we explore a

novel mathematical fairness notion using an auditor based on the principle of non-comparative justice. Auditor aims to classify the individuals using an intrinsic fair relation based on non-comparative justice. The major downside in such framework is determining whether the auditor is discriminatory or not. Hence, to compute the bias of an auditor, we investigate the relationship between comparative fairness notions (individual fairness and statistical parity) and our proposed notion of non-comparative fairness. We prove that, a system/entity satisfies comparative justice if it satisfies non-comparative fairness with respect to the auditor. We also show that converse holds true in the case of individual fairness. Once an auditor is deemed fair, his/her fair relation can be utilized to examine an unknown system.

2 Comparative Justice for Algorithmic Fairness

Comparative justice arguments require to articulate either a comparison or a contrast between the way in which some system has treated two or more individuals or groups. The principle of comparative justice can be formulated as a combination of two standards: (1) Similar cases must be treated similarly, and (2) dissimilar cases must be treated differently. In some contexts, the system might treat individuals differently in a situation in which they should be treated similarly. This leads to comparative injustice. For instance, imagine a situation where two individuals approach a corporate bank to apply for a loan. Even though both of them have relevantly similar qualifications such as credit history, employment, and salary, the bank decides to grant the loan for only one of the individuals. This argument violates the principle of comparative justice because it violates the normative rule, "similar cases must be treated similarly". The bank should either grant both or refuse both. On the contrary, suppose there exist two industries A and B which cause pollution to the environment in varying amounts. Though B produces twice the amount of pollution compared to A, the government imposes the same amount of tax on both the industries. This argument violates the principle of comparative justice as it violates the *second standard*, "dissimilar cases must be treated differently".

As mentioned earlier, algorithmic fairness literature has focused on the principle of comparative justice, mainly, statistical parity and individual fairness. The notion of *statistical parity* says that, the prediction outcomes across different groups are ensured to have equal probabilities. More formally,

DEFINITION 1 (STATISTICAL PARITY). *Given protected attributes A, if f is a predictor and $Y = f(X)$, where, X is the multi-attribute variable and Y is the outcome, f achieves statistical parity when*

$$P[f(x) = y \mid A = a] = P[f(x) = y \mid A = a'] \tag{1}$$

holds true for all $y \in Y$, $x \in X$, and $a, a' \in A$.

Here, we can observe that the notion of statistical parity compares two different groups of protected attributes a and a'. Different versions of group-conditional metric led to different notions of statistical parity. M. Hardt *et al.* in [13] introduced the notion *equality of opportunity* which states that the true positive rate should be the same for all the groups. Similarly, Zafar *et al.* in [31] proposes *disparate mistreatment* which states that the accuracy of decisions is equal across groups of sensitive attributes with respect to false positive rates. On the other hand, Zafar in [32] says that the existing notions of fairness, based on equality in treatment, tends to limit the decision making accuracy, and recommends *preference-based fairness* giving the liberty to a group of users to prefer its treatments or outcomes. Though the notion gives the liberty for a group of users, it does not emphasize on individual preferences. It is possible that few individuals in a group may prefer another option than the one preferred by most of them in the group.

On the contrary, the notion of *individual fairness* [6] states that similar individuals should be treated similarly concerning a particular task. The notion can be defined as follows.

DEFINITION 2 (INDIVIDUAL FAIRNESS). *Given a metric $\mathcal{D}$ for a classification task T with an outcome set O, a randomized predictor $f(X) = Y$ and distance measure $d : \Delta(O) \times \Delta(O) \rightarrow \mathbb{R}$, f is individually fair if and only if*

$$d(f(x_i), f(x_j)) \leq \mathcal{D}(x_i, x_j) \tag{2}$$

for all $x_i, x_j \in X$.

Although individual fairness compares two individuals, it does not adhere to the principle of comparative justice. In other words, the notion does not enforce on treating dissimilar individuals differently. Also, the formulation relies on a suitable similarity metric which is difficult to construct in reality. As pointed out by Speicher *et al.* in [27], individual fairness does not regard individual's merits or desires for different decisions. Speicher proposed a way to measure individual unfairness using inequality indices and taking a person's merit into consideration. On the other hand, M. Kusner *et al.* in [19] developed a novel framework from causal inference, where a decision or an outcome is considered fair for an individual, only if it is the same in the actual as well as the counterfactual world. For more details, interested readers may refer to a detailed survey on fairness in algorithmic decision-making in [5, 20].

3   Non-comparative Fairness in Human Judgements

Comparative fairness notions are defined based on a comparison between two persons, or two groups of people. However, non-comparative justice notions deviate from this assumption and define justice based on the appropriate treatment to each individual.

The principle of non-comparative justice can be formulated as, "Treat each person as he/she deserves or merits". Such an argument does not depend upon any comparison or contrast with the way in which some system treats two or more individuals or groups. For instance, historically, African-Americans are said to commit more crimes when compared to other races [9]. As a result, the COMPAS tool is biased towards African-Americans even though the severity of their crimes is less compared to other races. Since COMPAS *compares* the historical information to present data, it is *non-comparatively unfair*. Hence, we demonstrate the notion of non-comparative fairness with the help of an expert auditor who classifies inputs based on a fair relation regardless of the historical information or any comparisons. We define non-comparative fairness as follows.

DEFINITION 3 (ϵ-NONCOMPARATIVE FAIRNESS). *Let f denote a fair assessment (i.e. input-output relationship) of a given system, i.e. $y = f(x)$, which is evaluated subjectively by an expert auditor. If g is an alternative representation, i.e. $\tilde{y} = g(x)$ (e.g. classification algorithms minimizing loss function, recommender systems), then g is called $\epsilon-$noncomparatively fair w.r.t. f if*

$$d\left(g(x), f(x)\right) < \epsilon, \text{ for all } x \in X. \tag{3}$$

Though the above notion compares the auditor's fair assessment f with the classifier/recommender g, it does not compare two different individuals thereby adhering to the principle of non-comparative justice. However, note that non-comparative fairness notions have their drawbacks. For example, if the auditor is discriminatory, then f is no longer a fair relation. Therefore, it is necessary to investigate how traditional fairness notions are related to the notion

of non-comparative fairness. One important assumption in our analysis is that we assume that both the human auditor and the algorithm employ the same distance metric d in evaluating the gap between input-output relationships.

3.1 Relation with Individual Fairness

For the sake of this discussion, we define a weaker definition for individual fairness in the following:

DEFINITION 4 (WEAK INDIVIDUAL FAIRNESS). *Given any two individuals $x_i, x_j \in X$ with $\mathcal{D}(x_i, x_j) \geq \kappa$, then g is (κ, δ)-individually fair if $d\left(g(x_i), g(x_j)\right) \leq \kappa + \delta$.*

Note that the above definition reduces to the original individual fairness notion in Definition 2 when $\delta = 0$. Recall that individual fairness partially adopts the principle of comparative fairness by comparing two individuals. In the following proposition, we show how the relation g can be evaluated based on the notion of (κ, δ)-individual fairness, when g is non-comparatively fair with respect to another individually fair relation f.

PROPOSITION 1. *g is $(\kappa, 2\epsilon + \delta)$-individually fair, if g is ϵ-noncomparatively fair with respect to f, and f is (κ, δ)-individually fair.*

PROOF. Given (x_1, y_1) and (x_2, y_2), if $\mathcal{D}(x_1, x_2) \geq \kappa$ (the two individuals are κ-similar), then f is (κ, δ)-individually fair if $d\left(f(x_1), f(x_2)\right) < \kappa + \delta$. However, note that if g is ϵ-noncomparatively fair 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 obtain

$$
\begin{aligned}
d\left(g(x_1), g(x_2)\right) &\leq d\left(g(x_1), f(x_1)\right) + d\left(f(x_1), f(x_2)\right) + d\left(f(x_2), g(x_2)\right) \\
&< 2\epsilon + \kappa + \delta.
\end{aligned}
\tag{4}
$$

□

We illustrate this result using the following example from the banking domain. For instance, consider two individuals who are looking to apply for a loan. An individually fair banking system evaluates both the applications via collecting various customer's attributes 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 is individually fair. If the banking evaluation system is relatively similar to the auditor's fair relation, from Proposition 1, the banking system is also individually fairness. Otherwise, if the banking system is not non-comparatively fair with respect to the auditor, then the system itself is also not individually fair.

PROPOSITION 2. *If f is not individually fair and if g is ϵ-noncomparatively fair with respect to f, then g is also not individually fair.*

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

$$
d(f(x_1), f(x_2)) \leq d(g(x_1), f(x_1)) + d(g(x_2), f(x_2)) + d(g(x_1), g(x_2))
\tag{5}
$$

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

$$
\begin{aligned}
2\epsilon + d\left(g(x_1), g(x_2)\right) \quad &> \quad d\left(g(x_1), f(x_1)\right) + d\left(g(x_2), f(x_2)\right) + d\left(g(x_1), g(x_2)\right) \\
&\geq \quad d\left(f(x_1), f(x_2)\right) > \kappa + \delta
\end{aligned}
\tag{6}
$$

for all $\kappa, \delta \in \mathbb{R}$. Therefore, we also have

$$
d\left(g(x_1), g(x_2)\right) \quad > \quad \tilde{\kappa}
\tag{7}
$$

for all $\tilde{\kappa} \in \mathbb{R}$. $\square$

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 rule seem to be similar, they violate the precept, "treat similar individuals similarly". Moreover, the outcome violates the principle of non-comparative justice, since A is treated in a way that A does not merit. Hence, we can assert that banking evaluation does not satisfy individual fairness.

PROPOSITION 3. *Even though f is (κ, δ)-individually fair, g is also not individually fair if g is not noncomparitively fair with respect to f.*

PROOF. If g is not noncomparitively fair with respect to f, then there exists some $x \in \mathcal{X}$ such that $d\left(g(x), f(x)\right)$ is unbounded. Let x_1 be the input for which $d\left(g(x_1), f(x_1)\right)$ is unbounded, i.e. $d\left(g(x_1), f(x_1)\right) > \pi$ for any $\pi \in \mathbb{R}$. At the same time, let x_2 denote another input for which $d\left(g(x_2), f(x_2)\right)$ is bounded by ϵ.

By applying a chain of triangle inequalities, we know

$$
d\left(g(x_1), f(x_1)\right) \quad \leq \quad d\left(g(x_1), g(x_2)\right) + d\left(g(x_2), f(x_2)\right) + d\left(f(x_2), f(x_1)\right)
\tag{8}
$$

Substituting the bounds for the terms $d\left(g(x_2), f(x_2)\right)$ and $d\left(f(x_1), f(x_2)\right)$, we obtain

$$
\begin{aligned}
d\left(g(x_1), g(x_2)\right) + \epsilon + \kappa + \delta \quad &> \quad d\left(g(x_1), g(x_2)\right) + d\left(g(x_2), f(x_2)\right) + d\left(f(x_2), f(x_1)\right) \\
&> \quad \pi,
\end{aligned}
\tag{9}
$$

for any $\pi \in \mathbb{R}$. In other words, we have

$$
d\left(g(x_1), g(x_2)\right) > \pi - (\epsilon + \kappa + \delta),
\tag{10}
$$

for any $\pi \in \mathbb{R}$. $\square$

To illustrate Proposition 3, consider the recent accusation of inequality faced by the Apple Card (credit card by Apple Inc.) [14]. The government has launched an investigation into the company's practices after a user accused Apple of gender bias in determining credit limits. The user received a limit 20 times higher than his wife's even though both the profiles are relatively similar. As per Apple's practices, similar individuals are *not* treated similarly. Now, consider an auditor who determines the credit limits of the users using the rule: "If the credit score of the user is greater than x and payments are paid duly, the credit line must be \$3000. If not, the credit line can be \$1000". Note that, the auditor's fair relation treats similar individuals similarly. We can observe that Apple's evaluation of determining credit limits

is significantly different from auditor's rule. Therefore, Apple does not satisfy non-comparative fairness and thereby, dissatisfying individual fairness.

3.2 Relation with Statistical Parity

We define a weaker definition for group fairness in the following:

DEFINITION 5 (WEAK STATISTICAL PARITY). *Given set of protected attributes A, f satisfies δ-statistical parity when*

$$P[f(x) = y \mid A = a] - P[f(x) = y \mid A = a'] \leq \delta \tag{11}$$

holds true for all $y \in Y$, $x \in X$, and $a, a' \in A$.

Note that the above definition equates Definition 1 when the difference between the probabilities is equal to zero. As discussed earlier, statistical parity resembles the principle of comparative justice by comparing two different protected groups. In the remaining section, we will focus on the relationship between statistical parity and non-comparative fairness. For the sake of convenience, let us denote $p_{x,y}(g, a) = P[g(x) = y \mid A = a]$.

PROPOSITION 4. *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-noncomparatively fair 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

$$
\begin{aligned}
p_{x,y}(g, a) - p_{x,y}(g, a') &= \left[p_{x,y}(g, a) - p_{x,y}(f, a) \right] - \left[p_{x,y}(g, a') - p_{x,y}(f, a') \right] \\
&\quad + \left[p_{x,y}(f, a) - p_{x,y}(f, a') \right]
\end{aligned}
\tag{12}
$$

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

An illustrative example is the *reservation system* enforced by the Indian Constitution which sets quotas to socially and educationally backward communities to offer fair representation in the field of education, employment, and politics [30] from a group fairness perspective. Reservations are employed to 3 minority groups: Scheduled Castes (7.5% quota), Scheduled Tribes (15% quota), and Other Backward Classes (27% quota) ranked in decreasing order of their income cap. The remaining quota of jobs/admissions are allotted to the merit group of citizens. Now, suppose an auditor presents fair judgements based on the rule: "If group A's annual income is less than $8000, then 30% of the quota of jobs/college admissions are allotted to A. Similarly, if group B's annual income is greater than $15000, then 17% of the quota is assigned to B". Note that, the auditor's fair relation is somewhat similar to that of the government's policy. Since the government's policy is known to be fair, the auditor is also unbiased from a group fairness perspective.

4 Identification of Fair Auditors

Several attempts have been made to identify how people perceive fairness in order to automate the process of fixing fairness from an algorithmic standpoint [3, 11, 12, 25, 28]. However, the underlying assumption in most of the fairness literature is that these systems are evaluated by fair and unbiased auditors. This is not always true because people exhibit a wide range of biases based on diverse prior experiences. As a solution to this problem, we will demonstrate

how the proposed non-comparative fairness notion can be used to identify fair auditors. A direct approach is to compare an auditor's evaluation to a benchmark entity (e.g. a well-studied system, a known human expert) whose biases are well-quantified in terms of both weak individual fairness and weak group fairness notions as defined in Section 3. Using Propositions 1 and 4, we can evaluate the biases of an unknown auditor as a function of biases present in the benchmark entity, as stated below.

COROLLARY 1 (TO PROPOSITION 1). *If the benchmark entity is (κ, δ)-individually fair, then the unknown auditor can be deemed (κ, δ')-individually fair when the auditor is ϵ-noncomparitively fair with respect to the benchmark entity such that*

$$\epsilon < \frac{\delta' - \delta - \kappa}{2}.$$

PROOF. Let g denote the unknown auditor and f denote the benchmark entity. From Proposition 1, if f is (κ, δ)-individually fair and g is ϵ-noncomparatively fair with respect to f, we have

$$d\left(g(x_1), g(x_2)\right) \quad < \quad 2\epsilon + \kappa + \delta \tag{13}$$

However, our goal is to identify auditors who are (κ, δ')-individually fair. In other words, we need δ' to be at least as large as $2\epsilon + \kappa + \delta$. In other words,

$$\delta' \quad > \quad 2\epsilon + \kappa + \delta. \tag{14}$$

Upon rearranging the terms, we get the bound on ϵ stated in this corollary.

□

Similarly, we can also quantify the tolerable bias in an unknown auditor if the goal is to identify a fair auditor from a weak statistical parity sense. This is discussed in the following corollary.

COROLLARY 2 (TO PROPOSITION 4). *Assuming M-Lipschitz continuity in probability distributions at both the benchmark entity and the unknown auditor, if the benchmark entity satisfies δ-statistical parity, then the unknown auditor can be deemed to satisfy δ'-statistical parity when the auditor is ϵ-noncomparitively fair with respect to the benchmark entity such that*

$$\epsilon < \frac{\delta' - \delta}{2M}.$$

PROOF. Let g denote the unknown auditor and f denote the benchmark entity. From Proposition 4, if all probability distributions are M-Lipschitz continuous, f satisfies δ-statistical parity and g is ϵ-noncomparatively fair with respect to f, we have

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

However, our goal is to identify auditors who satisfy δ'-statistical parity. In other words, we need

$$\delta' > 2M\epsilon + \delta. \tag{16}$$

Upon rearranging the terms, we obtain the bound on ϵ as stated in the corollary statement.

□

Thus, any given auditor can be evaluated in terms of traditional fairness notions (e.g. individual fairness, statistical parity) using our proposed non-comparative fairness notion. Once the biases within a benchmark system is quantified, we can identify fair auditors via screening them using a threshold-based rule on their biases from a non-comparative fairness perspective. Note that the threshold used in this classifier can be computed based on the designer's bias tolerance. Having identified fair auditors along with quantifying their respective biases, we can leverage them as benchmark entities to evaluate an unknown recommender system via computing the system's biases using the auditor's evaluation.

5 Conclusion and Future Work

We introduced a non-comparative fairness notion which complements the existing comparative fairness notions proposed in the algorithmic fairness literature. We showed that any system can be deemed fair from the perspective of comparative fairness (e.g. individual fairness and statistical parity) if it is non-comparatively fair with respect to an auditor who has been deemed fair with respect to the same fairness notions. We also proved that the converse holds true in the context of individual fairness. We discussed how the proposed non-comparative fairness notion can be used to identify fair auditors who are hired to evaluate latent biases in decision-support systems. In our future work, we will develop novel algorithms to identify fair auditors from the perspective of multiple (potentially intransitive) attributes, and also validate our theoretical findings using real data.

References

[1] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. 2016. Machine Bias. *ProPublica* (May 23 2016).

[2] Anna Maria Barry-Jester, C. Ben, and G. Dana. 2015. The New Science of Sentencing. *The Marshall Project* (August 08 2015).

[3] Reuben Binns. 2019. Human Judgement in Algorithmic Loops; Individual Justice and Automated Decision-Making. *Individual Justice and Automated Decision-Making (September 11, 2019)* (2019).

[4] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 339–348.

[5] Alexandra Chouldechova and Aaron Roth. 2018. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810* (2018).

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

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

[8] Ray Fisman and Michael Luca. 2016. Fixing Discrimination in Online Marketplaces. In *Harvard Business Review*.

[9] New Century Foundation. 1999. The Color of Crime - Race, Crime and Violence in America. (1999). https://2kpcwh2r7phz1nq4jj237m22-wpengine.netdna-ssl.com/wp-content/uploads/2011/12/1999-Color-of-Crime-Report.pdf

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

[11] Jerald Greenberg. 1986. Determinants of perceived fairness of performance evaluations. *Journal of applied psychology* 71, 2 (1986), 340.

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

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

[14] Jon Fingas. 2019. New York Investigates Claims of Sexism in Apple Card Credit Limits. https://www.engadget.com/2019-11-09-new-york-investigates-apple-card-credit-limit-sexism.html.

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

[16] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2017. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144* (2017).

[17] Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Fairness through computationally-bounded awareness. In *Advances in Neural Information Processing Systems*. 4842–4852.

[18] James Rufus Koren. 2016. What Does That Web Search Say About Your Credit? *Los Angeles Times* (July 17 2016).

[19] Matt J Kusner, J. Loftus, C. Russell, and R. Silva. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems*. 4066–4076.

[20] B. Lepri, N. Oliver, E. Letouzé, A. Pentland, and P. Vinck. 2018. Fair, Transparent, and Accountable Algorithmic Decision-making Processes. *Philosophy & Technology* 31, 4 (01 Dec 2018), 611–627.

[21] Raleigh Hannah Levine and Russell Pannier. 2005. Comparative and noncomparative justice: some guidelines for constitutional adjudication. *Wm. & Mary Bill Rts. J.* 14 (2005), 141.

[22] Phillip Montague. 1980. Comparative and non-comparative justice. *The Philosophical Quarterly (1950-)* 30, 119 (1980), 131–140.

[23] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* 33–44.

[24] Ya'acov Ritov, Yuekai Sun, and Ruofei Zhao. 2017. On conditional parity as a notion of non-discrimination in machine learning. *arXiv preprint arXiv:1706.08519* (2017).

[25] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. 2019. 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.* 99–106.

[26] Brent Smith and Greg Linden. 2017. Two Decades of Recommender Systems at Amazon.com. *IEEE Internet Computing* 21, 3 (2017), 12–18.

[27] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* 2239–2248.

[28] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. 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.* 2459–2468.

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

[30] Wikipedia contributors. 2020. Reservation in India — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Reservation_in_India&oldid=969093785. [Online; accessed 29-July-2020].

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

[32] Muhammad Bilal Zafar, I. Valera, M. Rodriguez, K. Gummadi, and A. Weller. 2017. From Parity to Preference-based Notions of Fairness in Classification. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 229–239.

[33] Zhe Zhang and Daniel B. Neill. 2016. Identifying Significant Predictive Bias in Classifiers. arXiv:stat.ML/1611.08292